Statistical Mechanics of Neural Networks

William Whyte

St Peter's College,

Oxford

Theoretical Physics,

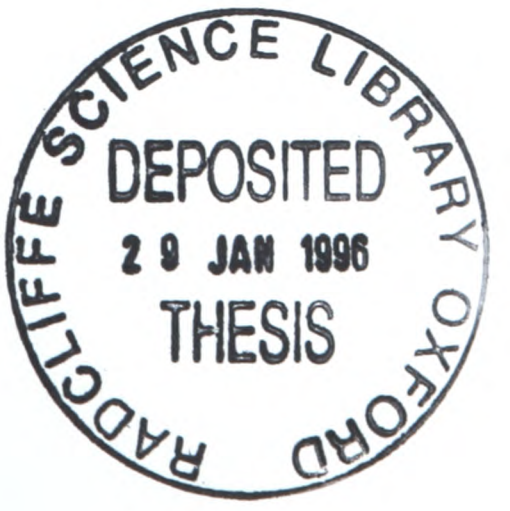
Department of Physics,

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

in the University of Oxford

Trinity Term, 1995



Abstract

We investigate five different problems in the field of the statistical mechanics of neural networks.

The first three problems involve attractor neural networks that optimise particular cost functions for storage of static memories as attractors of the neural dynamics. We study the effects of replica symmetry breaking (RSB) and attempt to find algorithms that will produce the optimal network if error-free storage is impossible.

For the Gardner-Derrida network we show that full RSB is necessary for an exact solution everywhere above saturation. We also show that, no matter what the cost function that is optimised, if the distribution of stabilities has a gap then the Parisi replica ansatz that has been made is unstable.

For the noise-optimal network we find a continuous transition to replica symmetry breaking at the AT line, in line with previous studies of RSB for different networks. The change to RSB1 improves the agreement between "experimental" and theoretical calculations of the local stability distribution $\rho(\lambda)$ significantly. The effect on observables is smaller.

We show that if the network is presented with a training set which has been generated from a set of prototypes by some noisy rule, but neither the noise level nor the prototypes are known, then the perceptron algorithm is the best initial choice to produce a network that will generalise well. If additional information is available more sophisticated algorithms will be faster and give a smaller generalisation error.

The remaining problems deal with attractor neural networks with separable interaction matrices which can be used (under parallel dynamics) to store sequences of patterns without the need for time delays. We look at the effects of correlations on a single-sequence network, and numerically investigate the storage capacity of a network storing an extensive number of patterns in such sequences.

When correlations are implemented along with a term in the interaction matrix designed to suppress some of the effects of those correlations, the competition between the two produces a rich range of behaviour. Contrary to expectations, increasing the correlations and the operating temperature proves capable of improving the sequence-processing behaviour of the network.

Finally, we demonstrate that a network storing a large number of sequences of patterns using a Hebb-like rule can store approximately twice as many patterns as the network trained with the Hebb rule to store individual patterns.

Contents

Chapter 1

Introduction

1.1 Motivations and Basic Definitions

1.1.1 Introductory Remarks

The study of neural networks has been a massively fruitful collaboration between scientists from such diverse disciplines as biochemistry, experimental psychology, computer science and theoretical physics. The interest that neural networks have provoked is due mainly to two phenomena: their ability to function as an associative memory and their ability to generalise ([A89], [MR91]). An associative memory is one that recalls stored information in response to a stimulus that resembles that information in some way; in simpler terms, it is a process of *recognition.* The ability to generalise means that a neural network, when trained to produce the correct responses to an appropriately chosen *training set* of example stimuli and responses, will produce the correct response to a stimulus it has not experienced before, but which is similar to a learned example. Both of these are features of human intelligence; both are features that it is hard to implement on a traditional von Neumann type of computer. Neural networks also display other desirable attributes, such as robustness against small disruptions of the structure of the network or against an element of randomness in the dynamics, which are suggestive of the properties of the brain.

In this thesis, we use the methods of statistical physics to research the behaviour of certain kinds of neural networks. The use of these methods in this subject is inspired by the philosophy that we can gain insight into the operation of a system by considering an extremely simple model of that system which appears to preserve the system's basic structure. This has the twin advantages that we reduce the complexity of the system to the extent that we can consider a mathematical treatment of it, and that we can determine which features of the system behaviour are due to the underlying structure and which are due to causes that only become apparent after more detailed consideration of the specifics of the system.

A further motivation for the use of statistical physics comes from research into disordered magnetic systems, known as "spin glasses". We will go into this in more detail in later sections; in brief, however, spin glasses provide an example of a disordered system with many metastable states (to be compared to the multiplicity of memories in a neural network) to which the system will iterate under its dynamics (to be compared to recognition). It was the demonstration of Hopfield [H82] of an exact analogy between certain spin glass models and certain models of neural systems that opened the door to the current advances that are being made in the field. For now, however, we introduce a discussion of neural networks inspired by biology and by the philosophical intuition just mentioned; the analogy to spin glasses will be made only where relevant.

1.1.2 A Model of a Neuron

A neural network is specified by the individual neuron dynamics, the structure of the network, and a rule determining the order in which the neurons are updated. We discuss each of these in turn, starting with the model of a neuron that we will use.

The most general description of a formal neuron is that it is a device that receives stimuli from other neurons and in response produces an output, often referred to as its "state". We can represent this mathematically as

$$S_i[t] = g\left(\{S_j[t']\}\right) \tag{1.1}$$

where $j = 1, \ldots, N$ runs over all the neurons in the system, and $t' < t$ runs over all previous times.

We look to the human brain for guidance in attempting to make this definition more specific.

In the cerebral cortex there are about 10^{10} neurons, each connected to about 10^4 others via *synapses*. As a first approximation, we can say that neurons operate by summing the inputs from the neurons connected to them; that if the total input is greater than some threshold, the neuron emits an electric signal or "spike"; and that all spikes are the same as each other [C86]. Inspired in part by the physiology just described and in part by a desire for simplicity, we adopt the formulation of McCullough and Pitts [MP43] and consider binary neurons $S_i = \pm 1$ updated according to the rule

$$S_i[t] = \text{sign}\left(\sum_{ij} J_{ij} S_j[t-1] + \theta_i\right), \tag{1.2}$$

where the J_{ij}'s are the pairwise interactions between the neurons, known as the *connection strengths* or *synaptic weights*, and we refer to the entire set of $\{J_{ij}\}$ as the *synaptic matrix.* We generalise this to a stochastic rule in which the probability P of a neuron S_i taking the value ± 1 at a time t is

$$P(S_i[t]) = \tfrac{1}{2}(1 + S_i[t] \tanh[\beta\left(\sum_{j \neq i} J_{ij} S_j[t-1] - \theta_i + h_i^{\text{ext}}[t]\right)]) \tag{1.3}$$

where the inverse operating temperature $\beta \equiv T^{-1}$ is a measure of the noise in the system and the thresholds θ_i and the external fields $h_i^{\text{ext}}[t]$ are included for completeness. The "memory" of the system is thus contained in the synaptic weights J_{ij} and the thresholds θ_i. Many other forms for the update rule could plausibly be considered; for example, a neuron could be modelled by a positive real number, representing its firing rate, or more complicated interactions than pairwise ones could be taken into consideration, adding terms of the form $\sum_{j_1 \ldots j_r} J_{i,j_1 \ldots j_r} S_{j_1} \ldots S_{j_r}$ to the update rule above, or we could lose the assumption that all interactions are instantaneous and instead consider the effects of

different stimuli arriving at different times. However, despite its simplicity, the update rule described remains a strong foundation for interesting behaviour, enlightening analogies to the brain, and (most importantly from the point of view of the D Phil student) unanswered questions. It also has a form highly similar to that for the single spin update rule in certain spin glass models [EA75], thus creating the possibility of using methods from spin glass theory to study neural networks.

1.1.3 Network Architecture

Having defined our neurons, we consider the arrangement of the connections between them, also referred to as the *network architecture*. Two forms of network are commonly considered: *feed-forward* or *hetero-associative* networks, and *feedback* or *auto-associative* networks.

Feed-forward networks are characterised by having "layers" of neurons through which information is passed, with each neuron only receiving input from the previous layer and only passing output to the following layer. In the context in which they are most frequently used, that of providing an answer to a question or (equivalently) a response to a stimulus, there is an initial "input" layer and a final "output" layer, and we can consider the network to be a mapping $\{\pm 1\}^{N_{in}} \rightarrow \{\pm 1\}^{N_{out}}$. It has been shown that, provided there is at least one intermediate layer, such a network with one output neuron is capable of reproducing any Boolean function [MP43]. Our discussion of feedforward networks will revolve mainly around the *perceptron* [MP88], the simplest form of feedforward network, which consists of a single output neuron with N inputs. Although this network is not capable of as wide a range of behaviour as the network with intermediate layers, its analytic tractability makes it an extremely useful research tool.

In a feedback network, by contrast, there is no sense of an overall direction to the information flow; a signal may (in theory) go from any neuron to any other neuron and two neurons may each provide input to the other. If we look on the network as a mapping, it is from $\{\pm 1\}^N \rightarrow \{\pm 1\}^N$. In this case our interest will be in the global state of the

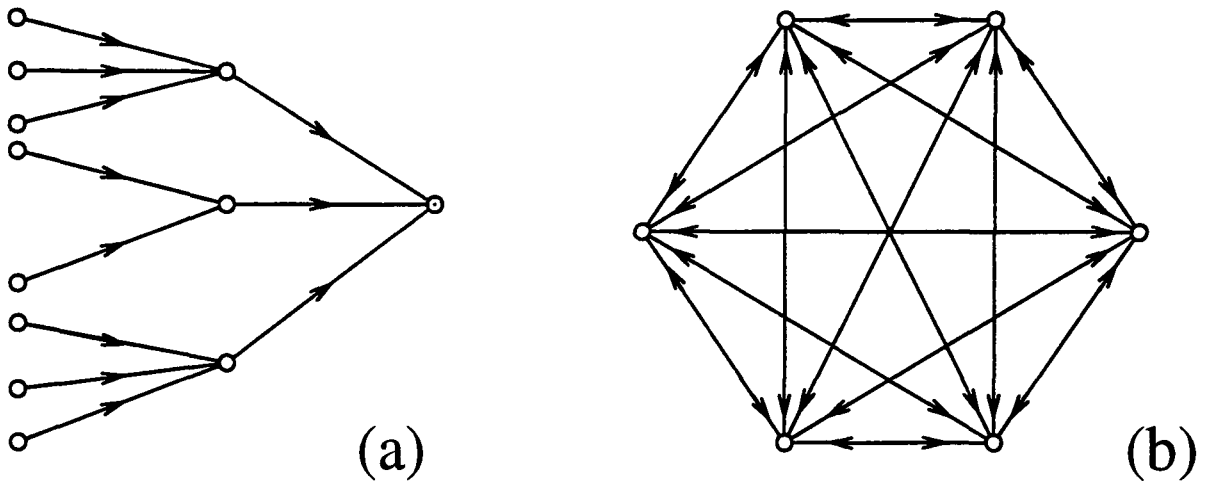


Figure 1.1: Schematic diagrams of two neural network layouts: (a) the tree-like feed-forward net; (b) the fully-connected feedback net.

network and how it relates to specified states or "patterns".

Extreme cases of these two forms of network are illustrated in Figure 1.1. The tree-like network of Figure 1.1a is characterised by the fact that information from a given input neuron can only reach the output neuron by one route. In the fully-connected feedback network of Figure 1.1b every neuron is connected to every other. As would be expected from the conceptual simplicity of their structure, these two kinds of networks are the most easily analysed ones. However, we could also consider intermediate cases, such as feedforward nets in which a neuron in one layer sends a signal to more than one neuron in the next layer, or networks combining feedforward and feedback aspects in which the neurons within a layer receive input from the previous layer and send output to the following layer, but are also connected to other neurons within their own layer. A particularly useful intermediate case, as we shall see, is the feedback net of such sparse connectivity that it resembles the tree-like feed-forward net.

1.1.4 Dynamics

As previously stated, the physical structure of the network, the synaptic matrix $\{J_{ij}\}$ and the individual neuron update rule (1.3) do not, on their own, fully determine the behaviour of the network. The final piece of information that must be provided is a rule stipulating the order in which the neurons are updated.

For the feedforward network, we usually simply assume that all neurons in one layer

are updated before any neuron in the next is. For the feedback network, there are two standard choices for the dynamics, *parallel* or *synchronous* and *sequential* or *asynchronous*. Under parallel dynamics, all sites are updated simultaneously; under sequential dynamics, the sites are updated one at a time in a random order. The dynamics are more complicated than for the feedforward network because a neuron can, through the interactions with other neurons, be indirectly influenced by its own past states. Dynamic calculations must therefore be treated with care, and in large measure the calculations in this thesis concern the fixed-points of the dynamics rather than the dynamical processes themselves, which we simulate numerically. We therefore consider next the uses to which these fixed-points can be put in the context of considering the network as an associative memory.

1.2 Equilibrium States: Storage of Patterns

Our interest in the first part of this thesis is primarily in associativity in feedback neural networks. We introduce this topic through consideration of the fixed-points of the dynamics, just discussed.

Consider a set of binary patterns $\{\xi_i^\mu\}$ $(1 \leq \mu \leq p; 1 \leq i \leq N)$, which correspond to desired states of the network. From the example of spin glasses [EA75], we know it is possible for the system to have many metastable states; we now wish to construct a set of J_{ij}'s such that these metastable states correspond to the specified patterns. In other words (referring to (1.3)), the patterns satisfy

$$\xi_i^\mu = \text{sign}\left(\sum_j J_{ij}\xi_j^\mu\right) \quad \forall\mu. \tag{1.4}$$

(note that here, and throughout the thesis, we are taking the thresholds and external fields to be zero).

In order to measure the similarity between the network state $\vec{S}$ and a desired state

(pattern) $\vec{\xi}^{\mu}$, we define the *overlap* q_{μ} as

$$q_{\mu} \equiv \frac{1}{N} \sum_{i} S_i \xi_i^{\mu}. \tag{1.5}$$

The overlap q_{μ} thus ranges from 1 if the system state is exactly the same as the pattern $\vec{\xi}^{\mu}$, to 0 if it is uncorrelated, to -1 if the neural value on each site is the opposite to what it is in the pattern.

We can restate the storage criterion in terms of q, at the same time taking the opportunity to slightly relax the condition that the system state and the pattern must agree on all sites: we now require that, in the absence of correlations between patterns, with every pattern there is a stable state of the network associated which has $q_{\mu} = \mathcal{O}(1), q_{\nu \neq \mu} = \mathcal{O}(1/\sqrt{N})$ (In practice there are many networks, such as the Gardner perceptron described in Section 3.2, where the stricter condition will be satisfied; but for other forms of network, such as the Hopfield network of Section 3.1 or any feedback network at finite operating temperature, the relaxation of the definition is necessary). Of course, under the stochastic dynamics of (1.3), no state of the network will be stable in the sense of being permanent and unchanging; in this context we rather mean that there is an attractor of the dynamics such that the network state $\vec{S}$ only explores a small region of its state space, and that within this region fluctuations in the measured overlap are small. This overlap is then referred to as the *retrieval overlap.*

A primary quantity of interest for any neural network is the number of patterns per synapse that can be stored by the criteria just defined. This is known as the *storage capacity.* Later in this thesis we will consider issues of storage capacity for various different forms of neural network.

1.3 Dynamics I: Basins of Attraction and Associativity

Merely storing the patterns is not enough to guarantee the functioning of the network as an associative memory. We also require that the patterns are stored as attractors of the dynamics. If this is the case, a network started in a state sufficiently similar to (in other words, having a high enough overlap with) a stored pattern will then iterate under the network dynamics to the retrieval state corresponding to that stored pattern, displaying the property of pattern recognition that has already been mentioned.

Any initial state that iterates to a stored pattern is referred to as lying within that pattern's *basin of attraction.* We can therefore measure the associativity of a network by the size of the basins of attraction, which in turn is given by the minimum overlap necessary for retrieval. The associativity has then become a quantity which we can calculate and attempt to maximise, subject to the proviso, already mentioned, that the exact dynamical equations are very difficult to solve due to the correlations between sites. However, calculations are possible in certain limiting cases, which we now discuss.

1.3.1 Initial Change in the Overlap

Various authors [KA88], [A&a] have given a method for calculating the initial dynamical behaviour for any specified neural network, given an input overlap m with one pattern and microscopic overlaps with the other patterns. In order to explain this, we introduce a natural measure of how "strongly" the patterns are stored, the *pattern stability* or *local aligning field* at a site i, λ_i^μ, defined as follows:

$$\lambda_i^\mu = \xi_i^\mu \sum_j J_{ij}\xi_j^\mu / |J_i| \tag{1.6}$$

(the normalisation $1/|J_i|$, $|J_i| \equiv \sqrt{\sum_j J_{ij}^2}$ is included in view of the fact that the storage criterion (1.4) is invariant under the rescaling $J_{ij} \rightarrow cJ_{ij} \forall i,j$). For a given synaptic

prescription, one can calculate the *local stability distribution* $\rho(\lambda)$,

$$\rho(\lambda) = \langle \lambda^{\nu} - \lambda \rangle. \tag{1.7}$$

The inital change in the overlap for a system with operating temperature $T = \beta^{-1}$ is then given by

$$\begin{aligned} m[t+1] &= f(m[t]) \qquad \text{(parallel dynamics)} \\ \tfrac{d}{dt} m[t] &= f(m[t]) - m, \quad \text{(sequential dynamics)} \end{aligned} \tag{1.8}$$

where

$$f(m[t]) = \int d\lambda \rho(\lambda) \int Dy \tanh\left(\beta\left[m[t]\lambda + \sqrt{1 - m^2[t]}y\right]\right), \tag{1.9}$$

$$Dy = \frac{dy}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}. \tag{1.10}$$

In the zero-temperature case the expression for $f(m[t])$ simplifies to

$$f(m[t]) = \int d\lambda\, \rho(\lambda) g(m[t], \lambda), \tag{1.11}$$

where $g(m, \lambda) = \text{erf}\,(m\lambda/\sqrt{2(1-m^2)})$, $\text{erf}\,(x) = \int_0^x e^{-u^2} du$. The stability distribution $\rho(\lambda)$ thus determines the initial dynamical behaviour for any network.

1.3.2 The Mapping $f(m)$ and the Highly Dilute Network

This rule cannot be used to determine the full dynamics of an arbitrary network; the form (1.8) for the change in overlap was obtained using the central limit theorem and is therefore inapplicable if correlations build up between sites. However, if the network is so dilute that a signal from a neuron takes an infinite length of time to get back to that neuron, the correlations will remain effectively zero and the equation (1.8) does indeed fully determine the dynamics at all times.

We now derive the criteria for this to be the case more precisely. If each neuron is

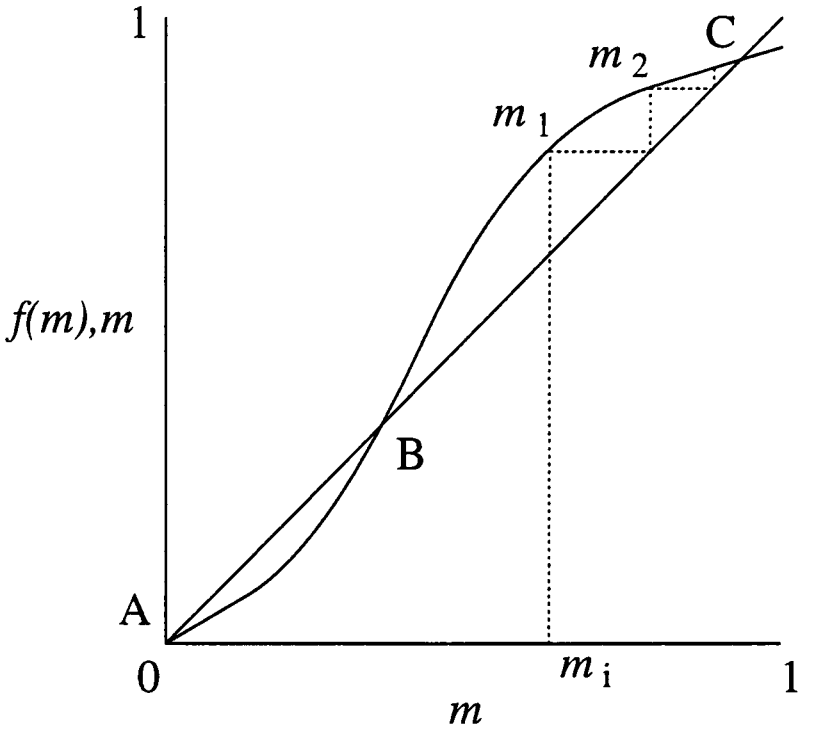


Figure 1.2: Qualitative representation of an iterative mapping $f(m)$; see the text for further details.

connected to C other neurons, then after t timesteps each neuron will have been influenced by C^t other neurons. In order to avoid correlations, we require that no neuron appears twice in this set; this condition is satisfied if $C^t \ll N$. We thus require $t \ll \ln N / \ln C$, which implies that $N/C \to \infty$ as $N \to \infty$. This is satisfied for $C \sim \mathcal{O}(\ln N)$, which therefore becomes the criterion for the retrieval dynamics to be determined by the mapping $f(m)$.

A network of this sort is called a "highly dilute network". It was first described by Derrida *et al* [DGZ87].

A qualitative example of the use of the mapping $f(m)$ to determine the dynamics of the network is given in Figure 1.2. Fixed-points of the mapping are those points where $f(m) = m$; the diagonal line is included to explicitly show these points. The fixed-points are the same for both parallel and sequential dynamics. A fixed-point is unstable or stable depending on whether or not at the fixed-point $f(m)$ is increasing slower or faster than m. In the figure, A and C are stable fixed-points of the mapping, and B is an unstable fixed-point. Any initial overlap below B will iterate to A; any initial overlap over it will iterate to C. The dotted line shows the evolution of an initial overlap m_i towards the fixed-point C under parallel dynamics.

The mapping $f(m)$ and its fixed points enable us to characterise the retrieval behaviour of the network. We refer to "wide retrieval" if any finite overlap with a stored pattern is enough to guarantee retrieval. We refer to "narrow retrieval" if a larger overlap is required. As we shall see, the aims of having wide retrieval in as much as possible of the phase diagram and of having the maximum possible retrieval overlap are not necessarily compatible.

1.4 Dynamics II: Dynamical Memories

In discussing overlaps and basins of attraction we have developed a language which is capable of describing far more complex phenomena than the simple case we have considered up till now, of static attractors giving stable overlaps. An equally valid area of investigation is to consider networks where our aim is to have the state evolve in time, following some specified path. In the second part of the thesis we will consider in more detail the problem where we wish the network to "retrieve" a sequence of patterns $\{\vec{\xi}_1, \ldots, \vec{\xi}_l\}$ in order; in this case, we use "retrieve" in a looser sense than previously, to mean "have a high overlap with". The attractor in this case will not be any one of the specific patterns, but rather the entire sequence. In this case we are dealing with *dynamical* memory structures, rather than the *static* ones we have previously considered.

1.5 Outline of the Thesis

In the first part of this thesis we consider issues of basins of attraction and storage of static memories.

First, we outline, in some detail, the mathematical technique known as the "replica method", which is commonly used in consideration of disordered systems. We apply this, for illustrative purposes, to a generalised disordered magnetic system known as the *Sherrington-Kirkpatrick* spin glass, and then to neural networks. We discuss the *replica-symmetric* and *replica symmetry breaking* ansatzes used to obtain solutions. For two

specific cases, we perform replica symmetry broken calculations at a higher level than has been done before; we also provide a proof that an infinite degree of replica symmetry breaking is necessary for the exact solution of the network trained with the *Gardner-Derrida* rule, which is the rule that stores the maximum possible number of patterns.

Second, we consider the practicalities of training a network when there are more patterns than the network can store correctly. For the case where the actual patterns are not available at the training stage of the network and we must instead use corrupted examples, we determine the appropriate training algorithm.

In the second part of the thesis, which can be read separately from the first part, we consider the storage of sequences of patterns. We examine two cases in which sequences of patterns are subjected to the type of disruptive influences that might occur in nature – in the first case due to correlations between the patterns in the sequence, in the second case due to noise from the storage of other sequences.

In the first case, we find that if the synaptic matrix has a term which acts to accentuate the differences between patterns, then the introduction of correlations produces interesting effects as a result of the competition between the two influences. We even find a region of the phase diagram within which an increase in the correlations between patterns improves the network's ability to distinguish between the patterns better, rather than causing it to worsen as would intuitively be the case.

In the case of storage of large numbers of sequences of patterns, we find that the system is approximately twice as robust against noise as the corresponding non-sequence-processing system, and offer an explanation for this.

All of the work presented in this thesis was performed under the supervision of Professor David Sherrington. The research into replica-symmetry breaking in noise-optimal networks (section 3.4) was done in collaboration with David Sherrington and Dr. Michael Wong. The research into storing sequences of correlated patterns was done in collaboration with David Sherrington and Dr. Ton Coolen, and has previously been published in [WSC95].

Part I

Chapter 2

The Method of Replicas in Disordered Systems

2.1 Overview

Neural networks are typical of a broad class of disordered systems which can be treated within a statistical physics framework, namely systems described by two sets of random variables, $\{S\}$ and $\{J\}$ say, which may vary over different timescales. In neural networks these variables are the instantaneous states of the neurons and the synaptic efficacies that connect them, while in the study of disordered magnetic systems ("spin glasses") the variables are the spin states at the various spin sites and the interaction strengths between them.

In general, we assume that we can analyse these systems by reference only to the average values of certain quantities; after all, one sample of a disordered material is very like another sample of the same disordered material, or like the original sample at a different time. We are therefore implicitly assuming that the quantities whose averages matter are those whose averages behave in a "reasonable" fashion, and are not dominated by rarely-occurring extreme values. These quantities are called "self-averaging".

More precisely, a variable is self-averaging if the size of the fluctuations in its mean

value goes to 0 as $\mathcal{O}(1/\sqrt{N})$ or faster as $N \to \infty$. Self-averaging variables will therefore tend to be the *extensive* variables of the system, which scale as N; for a system such as a neural network, where the partition function Z scales roughly as e^N, the quantities that one would expect to be self-averaging are the ones related to $(\ln Z)$ rather than Z itself.

This creates a technical difficulty, in that, while in general Z is easy to average over a given probability distribution, the average of $(\ln Z)$ is harder to obtain. In performing this average we must take into account whether or not the two sets of disordered variables vary on different timescales. We consider a system with model Hamiltonian

$$H = \sum_{ij} J_{ij} \vec{S}_i \cdot \vec{S}_j, \tag{2.1}$$

where the $\vec{S}_i$'s are the instantaneous spins and the J_{ij}'s are the interaction strengths. If the two vary on the same time-scale the disorder is referred to as "annealed" and it is correct to take the *annealed average*,

$$\langle F \rangle = T \ln \langle Z \rangle, \tag{2.2}$$

where $\langle \ldots \rangle$ refers to the average over both sets of disordered variables. However, the features of neural networks and spin glasses that make them interesting arise because one set of variables – the spins – changes much faster than the other – the interactions. The interactions thus provide constraints on the dynamics of the spins. When not all of these constraints can be satisfied simultaneously the system is called "frustrated", and this frustration is the source of the deep behaviour displayed by neural networks. For this kind of system it is appropriate to perform the *quenched average*,

$$\langle F \rangle = T \langle \ln Z \rangle. \tag{2.3}$$

The standard method for doing this is known as the Replica Method or the "replica trick". It was introduced in this context by Edwards and Anderson [EA75], and is inspired

by the identity

$$x^n = e^{n \ln x} = 1 + n \ln x + \tfrac{1}{2}(n \ln x)^2 + \ldots,$$
$$\Rightarrow \ln x = \lim_{n \to 0} \tfrac{1}{n}(x^n - 1) \tag{2.4}$$

In order to average $(\ln Z)$, then, we find an expression for Z^n for integer values of n and make the analytic continuation $n \to 0$. Z^n is almost as easy to average as Z. For the case where the J_{ij}'s are the slow or "quenched" variables, we express Z^n as

$$Z^n = Tr_S \prod_\alpha e^{-\beta H_\alpha}$$
$$H_\alpha = \sum_{ij} J_{ij} \vec{S}_i^\alpha \cdot \vec{S}_j^\alpha. \tag{2.5}$$

This is known as the "replica method", because we can look on the integer powers of Z^n as representing the partition functions of n systems which are physically isolated but linked by all having the same realisation of the quenched variables. In this sense, the systems are "replicas" of each other.

We now develop the replica method in the context of a simple disordered magnetic spin system (a "spin glass"). Within this simple model we will be able to discuss problems which also arise when doing replica calculations for neural networks.

2.2 The Replica Method in the SK Spin Glass

We discuss the replica method in the context of the Sherrington–Kirkpatrick (SK) Spin Glass [SK75], which is mathematically one of the simplest interesting systems involving Ising spins and quenched disorder, and thus a useful training ground for the techniques we will later use in the context of neural networks. The SK model is a spin-glass model in which the N spins S_i take Ising values ± 1 which are updated in accordance with the stochastic rule

$$\begin{aligned} P(S_i(t+1)) &= \tfrac{1}{2}(1 + S_i(t+1)\tanh[\beta h_i]), \\ h_i &= \textstyle\sum_j J_{ij} S_j + h, \end{aligned} \tag{2.6}$$

where h is an external field and the pairwise connections J_{ij} are drawn from the Gaussian probability distribution

$$P(J_{ij}) = \left(\frac{N}{2\pi J^2}\right)^{1/2} \exp\left(-\frac{N(J_{ij} - J_0/N)^2}{2J^2}\right) \tag{2.7}$$

The J_{ij}'s are therefore independent of i, j. J_0 is a measure of the ferromagnetic order in the system. This is an infinite-range interaction, so mean-field theory is exact. The interactions are symmetric. Detailed balance will hold in equilibrium and the probability distribution will be given by the Gibbs measure $P(\vec{S}) \sim e^{-\beta H(\vec{S})}$.

The appropriate Hamiltonian for the system is

$$H = -\tfrac{1}{2}\sum_{ij} S_i J_{ij} S_j - h\sum_i S_i, \tag{2.8}$$

which is effectively a measure of the degree of frustration in the system. We now attempt to find an expression for $\langle Z^n \rangle$, where $\langle \ldots \rangle$ denotes an average over the probability

distribution (2.7). Performing this average, we obtain

$$\langle Z^n \rangle = Tr_S \exp\left[\frac{1}{N}\sum_{ij}\left(\frac{1}{4}(\beta J)^2 \sum_{\alpha\beta} S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta + \beta J_0 \sum_\alpha S_i^\alpha S_j^\alpha\right) + \beta h \sum_{i\alpha} S_i^\alpha\right]. \quad (2.9)$$

Extracting the terms with $\alpha = \beta$ in the first sum gives

$$\langle Z^n \rangle = Tr_S \exp\left[\frac{(\beta J)^2}{4N}\sum_{\alpha\neq\beta}\left((\sum_i S_i^\alpha S_i^\beta)^2 + nN^2\right) + \frac{\beta J_0}{2N}\sum_\alpha(\sum_i S_i^\alpha)^2 + \beta h \sum_{i\alpha} S_i^\alpha\right]. \quad (2.10)$$

If we can reduce the squared sums $(\sum_i S_i^\alpha S_i^\beta)^2$ and $(\sum_i S_i^\alpha)^2$ to simple sums, we will have obtained an expression with no coupling between sites. The means of doing this is the well-known Hubbard-Stratonovich transformation,

$$\int_{-\infty}^{\infty} dz \exp(-az^2 + bz) = \left(\frac{\pi}{a}\right)^{1/2} \exp\left(\frac{b^2}{4a}\right). \quad (2.11)$$

When we use the transformation, as we do here, to turn an expression that is quadratic in b to one which is linear in b, z is referred to as the variable *conjugate to* b. Introducing the variables $q_{\alpha\beta}$ and x_α as conjugates to $(\sum_i S_i^\alpha S_i^\beta)$ and $(\sum_i S_i^\alpha)$ respectively, we obtain

$$\begin{aligned}\langle Z^n \rangle \;=\; & \exp(nN(\tfrac{1}{2}\beta J)^2) \int \prod_{\alpha\neq\beta} \frac{\beta J N^{1/2}}{\sqrt{2\pi}} dq_{\alpha\beta} \prod_\alpha \left(\frac{\beta J_0 N}{2\pi}\right)^{1/2} dx_\alpha \times \\ & \exp\left(-\tfrac{1}{2}N(\beta J)^2 \sum_{\alpha\neq\beta}(q_{\alpha\beta})^2 - \tfrac{1}{2}N\beta J_0 \sum_\alpha (x_\alpha)^2\right) \times \\ & Tr_s \exp\left((\beta J)^2 \sum_{i,\alpha\neq\beta} q_{\alpha\beta} S_i^\alpha S_i^\beta + \beta \sum_{i\alpha}(J_0 x_\alpha + h) S_i^\alpha\right). \end{aligned} \quad (2.12)$$

Now there is no coupling across sites and (2.12) can be expressed in the form

$$\langle Z^n \rangle = \exp(nN(\tfrac{1}{2}\beta J)^2) \int \prod_{\alpha\neq\beta} \frac{\beta J N^{1/2}}{\sqrt{2\pi}} dq_{\alpha\beta} \prod_\alpha \left(\frac{\beta J_0 N}{2\pi}\right)^{1/2} dx_\alpha \exp(-NG), \quad (2.13)$$

where

$$G = \tfrac{1}{2}(\beta J)^2 \sum_{\alpha\neq\beta}(q_{\alpha\beta})^2 + \tfrac{1}{2}\beta J_0 \sum_{\alpha}(x_\alpha)^2 - \ln Tr_S \left((\beta J)^2 \sum_{i,\alpha\neq\beta} q_{\alpha\beta} S_i^\alpha S_i^\beta + \beta \sum_{i\alpha}(J_0 x_\alpha + h) S_i^\alpha \right) \tag{2.14}$$

In the limit $N \to \infty$ we can thus solve this equation by a method known as the *saddle-point method* or the *method of steepest descent.* Here we use

$$\int dy \exp(-NG(y)) = \int dy \exp(-N[G(y_0) + \tfrac{1}{2}G''(y_0)(y-y_0)^2 + \ldots]), \tag{2.15}$$

where y_0 is the saddle-point, so that $G'(y_0) = 0$. In the limit $N \to \infty$, this integral will be dominated by the value at y_0, so long as $G''(y_0)$ is positive. If $G''(y_0)$ is negative the integral will diverge.

Note that in using this method we are taking $N \to \infty$ before we take $n \to 0$. This appears to invalidate the identity that inspired the replica approach, and was originally suspected to be one of the reasons for the shortcomings of the theory. However, these shortcomings have been overcome in other ways and, in the absence of any other empirically successful alternative to the replica method, most physicists are willing to accept it.

We can thus describe the free energy per spin, $f = \lim_{N\to\infty} \frac{1}{N} \ln Z$, in terms of the saddle-point values $\tilde{q}_{\alpha\beta}, \tilde{x}_\alpha$ of the variables $q_{\alpha\beta}, x_\alpha$, as follows:

$$\begin{aligned} -\beta f \;=\; & \lim_{n\to 0}\Big[(\tfrac{1}{2}\beta J)^2 \left(1 - \frac{1}{n}\sum_{\alpha\neq\beta}(\tilde{q}_{\alpha\beta})^2 \right) - \frac{\beta J_0}{2n}\sum_{\alpha}(\tilde{x}_\alpha)^2 \\ & + \frac{1}{n}\ln Tr_S \exp\left(\tfrac{1}{2}(\beta J)^2 \sum_{\alpha\beta}\tilde{q}_{\alpha\beta} S^\alpha S^\beta + \beta \sum_{\alpha}(J_0\tilde{x}_\alpha + h) S^\alpha \right) \Big]. \end{aligned} \tag{2.16}$$

The saddle-point condition allows us to make the identifications

$$\begin{aligned} \tilde{q}_{\alpha\beta} &= \langle S^\alpha S^\beta \rangle, \\ \tilde{x}_\alpha &= \langle S^\alpha \rangle. \end{aligned} \tag{2.17}$$

So $\tilde{x}_\alpha$ is a measure of the magnetisation of the αth replicated system, and $\tilde{q}_{\alpha\beta}$ a measure of the similarity of two replicas which are both in the lowest-energy state. If we were to find $\tilde{q}_{\alpha\beta} = 1 \forall \alpha, \beta$, it would imply that there is only one ground state for the system.

2.2.1 The Replica-Symmetric Ansatz

In order to obtain an expression that can actually be evaluated, we must make an ansatz about the form of the $\tilde{q}_{\alpha\beta}, \tilde{x}_\alpha$ in (2.16). We expect the magnetisation to be the same in each replica, implying that $\tilde{x}_\alpha = M \forall \alpha$. For the $\tilde{q}_{\alpha\beta}$'s we make the simplest possible assumption, that of *replica symmetry* (RS): $\tilde{q}_{\alpha\beta} = q \forall \alpha, \beta$. We are therefore assuming that the overlap $\langle S^\alpha S^\beta \rangle$ between any pair of equilibrium states is the same as the overlap between any other pair. This is a reasonable assumption at high temperatures (where we expect this overlap to be zero) or where ferromagnetic order exists, but we have no reason other than convenience to expect it to be the case when the frozen disorder is the dominant aspect of the system.

Using these assumptions and the identity $\sum_{\alpha \neq \beta} 1 = n(n-1)$, and introducing a new conjugate variable z to separate the spins in the $S^\alpha S^\beta$ term, we obtain for the free energy per spin

$$\begin{aligned} -\beta f &= (\tfrac{1}{2}\beta J)^2 (1-q)^2 - \tfrac{1}{2}\beta J_0 M^2 + \int Dz \ln \cosh \eta(z) \\ \eta(z) &= \beta(J\sqrt{q}\, z + J_0 M + h), \end{aligned} \tag{2.18}$$

where $\int Dz \equiv \int dz e^{-z^2/2}/\sqrt{2\pi}$ and we have made the replacement

$$\lim_{n \to 0} \ln \int Dx \phi(x)^n \sim n \int Dx \ln \phi(x). \tag{2.19}$$

From this we can obtain equilibrium values for q and M.

$$\begin{aligned} q(T,h) &= \int Dz \tanh^2 \eta(z) \\ M(T,h) &= \int Dz \tanh \eta(z). \end{aligned} \tag{2.20}$$

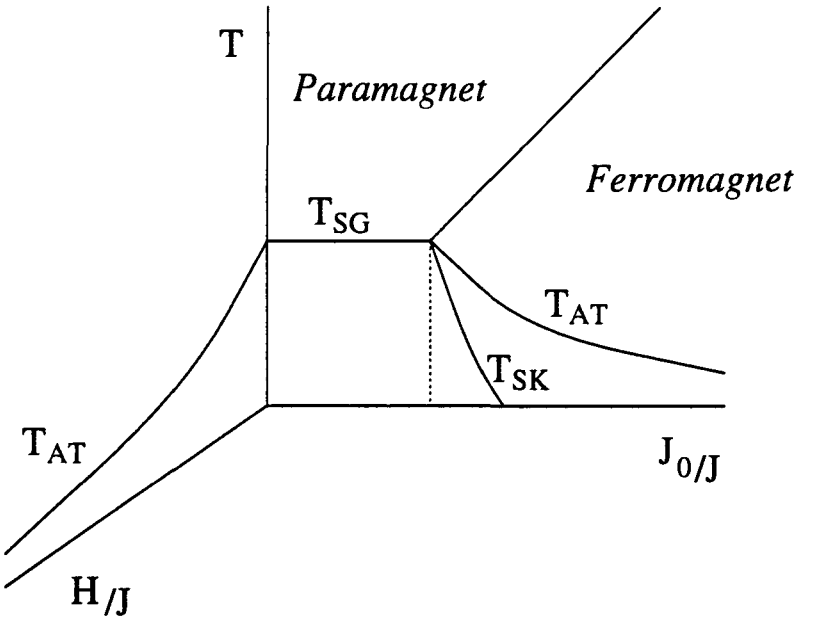


Figure 2.1: The qualitative phase diagram for the SK spin-glass. In the replica symmetric ansatz, the spinglass region is below the lines T_{AT} (on the left), T_{SG} and T_{SK}. In the full RSB solution the spinglass region is bordered on the right by the vertical dotted line rather than T_{SK}.

Ferromagnetic order exists when $M \neq 0$; in a spin glass, however, we expect to find states in which the individual spins are frozen but there is no long-range ferromagnetic order. This is indicated by $M = 0, q \neq 0$. Paramagnetic regions exist where both M and $q = 0$. The phase diagram of the SK model (shown in Figure 2.1) features all three of these regions. As might be expected, the spin glass region is to be found at low J_0, T, h.

2.2.2 Shortcomings of the Replica Symmetric Ansatz

The replica symmetric ansatz is straightforward and convenient, but it is unsatisfactory at low temperatures: the magnetic susceptibility calculated under this ansatz becomes negative below the lines T_{SG}, T_{AT}, and the entropy becomes negative in the limit $T \to 0$. These results are clearly unphysical, and if we look to the calculation for possible sources of error, we find three points where we have made possibly unjustified assumptions.

First, there is the assumption that the replica approach is valid at all; second, there is the point at which we take $N \to \infty$ before taking $n \to 0$; and third, there is the ansatz we make for the form of the saddle-point variables. The only assumption that it is possible to vary while keeping the broad form of the calculation is the last one. We will

proceed on the assumption that the replica method is correct but our use of the replica symmetric ansatz at low temperatures is wrong. This is equivalent to assuming that at low temperatures a more intricate form of spontaneous order arises in replica space as the number of replicas goes to zero.

2.2.3 Stability of the Replica Symmetric Ansatz

The replica-symmetric ansatz is an assumption about the form of the saddle-point variables that minimise the free energy. It is therefore only valid if the RS solution provides a global minimum. There are two different approaches to determining whether or not this is the case, which can be looked on as testing its global and local stability respectively.

Global Stability

In the first instance, we can make a different ansatz for the form of the solution, evaluate the free energy for this ansatz, and see if this yields a lower result. Although this approach is appealingly simple (apart from the problem of choosing an appropriate alternative ansatz), it suffers from two drawbacks.

Firstly, the meaning of "minimising" the saddle-point equations is not as straightforward as it seems. While it is true that it is necessary for the Hessian matrix in (2.15) to be positive definite, in the limit $n \to 0$ the dimensionality of this matrix $\frac{1}{2}n(n-1)$ becomes negative. We are therefore "minimising" the exponential in the saddle-point equations, but finding a *maximum* of the free energy. In the $J_0 = 0$ case it is possible to test the global stability of the RS solution by using another ansatz for the form of the $\tilde{q}_{\alpha\beta}$'s and maximising with respect to that. However, when $J_0 \neq 0$ we must simultaneously maximise with respect to the q's and minimise with respect to the x's.

Secondly, and more seriously, even if this subtlety is taken into account the approach has the drawback of all numerical approaches, namely that it runs the risk of revealing the answer to the specific question asked while concealing the structure of the problem. The global stability method is therefore better used as a way to check stability results

obtained by other methods than as a research tool in itself.

Local Stability

A more systematic approach is to perform the local stability calculation pioneered by de Almeida and Thouless (AT) [AT78]. As (2.15) demonstrates, the integral can only be solved by the saddle-point method if $G''(y_0)$ is positive. Since in this case we are performing a multi-dimensional integral, the quantities that must be positive are the eigenvalues of the Hessian matrix **A**:

$$\begin{aligned} A^{\alpha\beta,\gamma\delta} &= \frac{\partial^2 G}{\partial q_{\alpha\beta}\partial q_{\gamma\delta}} \\ &= \beta^2 J^2[\delta_{\alpha\beta,\gamma\delta} - \beta^2 J^2(\langle S^\alpha S^\beta S^\gamma S^\delta\rangle - \langle S^\alpha S^\beta\rangle\langle S^\gamma S^\delta\rangle)]. \end{aligned} \tag{2.21}$$

Note that we are implicitly taking $J_0 = 0$ in this calculation, and in the rest of the stability calculation to follow. This is partly for the reason given above – that we would be extremely surprised to find the optimal magnetisation varying from one replica to another – and partly because the slightly more laborious calculation taking $J_0 \neq 0$ demonstrates that the instability of the RS solution is due to fluctuations only in $A^{\alpha\beta,\gamma\delta}$. Thus, although the values of J, h at which the RS solution becomes unstable depend on J_0, the form of the calcuation does not.

There follows an outline of the AT calculation. The Hessian **A** has three different elements. The diagonal terms are given by:

$$(\beta J)^{-2} A^{\alpha\beta,\alpha\beta} = 1 - (\beta J)^2(1 - q^2). \tag{2.22}$$

The elements with one index in common between $\alpha\beta$ and $\gamma\delta$ are given by:

$$(\beta J)^{-2} A^{\alpha\beta,\alpha\delta} = -(\beta J)^2(q - q^2). \tag{2.23}$$

The remaining elements, with no indices in common, are:

$$(\beta J)^{-2} A^{\alpha\beta,\gamma\delta} = 1 - (\beta J)^2 (r - q^2), \tag{2.24}$$

where

$$r = \langle S^\alpha S^\beta S^\gamma S^\delta \rangle_{(\alpha \neq \beta \neq \gamma \neq \delta)} = \int Dz \tanh^4 \eta(z). \tag{2.25}$$

The eigenvectors also fall into three groups. First there is the fully symmetric eigenvector, whose eigenvalue when $n \to 0$ is

$$(\beta J)^{-2} \lambda_1 = 1 - (\beta J)^2 (1 - 4q + 3r). \tag{2.26}$$

The next group of eigenvectors, known as "anomalous modes", are those which are unchanged under exchange of all but one element. In the $n \to 0$ limit, these have the same eigenvalue as the symmetric eigenvector. The final group of eigenvectors, the "replicon modes", are unchanged under interchange of all but two elements and have eigenvalue

$$\begin{aligned} (\beta J)^{-2} \lambda_3 &= 1 - (\beta J)^2 (1 - 2q + r) \\ &= 1 - (\beta J)^2 \int Dz \ \mathrm{sech}^4 \eta(z). \end{aligned} \tag{2.27}$$

Of these eigenvalues, only λ_3 is ever negative. The area of the phase diagram in which it is negative is the area in which RS will be unstable.

Evaluating the condition for λ_3 to be positive derived from (2.27), we find this to be the same as the condition for the magnetic susceptibility to be positive. There is therefore a region of the phase diagram, shown in figure 2.1, in which the RS solution is locally unstable, and all of the unphysical features of the solution described in section 2.2.2 are contained within this region. This confirms the conjecture that our problems were due to our choice of the RS ansatz, and the interesting question is now to find an ansatz that is locally stable within the *replica-symmetry broken* (RSB) region. We explore this question in the next section.

2.3 Replica Symmetry Breaking

The hard-to-hyphenate phrase "replica symmetry breaking" covers a multitude of possibilities. In the initial rush to produce a workable ansatz for the form of the matrix $\mathbf{q}$, several possibilities were suggested. However, the symmetry breaking pattern put forward by Parisi [P80a] has gained widespread acceptance, not least because other patterns suggested ([B78], [BM78]) failed to satisfy one of the three requirements that Parisi suggested. These requirements are:

$$\begin{aligned} &\left|\lim_{n\to 0}\frac{1}{n}\sum_{\alpha\beta}q_{\alpha\beta}^2\right| < \infty \\ &\sum_{\alpha}q_{\alpha\beta} = \sum_{\alpha}q_{\alpha\gamma}, \quad \beta\neq\gamma \\ &\lim_{n\to 0}\frac{1}{n}\sum_{\alpha\beta}q_{\alpha\beta}^2 \leq 0 \end{aligned} \tag{2.28}$$

The first requirement comes from wanting the free energy to remain finite. The second preserves a reasonable kind of symmetry within the RSB scheme. The third is inspired by a physical feeling that at high temperatures the maximum of the free energy should be located at $q_{\alpha\beta}=0$. Parisi's scheme proceeds as follows.

For one-step replica symmetry breaking (1-step RSB or RSB1) we divide the $n\times n$ matrix $\mathbf{q}$ into submatrices of size $m\times m$, such that m divides n exactly. In the off-diagonal blocks we set $q_{\alpha\beta}=q_0$; in the diagonal blocks we set $q_{\alpha\beta}=q_1$. In the limit $n\to 0$, the requirement that m divides n simply becomes $0<m<1$. Before moving on to the full Parisi scheme, we outline the RSB1 solution to (2.16). For convenience we make the change in nótation $q_{\alpha\beta}\to q_{[\alpha_1,\alpha_2],[\beta_1,\beta_2]}$, where the first subscript of each pair denotes the block and the second denotes the replica index within the block. Using the RSB1 ansatz we obtain

$$\begin{aligned} -\beta f \;=\; &\lim_{n\to 0}\Big[(\tfrac{1}{2}\beta J)^2(1-(n-m)q_0^2-(m-1)q_1^2)-\tfrac{1}{2}\beta J_0 M^2 \\ &+\frac{1}{n}\ln Tr_S\exp\Bigg(\tfrac{1}{2}(\beta J)^2\Bigg[q_0(\sum_{\alpha_1\alpha_2}S_{\alpha_1\alpha_2})^2+(q_1-q_0)\sum_{\alpha_1}(\sum_{\alpha_2}S_{\alpha_1\alpha_2})^2-q_1\sum_{\alpha_1\alpha_2}S_{\alpha_1\alpha_2}^2\Bigg]\end{aligned}$$

$$+\beta(J_0+h)\sum_{\alpha_1\alpha_2} S_{\alpha_1\alpha_2}\Bigg) \tag{2.29}$$

Introducing z_0 and z_1 as conjugate variables to the two squared sums in the $\ln Tr_S$ terms enables us to obtain the solution

$$-\beta f = (\tfrac{1}{2}\beta J)^2(m(q_0^2-q_1^2)+(1-q_1)^2-\tfrac{1}{2}\beta J_0 M^2+\tfrac{1}{m}\int Dz_0 \ln \int Dz_1 \cosh^m \eta(z)$$
$$\eta(z) = \beta(J[\sqrt{q}\,z_0+\sqrt{q_1-q_0}\,z_1]+J_0M+h), \tag{2.30}$$

Further levels of RSB are constructed by an iterative process. At each step i, the diagonal blocks of size $m_{i-1}\times m_{i-1}$ are broken up into sub-blocks of size $m_i \times m_i$, the off-diagonal blocks are taken equal to q_{i-1} and the diagonal blocks are taken equal to q_i. In the limit $n \to 0$ the ordering $n > m_1 > m_2 > \ldots > 1$ is inverted to become $0 < m_1 < m_2 < \ldots < 1$. Although the structure of the resulting solutions to (2.16) is straightforward, it becomes extremely laborious to calculate the results explicitly if more than a small number of replica symmetry breaking steps have been taken. In the limit of full replica symmetry breaking, the m_i become a continuous variable x and we talk in terms of $q(x)$ rather than q_i. Parisi [P80b] has solved for $q(x)$ in the zero-temperature limit for the SK spinglass (and this result is directly transferable to certain neural network models), but in general the full RSB solution of any given problem is usually extremely technically difficult to obtain.

The structure of the successful RSB ansatz, shown schematically in Figure 2.3, is a very particular one. Two of its properties are especially worth commenting on: its *ultrametricity* and the fact that it is non-self-averaging.

A system is referred to as *ultrametric* if, given any three points within the system, the two greatest distances are equal. If, for any two replicas α, β, we define the distance as $(q_{\alpha\beta})^{-1}$, then the Parisi RSB scheme is ultrametric (it is necessary to take the inverse of $q_{\alpha\beta}$ because q is an overlap, not a distance). This suggests that there is a hierarchical valley structure to the n-dimensional energy landscape, with each valley (corresponding to one step of RSB) containing many sub-valleys, each of equal depth. In this case, of

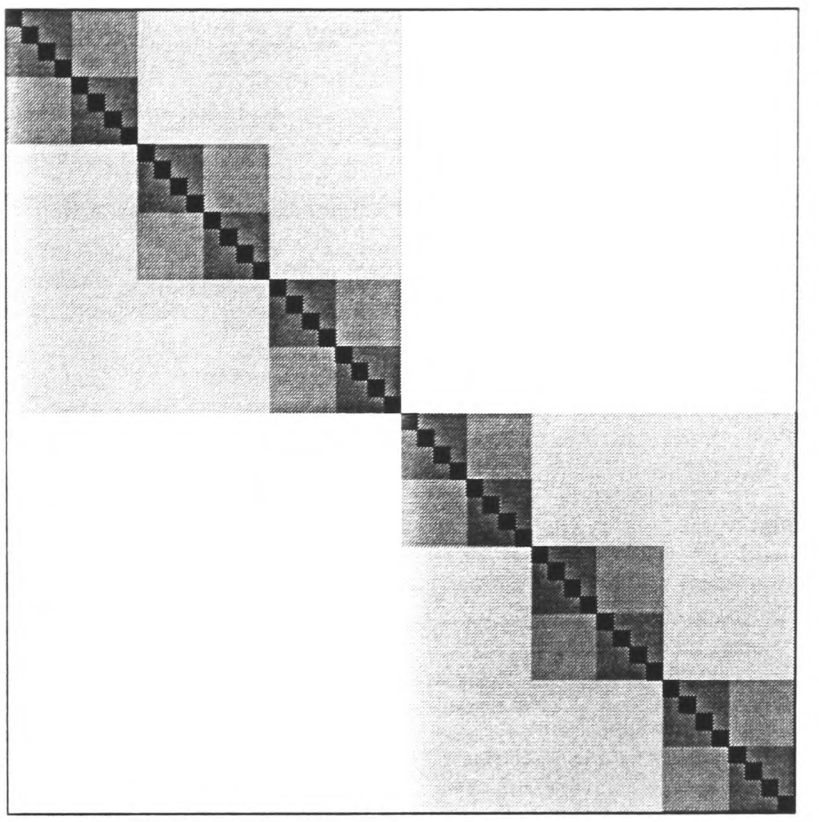


Figure 2.2: Schematic representation of 4-step replica symmetry breaking. The different shades of grey correspond to different values of q_i.

course, the ultrametricity is a form imposed by the ansatz and it is impossible to tell whether or not it reflects a genuine physical phenomenon; however, the success of this approach suggests that the ultrametricity may not be entirely artificial. Within the neural network context, ultrametricity offers aesthetically appealing analogies to the hierarchical structure of memories in the brain, though of course all such analogies must be treated with caution.

The scheme is non-self-averaging in that, while normal physical observables are still self-averaging, the distribution of the q's is not. While to an extent this is unexpected, we must bear in mind that the q's represent the overlap between two replicas in the limit of there being no replicas at all, and we should not be surprised if our intuition fails slightly in this case.

The Parisi model of RSB is the one we use throughout this thesis, but it is not the only one to have been successful. In the field of neural nets, for example, Tokita [T94] has obtained the full rsb solution for the Hopfield model [H82] using the De Dominicis and Kondor ansatz ([DGO81], [DGD82]). This model shares the attributes of being ultrametric and non-self-averaging, which suggests that these are requirements for any

successful RSB ansatz.

This concludes our introduction of the replica method and discussion of issues of replica symmetry in the context of the SK model. In the next chapter we discuss the application of the methods developed to the specific case of neural networks.

Chapter 3

Neural Networks as Associative Memories

In this chapter we discuss various aspects of the use of neural networks as associative memories. The bulk of the chapter is analytic and deals with issues arising from the use of the replica method in neural network calculations. For two different neural network models we perform analytic and numerical investigations of the degree of replica symmetry breaking required. We then investigate numerically some algorithms to realise one of the networks.

Historically, there have been two different approaches to the question of associativity. One can take a specific form for the synaptic connections and investigate its behaviour (*retrieval mode* studies); or one can take a task that needs to be performed (usually defined in terms of pattern bits) and attempt to find the J_{ij}'s that best perform it (*training mode* studies). There have been recent, dynamical, studies of networks in which both the J's and the S's are allowed to vary [PCS93], but even these fit broadly into the category of retrieval mode studies. In both the retrieval mode and the training mode approaches we commonly use the replica method to obtain results, but the choices for fast and slow variables are different: in the first case, the S's are the fast variables and the J's are held constant, while in the second the J's are the fast variables in an enviroment defined by the patterns ξ^{μ}. Since retrieval mode studies are a major part of neural network

research (making up, for example, the entire second part of this thesis), we first discuss an archetypal example of them. We then turn to the question of training mode studies which will occupy us for the bulk of this chapter.

3.1 Hopfield Model

As mentioned, studies of retrieval properties of neural networks can be divided between the categories of retrieval mode and training mode studies. The Hopfield model [H82] is the archetype for the field of retrieval modes and has formed a basis for many more realistic, or at least more complicated, models. As well as being analytically tractable, it has the advantage in practical terms of being a "one-step" learning rule which can be implemented quickly and easily.

In the basic Hopfield model, we take the pattern bits ξ_i^μ to be ± 1 with equal probability and take the synaptic matrix J_{ij} to be given by the Hebb rule [H49],

$$J_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu. \tag{3.1}$$

The model employs sequential dynamics, although the parallel dynamics calculation is a straightforward development from it [FK88].

This system has several useful properties. First, a simple signal-to-noise analysis of the local fields at each site shows that, for small numbers of patterns, the prescription (3.1) will indeed store the patterns as fixed-points of the dynamics. Second, the synaptic matrix is symmetric and detailed balance applies; therefore it is susceptible to the techniques of equilibrium thermodynamics. In the thermodynamic limit the equilibrium distribution is given by $P(\{S\}) = \epsilon^{-\beta H(\{S\})}/Z$, with H given by

$$H_{\text{Hopfield}} = \sum_{ij} S_i J_{ij} S_j + \sum_i h_i S_i. \tag{3.2}$$

Using the explicit expression (3.1) for the J_{ij}'s we obtain for the dynamics:

$$P(S_i[t+1]) = \tfrac{1}{2}(1 + S_i[t+1]\tanh[\beta \sum_{\mu} \xi_i^{\mu} q^{\mu}[t]]) \tag{3.3}$$

and for the Hamiltonian

$$H_{\text{Hopfield}} = N \sum_{\mu} (q_{\mu})^2 + \sum_{i} h_i S_i. \tag{3.4}$$

This demonstrates explicitly a third useful property of this system, namely that for the purposes of simulations it is not necessary to store the entire synaptic matrix. The "separable" nature of the dynamics means that it is possible to simulate the dynamics of a system of size N using only $\mathcal{O}(N)$ variables rather than $\mathcal{O}(N^2)$.

For the moment, however, we concentrate on the equilibrium properties of the network. In a calculation that has now become standard, Amit *et al* [AGS85] have calculated the free energy of the Hopfield model using the replica formalism in the RS ansatz. This calculation makes possible the evaluation of the *critical storage capacity* α_c, which is the storage level $\alpha = p/N$ at which it is no longer possible for the network to recall the stored patterns. The only major innovation in their calculation, compared to the SK calculation of the previous chapter, is the use of the "condensed ansatz" in which, following the average over the quenched disorder, we assume that the system has overlaps of $\mathcal{O}(1/\sqrt{N})$ with all but a small number of the stored patterns. This enables us to treat the $\mathcal{O}(1)$ overlaps with the "condensed" patterns as variables evolving in an environment in which, in addition to the noise due to the operating temperature of the system, additional "intrinsic noise" is contributed by the uncondensed patterns.

The RS expression for the free energy is

$$\begin{aligned} f &= \tfrac{1}{2}m^2 + \frac{\alpha}{2}\left[1 + \beta r(1-q) - \frac{q}{1-\beta(1-q)} + \frac{1}{\beta}\ln(1-\beta(1-q))\right] \\ &\quad -T\int Dz \langle \ln(2\cosh\beta[z\sqrt{\alpha r} + m\xi]), \end{aligned} \tag{3.5}$$

where the average is to be taken over the remaining condensed pattern bit ξ and the saddle-point conditions give the following values for m, r, q:

$$\begin{aligned} m = \quad \langle \xi S \rangle \quad &= \int Dz \langle \xi \tanh[\beta(z\sqrt{\alpha r} + m\xi)] \rangle \\ q = \quad \langle SS \rangle \quad &= \int Dz \langle \xi \tanh^2[\beta(z\sqrt{\alpha r} + m\xi)] \rangle \\ r = \quad \textstyle\sum_{\mu>1}(q_\mu)^2 \quad &= q[1-\beta(1-q)]^{-2}. \end{aligned} \tag{3.6}$$

We see that m, the pattern overlap, effectively obeys the familiar ferromagnetic equation $m = \tanh[\beta m]$, except for the inclusion of a Gaussian intrinsic noise term of width αr.

The calculation just outlined gives the result that the critical storage capacity for pattern recall $\alpha_c = 0.138$ at zero temperature. As α increases past α_c, the saddle-point value of m drops abruptly from approximately 0.97 to 0; this is known as the "memory blackout catastrophe". As T increases the saddle point value of m decreases, going continuously to 0 for $\alpha = 0$ at $T = 1$.

Use of the fully-connected Hopfield model as an associative memory is complicated by the existence of "spurious states", which are fixed-points of the dynamics other than the patterns. Some of the spurious states are easy to construct by hand: for example, it can easily be seen that we obtain a state that is stable at zero operating temperatures by considering any three patterns and taking S_i on each site to have the value determined by "majority vote" of the pattern bits at this site. As the operating temperature is raised, however, the associativity of these states is reduced, until for $T > 0.46$ the only stable attractors are the stored patterns. Increasing the temperature (or noise) in the system is thus actually beneficial to the retrieval performance, in terms of widening the basins of attraction.

3.1.1 Replica Symmetry Breaking in the Hopfield Model

A straightforward calculation of the stability of the replica-symmetric solution shows it to be unstable in the low T region, requiring replica symmetry breaking. The one-step

RSB calculation for the Hopfield model was carried out by Crisanti *et al* [CAG86], yielding an improved result for the maximum storage capacity at zero temperature of $\alpha_c = 0.144$, though this result has recently been questioned [C95]. More recently, Tokita [T94] has found the full replica-symmetry broken solution for the Hopfield model at zero operating temperature. In this solution $\alpha_c = 0.155 \pm 0.002$, which is slightly higher than results obtained by numerical investigation. However, the effects of replica-symmetry breaking are still much smaller than in the SK spin-glass; in general any significant effects are confined to the region near $T = 0, \alpha = \alpha_c$, and in this region the maximum value of $\Delta q \equiv |q_{\mathrm{rsb}} - q_{\mathrm{rs}}|$ for any RSB scheme was $\Delta q = 0.01, q_{\mathrm{rs}} = 1$.

3.1.2 Basins of Attraction and Dynamical Studies

The distribution of local stabilities $\rho(\lambda)$ for the Hopfield model is given by [KA88]:

$$\rho(\lambda) = e^{-\frac{1}{2}(\lambda - \frac{1}{\sqrt{\alpha}})^2}. \tag{3.7}$$

This is a Gaussian centred at $1/\sqrt{\alpha}$. For the highly dilute Hopfield network, we can use this in (1.8) to determine the dynamics. Surprisingly, we find that, although the storage capacity of the fully-connected network is $\alpha_c \sim 0.15$, the first-step dynamics show an initial increase in the overlap for $\alpha < 2/\pi$. For the highly dilute network the storage capacity is thus $\alpha_c = 2/\pi \sim 0.64$. ([DGZ87], [H&a89]), and this proves to give wide retrieval for all $\alpha < \alpha_c$. This is comparable to the theoretical maximum storage capacity (calculated in the following section) of $\alpha_{max} = 2.0$. Admittedly, in the highly dilute network even $0.64C$ isn't a lot of patterns, but we can take advantage of this in the fully-connected network too. We find that if the inital stimulus persists, in the form of an external field applied to the neurons as an extra input, then recall is possible for a stored pattern for α up to this value. (Intriguingly, an external field parallel to the stimulus is more successful in encouraging recall than an external field parallel to any of the stored patterns [EES89]). A diagram of the basin of attraction of the dilute Hopfield model is given in Figure 3.14.

Although there is a rich variety of interesting and useful questions that can be investigated regarding networks with fixed synaptic prescriptions, these networks will always remain the results of somewhat arbitrary choice unless we have a guide to how their properties compare to the properties of some optimal network. Fortunately, the techniques of equilibrium statistical mechanics, which we have just seen applied to the spin-states space of specified neural networks, can also be applied to the configuration space of all neural networks, thus determining their optimal properties. This is the approach we will pursue in the next section.

3.2 Theory of Optimized Networks

The late Elizabeth Gardner has established a general framework for determining the properties of a neural network optimising any cost function of the form $E = \sum_\mu g(\lambda^\mu)$. We here outline the replica calculation for a general cost function, and discuss issues of replica symmetry breaking for two specific cost functions: the Gardner-Derrida cost function, which produces optimal storage of noiseless patterns, and the cost function for "training with noise", which produces networks with the optimal retrieval behaviour in a noisy operating environment.

In this case we abandon the symmetry of the synaptic matrix, and consider optimising the input synapses on each neuron independently. The sites thus decouple and we can consider the network to be a collection of connected, but independently optimised, perceptrons. Our calculations are thus expressed only in terms of the number of inputs to the perceptron and can be applied with equal validity to the fully-connected or the extremely dilute network.

The system we are interested in here is a perceptron with N input nodes (labelled $i = 1, \ldots, N$) whose output node obeys the deterministic update rule

$$S_{output} = \text{sign}\left(\sum_i J_i S_i\right), \quad S_i \in \{\pm 1\}, \tag{3.8}$$

with the $\{J_i\}$ constrained by the spherical rule $\sum_i J_i^2 = N$. This perceptron is trained to optimise some measure of the performance of the perceptron on a set of $p = \alpha N$ patterns. We formulate the problem as "energy" minimisation, defining the energy as

$$E = -\sum_\mu g(\lambda^\mu), \tag{3.9}$$

where λ^μ is the local aligning field defined in the introduction (1.6), $\lambda^\mu = \xi_o^\mu \sum_i J_i \xi_i^\mu / |J|$.

Our results are initially presented in terms of the general cost function $g(\lambda)$. Typical choices of a specific cost function include the so-called *Gardner-Derrida* cost function

[GD88]

$$g(\lambda) = \theta(\kappa - \lambda), \tag{3.10}$$

the *Perceptron* cost function [GG91]

$$g(\lambda) = (\kappa - \lambda)\theta(\kappa - \lambda) \tag{3.11}$$

and the *Adaline* cost function ([WH60], [DO87])

$$g(\lambda) = (\kappa - \lambda). \tag{3.12}$$

Each of these cost functions results in a slightly different property of the system being optimised; crudely speaking, the Gardner-Derrida cost function attempts to store as many patterns as possible, the Perceptron cost function does the same but attempts to have the unstored patterns as near stability as possible, and the Adaline will, if κ is correctly chosen, reproduce the pseudo-inverse rule [PGD85], with all patterns being stored with exactly the same stability.

Returning to the general case, we calculate the minimised energy $\tilde{E}$ by minimising the free energy of the system,

$$\begin{aligned} f(\{\xi\}) &= -\lim_{N\to\infty} \frac{1}{N\beta} \langle \ln Z \rangle \\ &= -\lim_{N\to\infty} \frac{1}{N\beta} \langle \ln \int \prod_{i=1}^{N} dJ_i \delta(\vec{J}^2 - N) e^{-\beta E} \rangle, \end{aligned} \tag{3.13}$$

and finally taking the limit $\beta \to \infty$, in which $E = Nf$ (for the Gardner-Derrida cost function, the free energy in this limit gives the fraction of patterns that are stored incorrectly, and we thus identify it with the output error of the perceptron). This free energy is assumed to be self-averaging over the disorder in the patterns $\{\xi^\mu\}$. In order to perform the average, we employ the replica trick $\langle \ln Z \rangle = \lim_{n\to 0}(\langle Z^n \rangle - 1)/n$.

In outline, the calculation of $\langle Z^n \rangle$ proceeds as follows ([GD88], [WS90b]). We introduce a delta-function to enforce the definition of λ^μ (1.6), and express both the delta func-

tions that now figure in the equation in terms of the identity $\delta(x-y) = \int dz e^{iz(x-y)}/(2\pi)$. This gives

$$\begin{aligned}\langle Z^n \rangle &= \int \prod_{j\alpha} dJ_j^\alpha \prod_\alpha E_\alpha \prod_{\mu\alpha} d\lambda_\alpha^\mu \prod_{\mu\alpha} dx_\mu^\alpha \times \\ &\exp\left[i \sum_\alpha E^\alpha (\sum_i (J_i^\alpha)^2 - N) + \beta \sum_{\mu\alpha} g(\lambda_\alpha^\mu) + i \sum_{\mu\alpha} x_\mu^\alpha (\lambda_\alpha^\mu - \frac{\xi_o^\mu}{\sqrt{N}} \sum_j J_j^\alpha \xi_j^\mu) \right],\end{aligned} \tag{3.14}$$

where we have ignored multiplicative constants which will disappear in the $N \to \infty, n \to 0$ limits.

Performing the average over the stored patterns gives a term of the form

$$\exp\left(\sum_{j\mu} \ln\cos(\sum_\alpha x_{\alpha\mu} J_j^\alpha / \sqrt{N}) \right) \sim \exp\left(-\frac{1}{2N} \sum_{j\mu\alpha\beta} x_{\alpha\mu} x_{\beta\mu} J_j^\alpha J_j^\beta \right), \tag{3.15}$$

using the identity $\cos(x) \sim 1 - x^2/2$ for small x. We next introduce

$$q_{\alpha\beta} \equiv \frac{1}{N} \sum_i J_i^\alpha J_i^\beta \tag{3.16}$$

by means of yet another delta-function (with conjugate variable taken to be $iF_{\alpha\beta}$) and, factorising over the site and pattern indices as is clearly possible, we obtain

$$\begin{aligned}\langle Z^n \rangle &= \int \prod_{\alpha\beta} dq_{\alpha\beta} e^{N\Phi(q_{\alpha\beta})} \\ &= \int \prod_{\alpha\beta} dq_{\alpha\beta} \prod_{\alpha\beta} dF_{\alpha\beta} \prod_\alpha dE_\alpha \times \\ &\exp\left[N \left(\alpha G_\Lambda(\{q_{\alpha\beta}\}) + G_J(\{E_\alpha\}, \{F_{\alpha\beta}\}) - \sum_{\alpha<\beta} F_{\alpha\beta} q_{\alpha\beta} \right) \right],\end{aligned} \tag{3.17}$$

where

$$\exp[\alpha G_\Lambda(\{q_{\alpha\beta}\})] \equiv \int \prod_\alpha d\lambda_\alpha \prod_\alpha dx_\alpha \times$$

$$\exp\left[\sum_{\alpha}(\beta g(\lambda_\alpha) + ix_\alpha\lambda_\alpha - \tfrac{1}{2}x_\alpha^2) - \sum_{\alpha\beta} q_{\alpha\beta}x_\alpha x_\beta\right],$$

$$\exp[G_J(\{E_\alpha\},\{F_{\alpha\beta}\})] \equiv \int \prod_\alpha J_\alpha \exp\left[\sum_\alpha E_\alpha(1-J_\alpha^2) + \sum_{\alpha\beta} F_{\alpha\beta}J_\alpha J_\beta\right]. \quad (3.18)$$

Now $q_{\alpha\beta}$ is a measure of the similarity of the synapses in two replicated networks that both minimise the free energy.

We solve equation (3.17) by the saddle-point method in the limit $N \to \infty$. Note that our use of $iF_{\alpha\beta}$ as the conjugate variable ensures that we are actually seeking a minimum in the $n \to 0$ limit, in contrast to the case of the SK model.

Replica-Symmetric Theory

As before, we must make an ansatz about the form of the replica solution and, as before, we start by making the RS ansatz $q_{\alpha\beta} = q$, $F_{\alpha\beta} = F$, $E_\alpha = E$. We can thus perform a Hubbard-Stratonovich transformation to separate the $\sum_{\alpha\beta} x_\alpha x_\beta$ sum, and obtain in the limit $N \to \infty, n \to 0$:

$$\begin{aligned}\frac{\ln Z}{N} &= \mathrm{extr}_q\Big[\tfrac{1}{2}\ln(2\pi(1-q)) + \frac{1}{2(1-q)} \\ &\quad + \alpha\int Dt \ln\int \frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left[\beta g(\lambda) - \frac{(\lambda-\sqrt{q}t)^2}{2(1-q)}\right]\Big]. \end{aligned} \quad (3.19)$$

The first two terms are obtained by direct evaluation of the saddle-point conditions for F, E.

We make the assumption that there is only one network that will best satisfy the constraints imposed by the need to minimise the cost function. This implies that $q \to 1$ and enables us to impose a sensible scaling in the innermost square brackets by making the ansatz that as $\beta \to \infty$, $(1-q) \sim 1/\beta$. With this ansatz, the inner integral can be solved by the saddle-point method in the limit $\beta \to \infty$, obtaining

$$f = \lim_{\beta\to\infty}\frac{1}{\beta\alpha C}\langle \ln Z\rangle = \min_\gamma \int Dt\left[g(\lambda_0) - \frac{1}{2\gamma}(\lambda_0 - t)^2\right] + \frac{1}{2\alpha\gamma}. \quad (3.20)$$

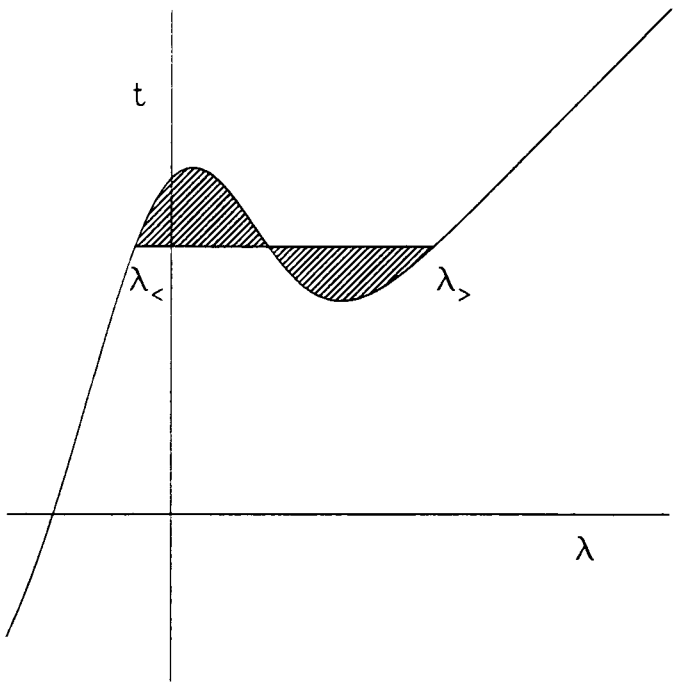


Figure 3.1: Illustration of the Maxwell's construction to obtain $\lambda(t)$ from $t(\lambda)$. In order for the function $\lambda(t)$ to be single-valued, we introduction a discontinuity in λ from $\lambda_<$ to $\lambda_>$, such that $t(\lambda_<) = t(\lambda_>)$ and the two shaded areas are equal.

Here $\gamma \equiv \beta(1-q)$ and λ_0 minimises

$$\beta g(\lambda) - \frac{(\lambda - \sqrt{q}t)^2}{2(1-q)}. \tag{3.21}$$

The integral over t thus requires us to invert the function $t(\lambda) = \gamma g'(\lambda) + \lambda$. If at any point this inverse function $\lambda(t)$ is multiple valued, we must perform a Maxwell's construction [WS90a] (see Figure 3.1) to determine which of the values for λ we shall accept. This will lead to a discontinuity in $\lambda(t)$ which is reflected in, for example, the distribution of local stabilities $\rho(\lambda)$. This quite simple procedure (it certainly strains credulity less than many other things connected with the replica method) proves to have profound effects on considerations of replica stability.

We can also calculate the local stability distribution $\rho(\lambda)$ which, as we have already said, is the relative volume of solution space with aligning field λ:

$$\rho(\lambda) = \lim_{\beta\to\infty} \frac{1}{Z} \int \prod_{i=1}^{N} dJ_i \delta((\mathbf{J})^2 - N) e^{-\beta \sum_\mu g(\lambda^\mu)} \delta(\lambda - \lambda^\nu)$$

$$= \int Dt\, \delta(\Lambda - \lambda(t)). \tag{3.22}$$

The saddle-point condition for γ is

$$\int Dt(\lambda(t) - t)^2 = \alpha^{-1}. \tag{3.23}$$

We can therefore express f in the form

$$f = \int d\lambda\, \rho(\lambda)\, g(\lambda). \tag{3.24}$$

We round off this survey of the RS result by evaluating the AT condition for local stability of this solution. In this case we must evaluate the replicon-like eigenvalues of the Hessian generated by differentiating Φ (3.17) with respect to both the q's and the F's. Inspection of (3.17) shows that the off-diagonal block elements $\partial^2\Phi/\partial F_{\alpha\beta}\partial q_{\gamma\delta} = 1$. It therefore only remains to calculate the eigenvalues of the blocks on the diagonal, and this calculation proceeds along much the same lines as for the SK spin glass. As before, each block has three different elements. For the q-block they are

$$\begin{aligned} P = \quad & \frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}^2} \quad = \langle\langle x^2\rangle_x^2\rangle_t - \langle\langle x^2\rangle_x\rangle_t^2 \\ Q = \quad & \frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}\partial q_{\alpha\gamma}} \quad = \langle\langle x^2\rangle_x\rangle_t\langle\langle x\rangle_x^2\rangle_t - \langle\langle x^2\rangle_x\rangle_t^2 \\ R = \quad & \frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}\partial q_{\gamma\delta}} \quad = \langle\langle x\rangle_x^4\rangle_t - \langle\langle x^2\rangle_x\rangle_t^2, \end{aligned} \tag{3.25}$$

where

$$\langle f(x)\rangle_x = \frac{\int dx d\lambda f(x) e^{\beta g(\lambda) + ix(\lambda - \sqrt{q}\,t - (1-q)x^2/2}}{\int dx d\lambda e^{\beta g(\lambda) + ix(\lambda - \sqrt{q}\,t - (1-q)x^2/2}} \tag{3.26}$$

$$\langle f(t)\rangle_t = \int Dt\, f(t) \tag{3.27}$$

(the F-block results are similar). The eigenvalue that governs RSB is the replicon-like

$$\gamma_3 = P - 2Q + R \tag{3.28}$$

and the condition for stability is that the eigenvalues of the matrix

$$\begin{pmatrix} \alpha\gamma_q & 1 \\ 1 & \gamma_F \end{pmatrix} \tag{3.29}$$

must have the same sign as for the known stable solution $\alpha \to 0$. Our stability condition thus turns out to be

$$\alpha\gamma_q\gamma_J < 1 \quad \Rightarrow \quad \alpha \int Dt[1-\lambda'(t)]^2 < 1. \tag{3.30}$$

1-step RSB

We now introduce the RSB1 ansatz defined in chapter 2, with $q_{\alpha\beta} = q_0$ in the diagonal blocks and q_1 in the off-diagonal blocks. The averaged minimum energy for any cost function $g(\lambda)$ in the limit $\beta \to \infty, q_0 \to 1$ is [MEZ93]:

$$\begin{aligned} e = \frac{\langle E_{min}\rangle}{N} &= \lim_{\beta\to\infty}\langle f\rangle = \lim_{\beta\to\infty}\min_{\gamma,q_1,w}\left[\frac{q_1}{2\gamma(1+w\Delta q)} + \frac{\ln(1+w\Delta q)}{2w\gamma} + \right. \\ &\left. \frac{\alpha}{w\gamma}\int Dz_1 \ln\int Dz_0 \exp\left(-w\gamma\left[g(\lambda_0) + \frac{(\lambda_0 - z_1\sqrt{q_1} - z_0\sqrt{\Delta q})^2}{2\gamma}\right]\right)\right] \end{aligned} \tag{3.31}$$

where $w = \beta m/\gamma$ and λ_0 minimises the final square bracket for given values of z_0, z_1, q_0, $\Delta q \equiv (1-q_0)$ and $\gamma \equiv \beta(1-q_1)$. As before, we are required to perform a Maxwell's construction to obtain $z_0(\lambda)$ in the innermost integral. This having been done, we solve equation (3.31) by minimising it numerically with respect to γ, q_0, w.

The RSB1 solution for the aligning field distribution is

$$\rho(\lambda) = \int Dz_0 \frac{\int Dz_1 \exp\left(-w\gamma\left[g(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right)\delta(\lambda-\lambda_0)}{\int Dz_1 \exp\left(-w\gamma\left[g(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right)}. \tag{3.32}$$

The values of the parameters are those obtained by minimisation of (3.31).

Summary

The result here obtained has been for any arbitrary cost function of the form $E = \sum_\mu g(\lambda_\mu)$, and for a fully connected network. This calculation may be varied in many different ways, for example by looking at biased patterns [G88] or by randomly diluting the network before [BEKS90] or after ([BEKS90], [H91]) training. The general method is also applicable to different forms of the couplings, such as binary couplings [KM89] or couplings that interpolate between the spherical case and the binary case [PS94a]. The simplest next step, however, and the one we restrict ourselves to, is to take a specific choice for the cost function. In the following two sections we investigate issues of RSB for neural networks that have been optimised with respect to two specific cost functions. The first, the Gardner-Derrida cost function, gives the optimal results for storage of noiseless patterns; the second, the noise-optimal cost function, gives the maximum first-step overlap increase for a network designed to operate in the presence of retrieval noise.

3.3 Replica Symmetry Breaking in the Gardner-Derrida Perceptron

The Gardner-Derrida perceptron is defined by the cost function

$$g(\lambda) \equiv \theta(\kappa - \lambda). \tag{3.33}$$

In other words, for some κ, usually taken to be ≥ 0, a pattern $\vec{\xi}^{\mu}$ is considered to be stored correctly if its aligning field $\lambda^{\mu} > \kappa$. The cost function counts the number of patterns that are not correctly stored. The minimised energy therefore gives the minimum possible number of patterns that can be stored incorrectly; this is zero if α is less than a critical storage capacity $\alpha_c(\kappa)$. In the special case where $\alpha = \alpha_c(\kappa)$ then the noiseless patterns are stored as well as is possible, and the perceptron is referred to as the *perceptron of maximal stability*. For $\kappa = 0$, the maximum storage capacity is found to be $\alpha = 2$, though for the fully-connected network useful memory operation is only observed for values of α up to $\alpha \sim 0.4$ [F88].

The replica symmetric solution has been shown ([B94],[MEZ93],[ET93]) to be the correct solution to the saddle point equations for $\alpha \leq \alpha_c$; however, for the perceptron above saturation a different ansatz must be used. A first step towards finding the correct structure for the solution space of these networks has been taken by the numerical studies of ([MEZ93],[ET93]), which show that RSB occurs whenever the perceptron is above saturation, and that the RS solution provides a global minimum whenever it is locally stable. The analysis turns on the point that there is a discontinuity in λ for the replica-symmetric solution, because, as mentioned previously, we have had to perform a Maxwell's construction. This discontinuity persists in the RSB1 case, and is reflected in a gap in $\rho(\lambda)$. We display a plot of $\rho(\lambda)$ for $\alpha = 1, \kappa = 1$ in the RSB1 ansatz for illustrative purposes (Figure 3.2).

We now take the work of these studies a step further by investigating the stability of the RSB1 solution, analytically and numerically; the analysis suggests strongly that full

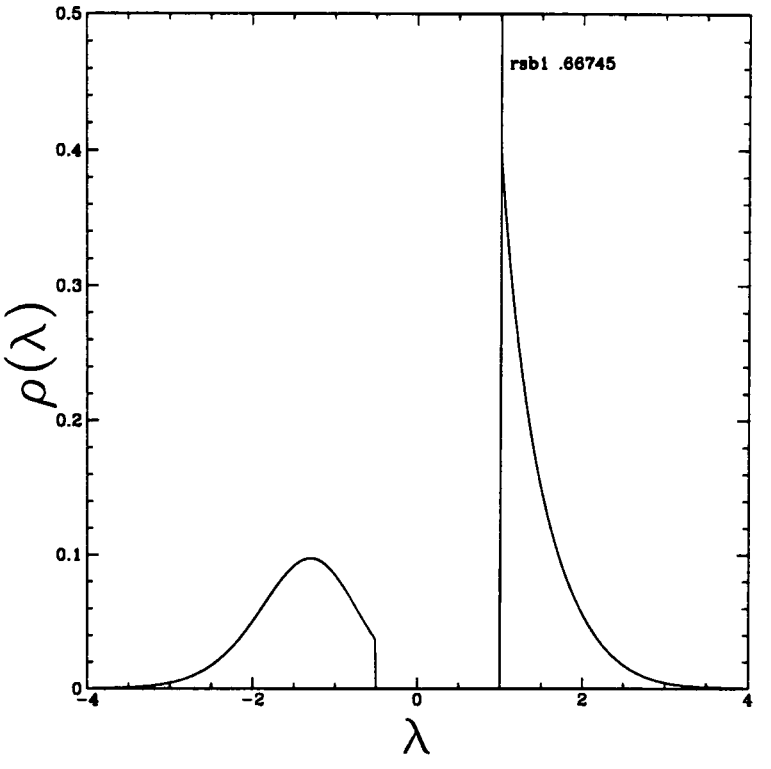


Figure 3.2: The local stability distribution $\rho(\lambda)$ in the RSB1 ansatz for $\alpha = 1$, $\kappa = 1$, showing the gap in the stabilities. There is a delta-function peak of weight 0.67 at $\lambda = 1$.

RSB is necessary whenever the perceptron is above saturation.

3.3.1 Stability of the 1-step RSB Solution

We first look at the fluctuations of the RSB1 solution (3.31) around its saddle-point values. As previously discussed, a change in the local stability of the RSB1 solution is indicated by a change in the sign of one of the eigenvalues of the matrix of the quadratic fluctuations in the saddle-point equation (2.15) ([GD88], [AT78]).

We define the matrix of fluctuations **H** as follows. The submatrices $H^{qq}_{(\alpha\beta)(\gamma\delta)}$ and $H^{Eq}_{\alpha(\beta\gamma)}$ are defined by

$$H^{qq}_{(\alpha\beta)(\gamma\delta)} \equiv \frac{\partial^2 \Phi}{\partial q_{\alpha\beta} \partial q_{\gamma\delta}}, H^{Eq}_{\alpha(\beta\gamma)} \equiv \frac{\partial^2 \Phi}{\partial E_\alpha \partial q_{\beta\gamma}}. \tag{3.34}$$

The other submatrices, $H^{FF}_{(\alpha\beta)(\gamma\delta)}$, $H^{qF}_{(\alpha\beta)(\gamma\delta)}$, $H^{FE}_{(\alpha\beta)\gamma}$, $H^{EE}_{\alpha\beta}$, etc, are defined analogously. The evaluation of the eigenvalues is complicated; however, we can make use of several simplifying arguments, used originally by Dorotheyev [D92] for investigating the stability

of the RSB1 solution for a pseudo-inverse synaptic matrix. First, we observe that the submatrices are invariant under the action of the Heirarchical Tree (HT) group. This group is defined in Appendix 1 of [D92] and in [MPV87]. We may represent it as

$$S_{n/m} \hat{\oslash} (S_m)^{\otimes n/m}, \tag{3.35}$$

with $(S_m)^{\otimes n/m}$ being the direct product of the k-element permutation group with itself n/m times, and $\hat{\otimes}$ being the semidirect product. We can thus express the eigenvectors in terms of the bases of the irreducible representations of the HT group. We further reduce the number of calculations necessary by using the fact that instability only arises in the direction of the replicon-like eigenvectors. In RSB1 there are four families of these eigenvectors, which we call $R^{(a)}$, $R_1^{(e)}$, $R_2^{(e)}$, $R_3^{(e)}$; their definition and the calculation of the corresponding eigenvalues $\gamma_q^{R^{(a)}}, \gamma_F^{R^{(a)}}$, etc, are given in the Appendix to this section. In order for the RSB1 solution to be stable, the sign of each eigenvalue of the matrix

$$\begin{pmatrix} \alpha\gamma_q^R & 1 \\ 1 & \gamma_F^R \end{pmatrix} \tag{3.36}$$

must be the same as its sign for a known stable solution, for any $R \in \{R^{(a)}, R_1^{(e)}, R_2^{(e)}, R_3^{(e)}\}$. We know that the RSB1 solution is stable in the limit $\alpha \to 0$, in which case all the eigenvalues in question are -1. We therefore need to find at what point any eigenvalue changes sign, or in other words where

$$\alpha\gamma_q^R\gamma_F^R \geq 1. \tag{3.37}$$

We now show that this occurs at the critical storage capacity. Our proof for RSB1 parallels that of Bouten [B94] for the instability of the RS solution.

We consider the family $R^{(a)}$, whose eigenvalues, following the notation used in the

Appendix, can be written as

$$\begin{aligned}\gamma^{R^{(a)}} &= K_1 - 2K_2 + K_3 \\ &= \left[\left[\left([f^2]_f - [f]_f^2\right)^2\right]_0\right]_1. \end{aligned} \tag{3.38}$$

Noting that

$$\frac{d}{dz_1}([1]_f^{m-1}[f]_f) = -i\sqrt{f_1}\left((m-1)[1]_f^{m-2}[f]_f^2 + [1]_f^{m-1}[f^2]_f\right), \tag{3.39}$$

where $f_1 = q_1$ if $f = x$ and $f_1 = F_1$ if $f = J$, it is easy to see that in the limit $m \to 0$,

$$\begin{aligned}\gamma_y^{R^{(a)}} &= -\frac{1}{f_1}\left[\frac{\int Dz_0 \left(\frac{d}{dz_1}[1]_f^{m-1}[f]_f\right)^2 [1]_f^{-m}}{\int Dz_0 [1]_f^m}\right] \\ &= \frac{1}{m^2 f_1^2}\left[\frac{\int Dz_0 \left(\frac{d^2}{dz_1^2}[1]_f^m\right)^2 [1]_f^{-m}}{\int Dz_0 [1]_f^m}\right], \end{aligned} \tag{3.40}$$

where $f = x$ for $y = q$ and $f = J$ for $y = F$. Using these, it is straightforward to obtain $\gamma_F^{R^{(a)}} = (1 - q_0)^2$ and hence

$$\alpha\gamma_F^{R^{(a)}}\gamma_q^{R^{(a)}} = \frac{\alpha}{w^2 q_1^2}\left[\frac{\int Dz_0 \left(\frac{d^2}{dz_1^2}[1]_x^m\right)^2 [1]_x^{-m}}{\int Dz_0 [1]_x^m}\right]. \tag{3.41}$$

The proof of instability above the critical storage capacity follows from there being a discontinuity in $d/dz_1([1]_x)$. From the Appendix,

$$\begin{aligned}\frac{d}{dz_1}[1]_x &= \frac{1}{2}\frac{e^G}{(1-\gamma g''(\lambda_0))^{3/2}}\gamma g'''(\lambda_0)\frac{d\lambda_0}{dz_1} \\ &\quad + \frac{e^G}{(1-\gamma g''(\lambda_0))^{3/2}}\Big(-w\gamma(g'(\lambda_0)\frac{d\lambda_0}{dz_1} - \frac{1}{\gamma}\frac{d}{dz_1}(\lambda_0 - \sqrt{q_1}z_1 + \sqrt{\Delta q}z_0)\Big), \end{aligned} \tag{3.42}$$

where $G \equiv w(\gamma g(\lambda_0) - (\lambda_0 - \sqrt{q_1}z_1 + \sqrt{\Delta q}z_0)^2/2)$. For the cost function $\tilde{g}(\lambda)$ (3.33) under

consideration here, λ_0 is easy to evaluate as a function of $t \equiv \sqrt{q_1}z_1 + \sqrt{\Delta q}z_0$:

$$\begin{aligned} \lambda_0 &= t \quad \text{for } t < \kappa - \sqrt{2\gamma} \\ \lambda_0 &= \kappa \quad \text{for } \kappa - \sqrt{2\gamma} < t < \kappa \\ \lambda_0 &= t \quad \text{for } \kappa < t \end{aligned} \tag{3.43}$$

Since for $\alpha > \alpha_c$ there is a discontinuity in λ_0 at $t = \kappa - \sqrt{2\gamma}$, its first and therefore its second derivative contain a delta-function at this point. Since this delta-function is then squared, it will contribute an infinite positive weight to $\gamma_q^{R_2^{(a)}}$. Therefore, whenever there is a discontinuity in λ_0, which is the case in the entire region above saturation, the sign of $\alpha\gamma_F^{R^{(a)}}\gamma_q^{R^{(a)}}$ is positive. We can therefore say that the the RSB1 ansatz is unstable.

Let us now consider qualitatively the situation for further levels of replica symmetry breaking. These will still produce a Hessian matrix that is invariant under some generalisation of the HT group; the rth level of RSB will have some replicon-like eigenvalue that can be expressed as

$$\gamma_q = c\Big[\dots\Big[\frac{\int Dz_0 \left(\frac{d^2}{dz_1^2}[1]_x^m\right)^2 [1]_x^{-m}}{\int Dz_0 [1]_x^m}\Big]_1 \dots\Big]_r, \tag{3.44}$$

to which a generalisation of the previous argument can be applied, showing that any replica-symmetry broken solution in which this term has a finite weight will be unstable. This will be the case if there is only a finite degree of RSB. We therefore conclude that for a perceptron above saturation the only exact solution is given by full replica-symmetry breaking. A study of the full replica-symmetry broken solution is, of course, outside the scope of this thesis; however, in the next section we perform a numerical study of the 2-step RSB solution, demonstrating that throughout the region above saturation it gives a lower minimum than the RSB1 solution and providing confirmation of the RSB1 result of this section.

3.3.2 2-step Replica Symmetry Breaking

2-step replica symmetry breaking (RSB2) is a relatively straightforward, but numerically complicated, extension of RSB1 [P79]. We change our notation in line with the convention introduced in chapter 2, so

$$q_{\alpha\beta} = q_{[\alpha_1,\alpha_2,\alpha_3],[\beta_1,\beta_2,\beta_3]}$$
$$\alpha_1,\beta_1 = 1,\ldots,n/m_1 \quad \alpha_2,\beta_2 = 1,\ldots,m_1/m_2 \quad \alpha_3,\beta_3 = 1,\ldots,m_2. \tag{3.45}$$

Under this notation, the RSB2 ansatz becomes

$$\begin{aligned}
q_{[\alpha_1,\alpha_2,\alpha_3],[\beta_1,\beta_2,\beta_3]} &= 1 \quad \text{if } \alpha_1 = \beta_1 \text{ and } \alpha_2 = \beta_2 \text{ and } \alpha_3 = \beta_3 \\
q_{[\alpha_1,\alpha_2,\alpha_3],[\beta_1,\beta_2,\beta_3]} &= q_0 \quad \text{if } \alpha_1 = \beta_1 \text{ and } \alpha_2 = \beta_2 \text{ and } \alpha_3 \neq \beta_3 \\
q_{[\alpha_1,\alpha_2,\alpha_3],[\beta_1,\beta_2,\beta_3]} &= q_1 \quad \text{if } \alpha_1 = \beta_1 \text{ and } \alpha_2 \neq \beta_2 \\
q_{[\alpha_1,\alpha_2,\alpha_3],[\beta_1,\beta_2,\beta_3]} &= q_2 \quad \text{if } \alpha_1 \neq \beta_1.
\end{aligned}$$

Considering the general cost function (3.9) and using the techniques introduced above, we can derive the following expression for the averaged minimum energy in the limit $\beta \to \infty$, $q_0 \to 1$:

$$\begin{aligned}
e = \lim_{\beta\to\infty} \min_{\gamma,q_1,q_2,w_1,w_2} \Bigg[& \frac{q_2}{2\gamma(1+w_1\Delta_0+w_2\Delta_1)} + \\
& \frac{\ln(1+w_1\Delta_0)}{2w_1\gamma} + \frac{\ln[1+w_2\Delta_1/(1+w1\Delta_0)]}{2w_2\gamma} + \\
& \frac{\alpha}{w_2\gamma}\int Dz_2 \ln \int Dz_1 \Bigg[\int Dz_0 \times \\
& \exp\left(-w_1\gamma\left[g(\lambda_0) + \frac{(\lambda_0 - z_2\sqrt{q_2} - z_1\sqrt{\Delta_1} - z_0\sqrt{\Delta_0})^2}{2\gamma}\right]\right)\Bigg]^{w_2/w1}\Bigg].
\end{aligned} \tag{3.46}$$

Here $\gamma = \beta(1-q_0)$ as before and $w_i \equiv \beta m_i/\gamma$; we also use the shorthand $\Delta_0 = q_0 - q_1$, $\Delta_1 = q_1 - q_2$. It can easily be seen that if we take $q_1 = q_2$, $\Delta_1 = 0$ and this formula reduces to the RSB1 case.

In the case under investigation here,with the cost function given by (3.33), the

innermost integral reduces to

$$I_{z_0} = \frac{\exp\left(\frac{1}{2}\frac{-w_1 A^2}{1+w_1\Delta_0}\right)}{\sqrt{(1+w_1\Delta_0)}}\left[H\left(\frac{A-\sqrt{2\gamma}(1+w\Delta_0)}{\sqrt{\Delta_0(1+w_1\Delta_0)}}\right) - H\left(\frac{A}{\sqrt{\Delta_0(1+w_1\Delta_0)}}\right)\right] + \exp(-w_1\gamma)H\left(\frac{\sqrt{2\gamma}-A}{\sqrt{\Delta_0}}\right) + H\left(\frac{A}{\sqrt{\Delta_0}}\right), \quad (3.47)$$

where $A = \kappa - z_2\sqrt{q_2} - z_1\sqrt{\Delta_1}$ and $H(x) = \int_x^\infty Du$. We evaluated (3.46) for a range of values of α and for $\kappa = 1$, performing the minimisation numerically. The results are shown in Figures 3.3-3.7. The transition from RSB1 to RSB2 causes the output error e to increase by an amount that is typically $\mathcal{O}(10^{-4})$; this is small, but greater than the numerical tolerance of our minimisation procedure which is $\mathcal{O}(10^{-6})$. Any lack of smoothness in the curves presented is due to a combination of two factors: first, the minima for this problem prove to be relatively wide; second, the function varies much faster in general with respect to γ, q_1, q_2 than w_1, w_2, meaning that the choice of starting values of w_1, w_2 is of great importance.

We find that throughout the region of replica symmetry instability the RSB2 solution is at a lower minimum energy (error) than the RSB1 solution, showing that RSB1 is not the true minimum anywhere in this region. Figure 3.3 shows the minimum errors evaluated within RSB1 and RSB2 and Figure 3.4 shows their differences. The other figures show the behaviour of the minimising values of γ, q_1, q_2, w_1, w_2, with the corresponding RSB1 values for comparison. As $\gamma \equiv \lim_{\beta\to\infty}\beta(1-q_0)$, and as we expect that q_0 will increase with increasing levels of RSB, γ^{rsb2} should be less than γ^{rsb1}; our results (Figure 3.5) confirm this. A decrease in γ corresponds to a slight narrowing of the gap in the distribution of local stabilities, which is of magnitude $\sqrt{2\gamma}$ ([GD88], [MEZ93]). We have calculated this distribution but the result is sufficiently similar to the RSB1 result that we do not reproduce it here. We find that q_1 and q_2 are respectively higher and lower than q_1^{rsb1} (Figure 3.6), and that $w_1\gamma$ and $w_2\gamma$ are respectively higher and lower than $w_1^{rsb1}\gamma^{rsb1}$ (Figure 3.7), as would be expected.

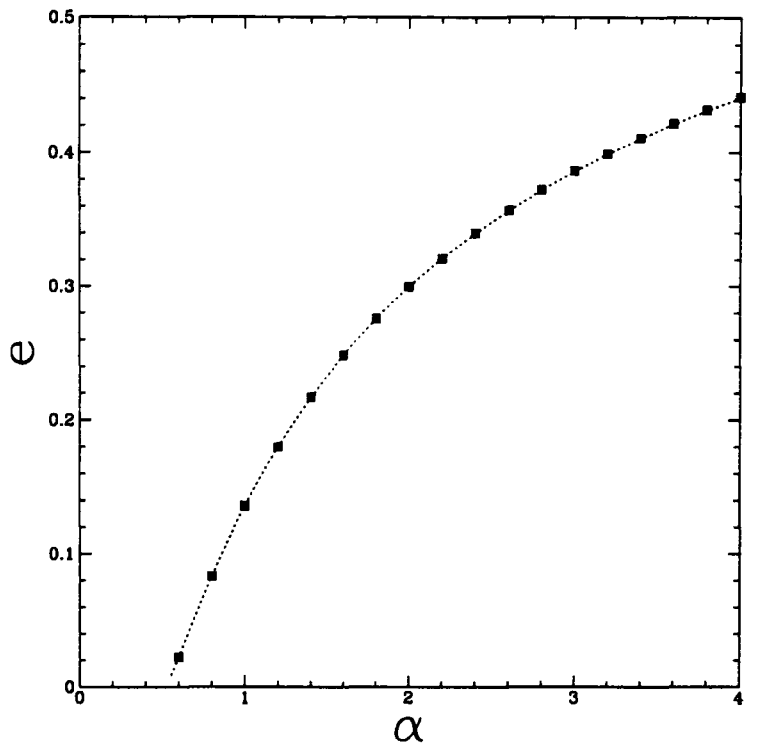


Figure 3.3: The minimum error e for RSB1 (dotted line) and RSB2 (points), for $\kappa = 1$.

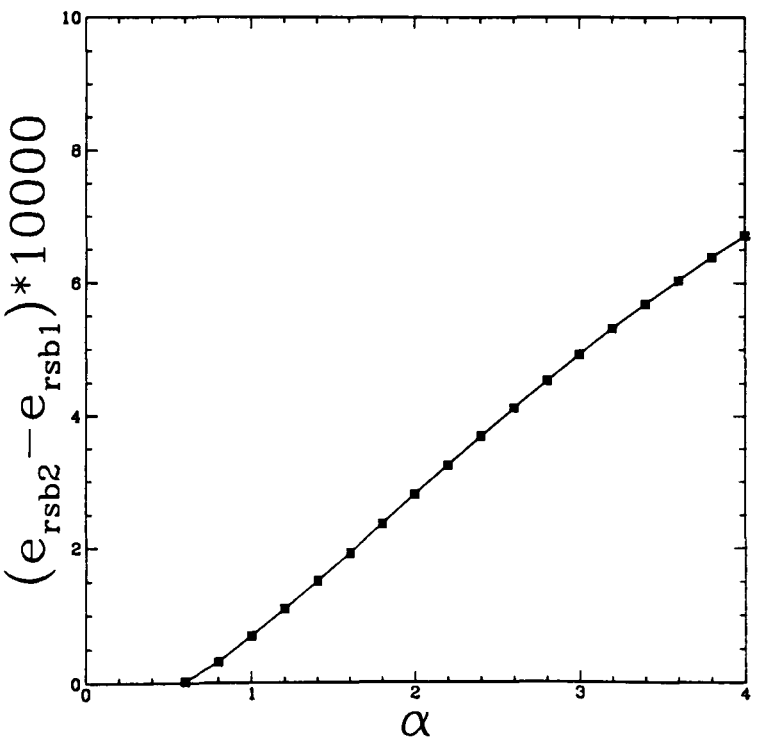


Figure 3.4: The difference between the minimum errors obtained for the RSB1 and the RSB2 solution, measured in units of 10^{-4}. The line is a guide to the eye. $\kappa = 1$.

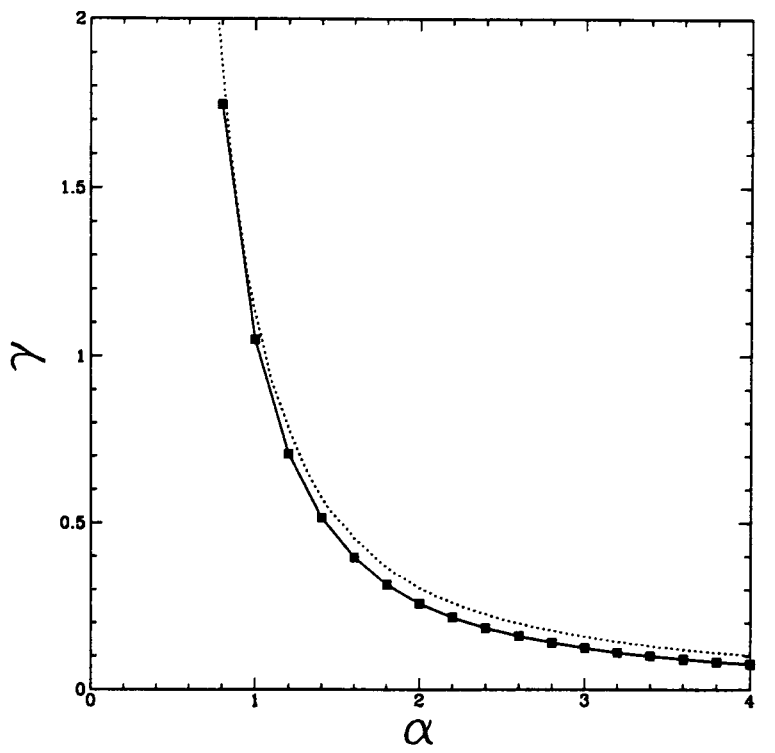


Figure 3.5: γ^{rsb1} (dotted line) and γ^{rsb2} (points). The full line is a guide to the eye. $\kappa = 1$.

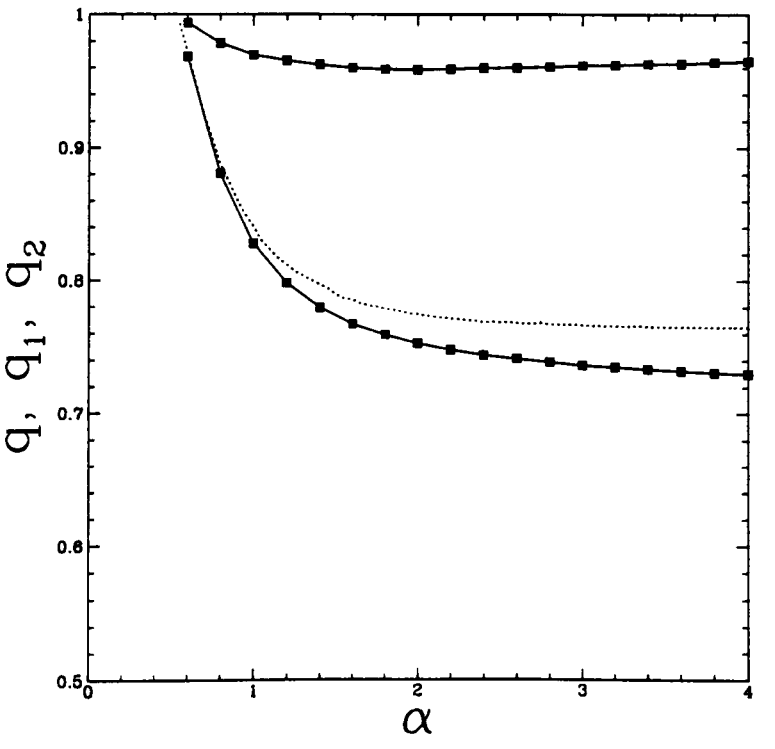


Figure 3.6: q_1^{rsb1} (dotted line). q_1^{rsb2} (upper set of points) and q_2^{rsb2} (lower set of points). The full lines are guides to the eye. $\kappa = 1$.

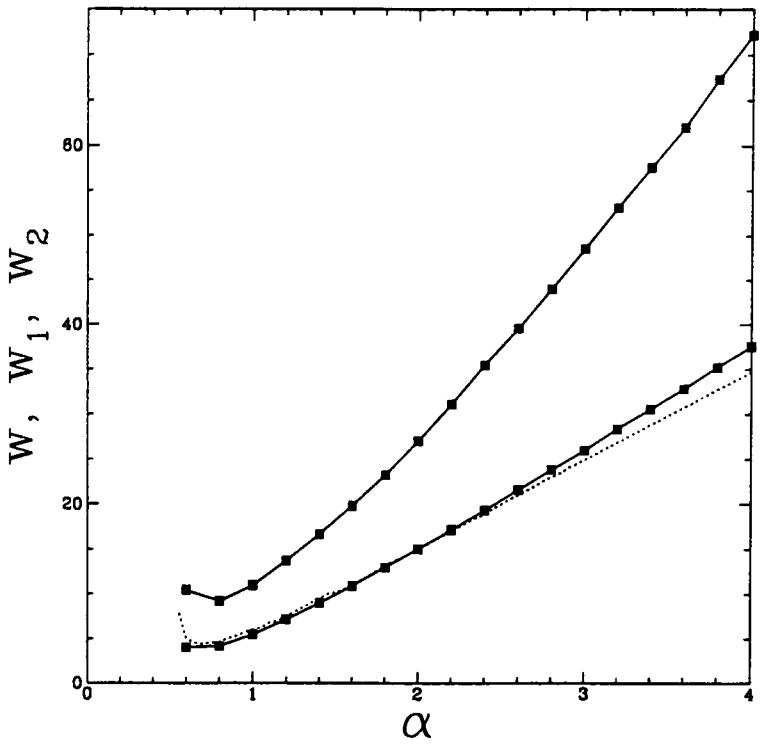


Figure 3.7: w_1^{rsb1} (dotted line), w_1^{rsb2} (upper set of points) and w_2^{rsb2} (lower set of points). The full lines are guides to the eye. $\kappa = 1$.

3.3.3 Conclusions

It has been known for some time that some degree of replica symmetry breaking is necessary for an exact solution to the problem of minimum-error storage of patterns in a perceptron above saturation. We have shown that such a solution requires full RSB. Our results imply that any finite level of replica symmetry breaking for any cost function $g(\lambda)$ will be unstable if the function $\lambda_0(z_i, q_i, \gamma)$ contains a discontinuity. The result for the perceptron is backed up by a numerical investigation of the global stability of the RSB1 solution, showing that it will not minimise the saddle-point equations anywhere in the region above saturation. This numerical study has, however, also shown that the observable effects of the transition from RSB1 to RSB2 are very small compared to the effects of the transition from RS to RSB1. so that for most purposes RSB1 may be considered sufficient for this case.

This section has concentrated on the free energy of the perceptron trained with the Gardner-Derrida cost function. This cost function is somewhat special, in that the discontinuity in λ_0 is present throughout the region $\alpha > \alpha_c$. For cost functions such as

the perceptron cost function, $g(\kappa - \lambda) = (\kappa - \lambda)\theta(\kappa - \lambda)$, this is not the case, and the transition to RSB2 is still an open question.

3.3.4 Appendix: Details of the Hessian Matrix Elements

This discussion in large part reproduces and expands on the results of [D92], making those changes necessary to make it relevant to the present case. Further discussion of the matrix calculations of RSB1 can be found in [PS94b].

By consideration of the group orbits, we can obtain the different matrix elements of the submatrices H^{qq}, H^{FF}. As in section 2.3, we change our notation to

$$q_{\alpha\beta} = q_{[\alpha_1,\alpha_2],[\beta_1,\beta_2]}, \tag{3.48}$$

where α_1, β_1 label the blocks and α_2, β_2 label the individual matrix elements within the blocks. We make the corresponding changes in notation for $F_{\alpha\beta}$, $H_{(\alpha\beta)(\gamma\delta)}$. The values of those submatrix elements which are relevant to the calculation of the replicon-like eigenvalues can be expressed as follows (subscripts refer to variables being averaged over, superscripts are exponents):

$$\begin{aligned}
H_{[1,1][1,2][1,1][1,2]} &\equiv K_1 \\
&= \left[\left[[f^2]_f^2\right]_0\right]_1 - \left[\left[[f]_f^2\right]_0\right]_1^2 \\
H_{[1,1][1,2][1,1][1,3]} &\equiv K_2 \\
&= \left[\left[[f^2]_f[f]_f^2\right]_0\right]_1 - \left[\left[[f]_f^2\right]_0\right]_1^2 \\
H_{[1,1][1,2][1,3][1,4]} &\equiv K_3 \\
&= \left[\left[[f]_f^4\right]_0\right]_1 - \left[\left[[f]_f^2\right]_0\right]_1^2 \\
H_{[1,1][2,1][1,1][2,2]} &\equiv L_1 \\
&= \left[\left[[f^2]_f\right]_0^2\right]_1 - \left[\left[[f]_f\right]_0^2\right]_1^2 \\
H_{[1,1][2,1][1,1][2,2]} &\equiv L_2 \\
&= \left[\left[[f^2]_f\right]_0\left[[f]_f\right]_0^2\right]_1 - \left[\left[[f]_f\right]_0^2\right]_1^2
\end{aligned}$$

$$
\begin{aligned}
H_{[1,1][2,1][1,2][2,2]} &\equiv L_3 \\
&= \Big[\big[[f]_f^2\big]_0^2\Big]_1 - \Big[\big[[f]_f\big]_0^2\Big]_1^2 \\
H_{[1,1][2,1][1,1][3,1]} &\equiv L_4 \\
&= \Big[\big[[f]_f^2\big]_0^2\Big]_1 - \Big[\big[[f]_f\big]_0^2\Big]_1^2 \\
H_{[1,1][2,1][1,2][3,1]} &\equiv L_5 \\
&= \Big[\big[[f]_f^2\big]_0\big[[f]_f\big]_0^2\Big]_1 - \Big[\big[[f]_f\big]_0^2\Big]_1^2 \\
H_{[1,1][2,1][3,1][4,1]} &\equiv L_6 \\
&= \Big[\big[[f]_f\big]_0^4\Big]_1 - \Big[\big[[f]_f\big]_0^2\Big]_1^2,
\end{aligned}
\tag{3.49}
$$

where $f \equiv x$ for H^{qq}, $f \equiv J$ for H^{FF}, and the averages $[\ldots]_1, [\ldots]_0, [\ldots]_J, [\ldots]_x$ are defined as follows:

$$
\begin{aligned}
\big[h\big]_1 &= \int Dz_1\, h(z_1) \\
\Big[\prod_{i=1}^{s} [h_i]_f\Big]_0 &= \frac{\int Dz_0\, [1]_f^{m-s} \prod_{i=1}^{s} [h_i]_f}{\int Dz_0\, [1]_f^m} \\
\big[h\big]_J &= \int dJ\, e^{\left(-\frac{1}{2}J^2(2E-F_0)+iJ(\sqrt{F_1}z_1+\sqrt{F_0-F_1}z_0)+E\right)} h(J) \\
\big[h\big]_x &= \int d\lambda\, dx\, e^{\left(\beta g(\lambda)+ix(\lambda-\sqrt{q_1}z_1-\sqrt{q_0-q_1}z_0)-\frac{1}{2}(1-q_0)x^2\right)} h(x)
\end{aligned}
\tag{3.50}
$$

The eigenvalues γ^R of the replicon-like eigenvector families R are as follows:

$$
\begin{aligned}
R^{(a)}: &\quad \gamma^{R^{(a)}} = K_1 - 2K_2 + K_3 \\
R_1^{(e)}: &\quad \gamma^{R_1^{(e)}} = L_1 - 2L_2 + L_3 + 2m(L_2 - L_3 - L_4 + L_5) + m^2(L_3 - 2L_5 + L_6) \\
R_2^{(e)}: &\quad \gamma^{R_2^{(e)}} = L_1 - 2L_2 + L_3 \\
R_3^{(e)}: &\quad \gamma^{R_3^{(e)}} = L_1 - 2L_2 + L_3 + m(L_2 - L_3 - L_4 + L_5)
\end{aligned}
\tag{3.51}
$$

We can also evaluate $[1]_J$ and (in the limit $\beta \to \infty$) $[1]_x$. These are:

$$
[1]_J = \frac{1}{\sqrt{2E - F_0}} \exp\left(E - \frac{(\sqrt{F_1}z_1 + \sqrt{F_0 - F_1})^2}{2(2E - F_0)}\right)
$$

$$[1]_x = \frac{1}{\sqrt{1-\gamma g''(\lambda_0)}} \exp\left(-w\gamma\left[g(\lambda_0) + \frac{1}{2\gamma}(\lambda_0 - \sqrt{q_1}z_1 - \sqrt{\Delta q}z_0)^2\right]\right). \quad (3.52)$$

3.4 Replica symmetry Breaking in Noise-Optimal Perceptrons

The fact that different cost functions are used in calculations of optimal neural networks reflects the fact that there are many different properties of these nets, and it is, generally speaking, not possible to optimise all of these properties simultaneously. The *principle of adaptation* ([WS90a]) states that the best network for a particular retrieval enviroment is one that was optimised in the same training environment. Thus, for example, the Maximally Stable network discussed above is trained on entirely noiseless patterns, and is the optimal network for storing or retrieving noiseless patterns in a noiseless environment. However, in real neural network implementations, it will be rare to find an entirely noiseless environment. On the one hand, there will be operating noise analogous to the operating temperature T discussed in the Hopfield model. On the other, the patterns presented for recall will almost certainly be corrupted in some way. It is therefore of interest to discuss a network that is optimally adapted for operation in the presence of noise.

The example of the maximally stable network provides another motivation for the study of training with noise. The rationale behind the introduction of the stability parameter κ was to attempt to increase the basins of attraction of the noiseless patterns by storing the patterns as "strongly" as possible. By considering the optimal retrieval quality possible for any given level of operating noise, we can judge whether or not the MSN is successful in this endeavour.

The *noise-optimal* network was introduced by Sherrington and Wong ([WS90a], [WS90b], [WS93]), who have studied it extensively in the RS regime. Of particular interest is the fact that the distribution of stabilities shows a gap for high m_t (illustrated in figure 3.10); it is as if, in order to perform as well as possible in those cases where the pattern can be recalled, the network deliberately ignores the stability of those patterns that can't. This "sacrificial effect" proves to be the key to the success of the noise-optimal network. A similar effect is displayed by the Garnder-Derrida perceptron above saturation, as we

have seen.

However, as we have also seen, the presence of this gap in $\rho(\lambda)$ means that RSB is occuring. It is therefore of interest to investigate the nature of RSB for this network, and the nature of the transition to RSB, in order to compare it to the results in the Gardner-Derrida perceptron.

3.4.1 The Model: Review of Previous Results

As before, the system we are interested in is a perceptron with N input nodes (labelled $i = 1, \ldots, N$) and one output node, this time obeying the stochastic update rule

$$S_{output} = \text{sign}\left(\sum_i J_i S_i + Tz\right), \quad S_i \in \{\pm 1\}, \tag{3.53}$$

where the $\{J_i\}$ are constrained by the spherical rule $\sum_i J_i^2 = N$ and z is a Gaussian.

In line with the principle of adaptation, we train the network by presenting it with examples featuring the level of noise which we expect in the input patterns in the retrieval stage. This perceptron is trained to store correctly as many as possible of an ensemble of Qp noisy examples $\{\eta^{\mu\nu}\}(\mu = 1, \ldots, Q, \nu = 1, \ldots, p)$, which are generated from a corresponding set of $p = \alpha N$ noiseless patterns $\{\xi^\mu\}$ by the rule

$$\mathbf{P}(\eta_i^{\mu\nu}) = (1 + \eta_i^{\mu\nu}\xi_i^\mu m_t)/2, \tag{3.54}$$

where m_t is known as the "training overlap"; correspondingly the training noise is the complement $d_t \equiv \frac{1}{2}(1 - m_t)$.

The noiseless patterns, or "prototypes" $\{\xi^\mu\}$ are drawn at random from $\{\pm 1\}^N$. We define the "aligning field" $\lambda^{\mu\nu}$ of an example $\eta^{\mu\nu}$ by $\lambda^{\mu\nu} = \xi^\mu_{\text{output}}/|J| \sum_i J_i \eta_i^{\mu\nu}$. by analogy with the stability of the noiseless patterns defined above. Given the aim of storing as many as possible of these examples, our cost function to be minimised is simply

$$E = -\sum_{\mu\nu} \text{sign}(\lambda^{\mu\nu}). \tag{3.55}$$

In the limit $Q \to \infty$, we can average over the quenched cost function above to obtain an annealed cost function defined only with respect to the stabilities of the noiseless patterns and the operating temperature:

$$E = \sum_{\mu} g(\lambda^{\mu}), \quad g(\lambda) = -\mathrm{erf}\left(\frac{m_t \lambda}{\sqrt{2[1 - m_t^2 + T^2]}}\right). \tag{3.56}$$

This function can be minimised by the methods outlined in the previous section. We summarise here some of the results from the RS case.

First, it was shown that for an arbitrary noise level it is not always possible to obtain error-free storage. This is a complicating factor when attempting to design an algorithm to create the noise-optimal network, as we shall see.

Second, as stated in equation (3.24), we can write f as $f = \int d\lambda\, \rho_m(\lambda)\, g_m(\lambda)$ (the subscripts are for later convenience). In this case $g(\lambda)$ also corresponds to the function used in the one-step updating equation (1.8) if we take $T = 0$ in (3.56). Therefore, the network we have found is the one which maximises the one-step update for an input stimulus with overlap m with a stored pattern.

Third, for a wide range of m_t, α, the aligning field distribution contains a gap, because we have had to perform a Maxwell's construction in evaluating (3.20). We interpret this as a "sacrificial effect"; rather than attempt to store correctly those patterns with high negative stability, the network attempts instead to increase $\rho(\lambda)$ at high values of λ in order to maximise $\int d\lambda\, \rho(\lambda)\, g(\lambda)$.

Fourth, the network displays wide retrieval in a greater area of the phase diagram than the maximally stable network discussed in the previous section (though to a lower value of the retrieval fixed-point). This will be discussed in more detail in the next section.

The local stability properties of the RS solution are given in [WS93]. However, we are concerned here with the global stability, which we assess by evaluating the minimal values of equation (3.31) for our specific cost-function. In theory, it is possible for the RS solution to be locally stable but yet not provide a global minimum of the free energy. Our interest will be to see if this is the case, and to investigate the extent to which RSB

affects the performance of the network.

3.4.2 Results

We have evaluated the average minimum cost at several values of α and m_t for $T = 0$, on both sides of the AT line; note that asymptotically, as $\alpha \to \infty$, the critical AT value of $m_t \sim 0.84$, but it goes to zero as $\alpha \to 0$. Results for $q_0(m_t)$ are exhibited in Fig. 3.8 for several values of α, along with the corresponding AT values. They clearly demonstrate that replica symmetry is broken for noise values less than the AT values and that its onset is continuous as the line is crossed from the noisier side. Fig. 3.9 shows results for $q_0(\alpha)$ as a function of α for three values of m_t. As expected, for $m_t = 0.8$ there is no RSB1 until α is reduced beneath the corresponding AT value, $\alpha_{AT} = 2.80$, whereas for $m_t = 0.99$ and 0.85, replica symmetry is broken for all values of α. In all the cases shown q_0 increases as α is reduced towards zero, in accord with the expectation that all optimal networks become Hebb-like for a finite number of patterns, and replica symmetry can at most be marginally broken [WRS92]. In the case of $m_t = 0.99$, q_0 approaches 1 as α is reduced to order 2, reflecting the fact that the $m_t = 1$ limit corresponds to using the Gardner-Derrida [GD88] minimal stability cost function which is replica symmetric for $\alpha < 2$. The decrease of q_0 as α is reduced from high values to order 1, seen for $m_t = 0.85$ and for $m_t = 0.8$ below the AT value, is a reflection of the migration of these m_t values further from the (decreasing) AT line and deeper into the RSB region.

Next we examine the effects of one-step RSB on the local aligning field distribution $\rho(\lambda)$. For this we take two points, $\alpha = 1.5, m_t = 0.9$ and $\alpha = 4, m_t = 0.85$, both of which are in the RSB1 region but which under RS respectively do and do not have a gap in $\rho(\lambda)$. The results for $\alpha = 1.5, m_t = 0.9$ under RS and RSB1 are shown in Figure 3.10. As observed in the corresponding noiseless study of Majer *et al*, [MEZ93], RSB1 has a marked effect on the aligning field distributions; in particular, the band-gap is narrowed significantly.

These results can be compared with those obtained by direct simulational training on

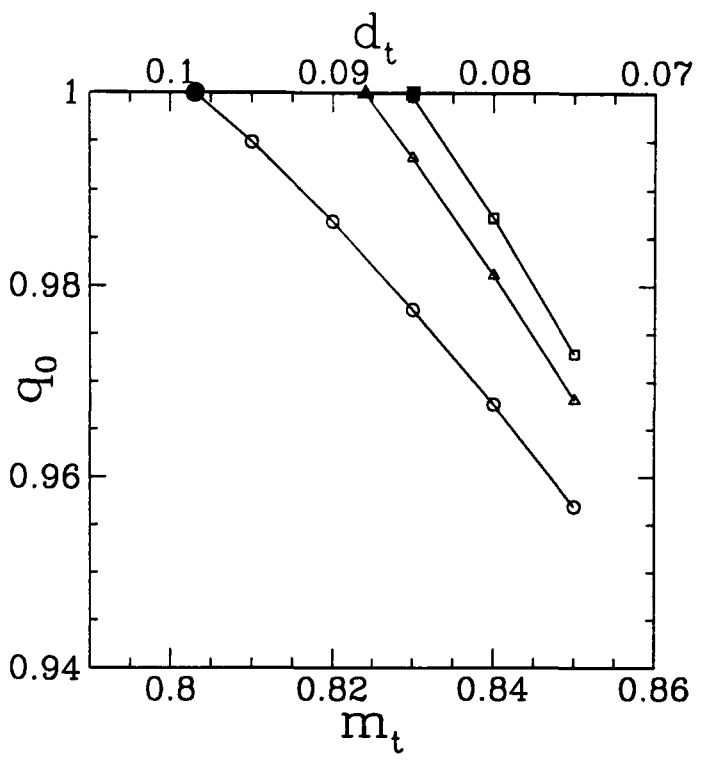


Figure 3.8: q_0, the measure of replica symmetry breaking, as a function of m_t for $\alpha = 3$ (circles), 6.5 (triangles), 10 (squares). The hollow points are results of numerical minimisation of (9); the full points are the calculated AT values. The lines are a guide to the eye.

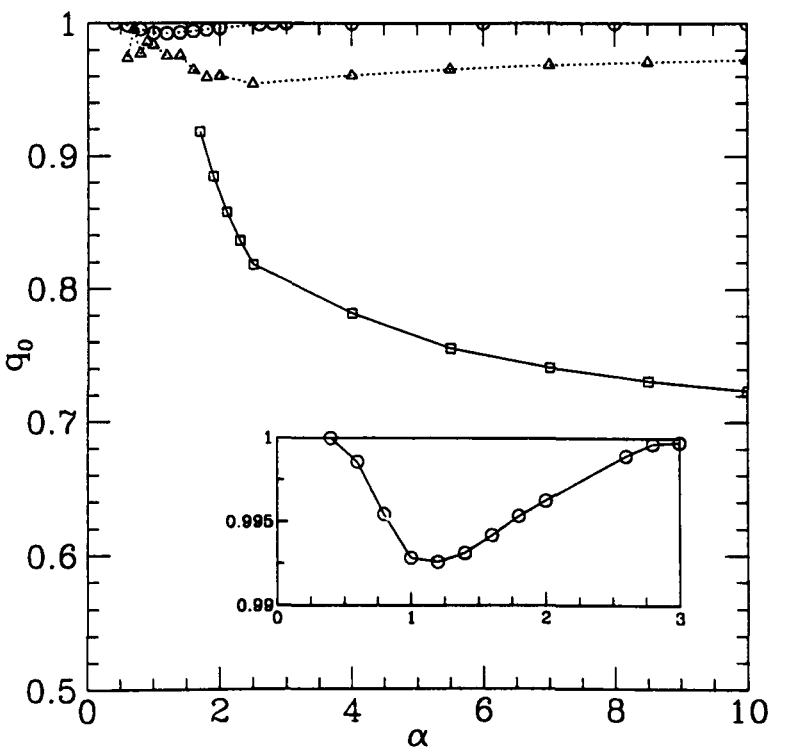


Figure 3.9: q_0 as a function of α for $m_t = 0.99$ (squares), 0.85 (triangles), 0.8 (circles). The lines are a guide to the eye. The inset shows the $m_t = 0.8$ result in more detail.

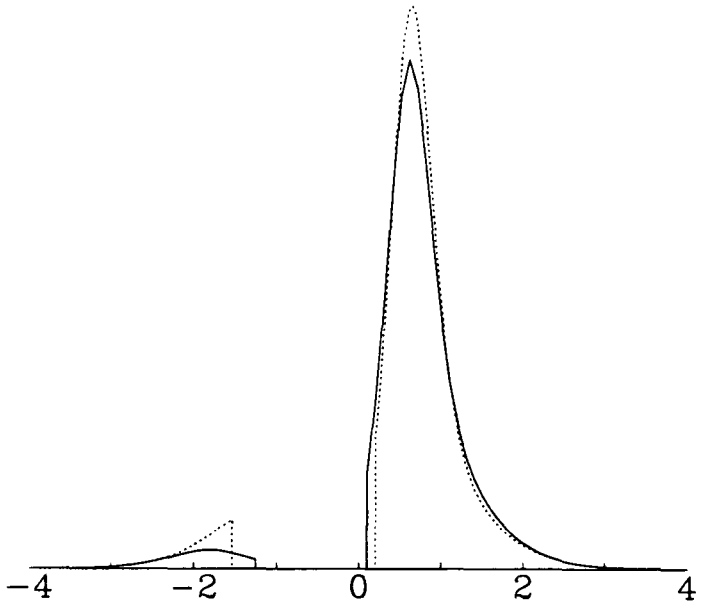


Figure 3.10: The distribution of aligning fields, $\rho(\lambda)$, for $m_t = 0.9$, $\alpha = 1.5$ for the replica symmetric (dotted line) and RSB1 (full line) ansatzes.

the annealed cost function for a realised set of patterns. The aligning field distributions displayed in Figure 3.11 were obtained on networks of 200, 400 and 800 neurons for $\alpha = 1.5, m_t = 0.9$. A simple gradient descent algorithm was used to minimise the cost function $E = -\sum_\mu \mathrm{erf}(m_t \lambda^\mu / \sqrt{2[1 - m_t^2]})$ (more details are given in the next section). As can be seen by comparison with Figure 3, RSB1 improves the agreement of the theoretical and experimental curves.

In the region where RSB exists but there is no gap in the aligning field distribution, the effect of moving from RS to RSB1 on the distribution is to raise the minimum in the pseudogap, which brings the results shown in Figure 3.12 for the case $\alpha = 4, m_t = 0.85$ more into agreement with those obtained from simulations.

Since RSB occurs when the replica symmetric ansatz no longer gives a global minimum for $\tilde{E}$, its effect will be to decrease the performance of the network. Figure 3.13 compares the performances under the RS and RSB1 ansatzes for the three values of m_t used above. For $m_t = 0.8$, 0.85 the effect of RSB is small, as might have been expected from the fact that q_0 is very near 1 in the RSB1 regime for these m_t values. For $m_t = 0.99$, however, the effect of RSB1 is to cause a marked decrease in the performance, particularly

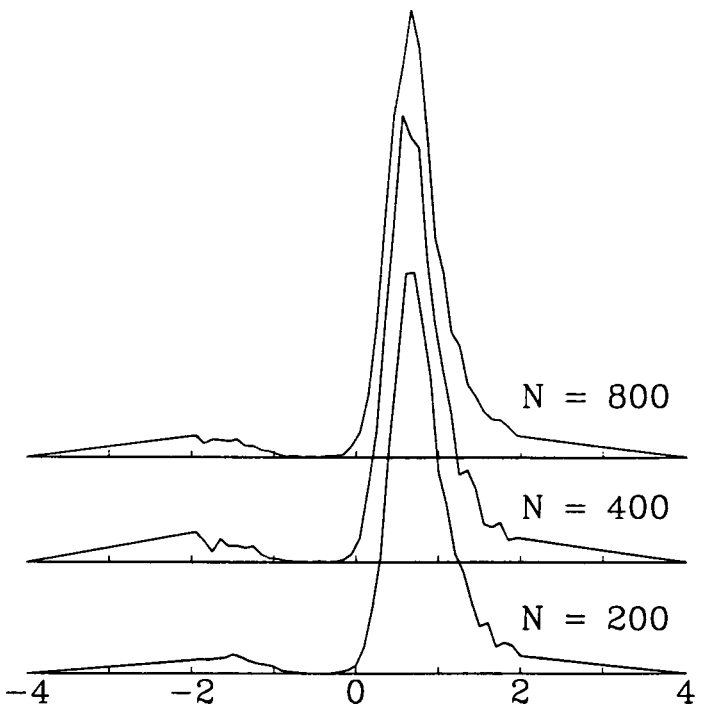


Figure 3.11: The aligning field distribution, $\rho(\lambda)$, obtained for $m_t = 0.9$, $\alpha = 1.5$, $-2 < \lambda < 2$ on networks of 200, 400 and 800 neurons by gradient descent on the cost function $g(\lambda)$.

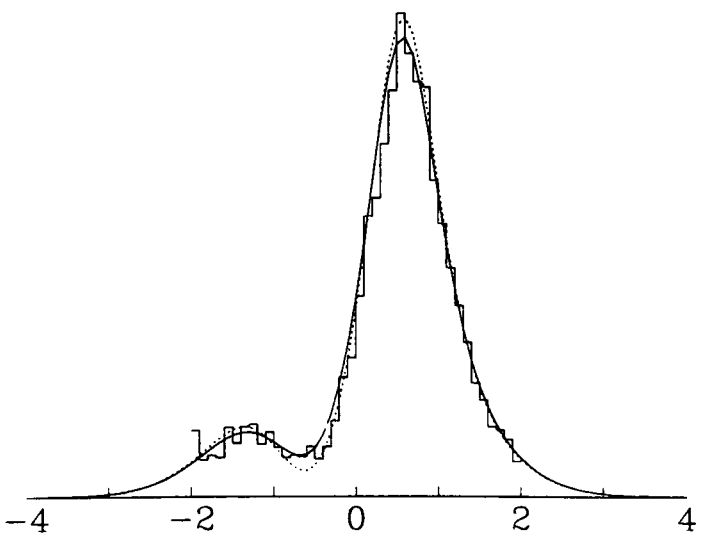


Figure 3.12: The distribution of aligning fields, $\rho(\lambda)$, for $m_t = 0.85$, $\alpha = 4$ for the replica symmetric (dotted line) and RSB1 (full line) ansatzes, and as obtained by direct simulational training on a network of 200 neurons (histogram represents average over ten training runs).

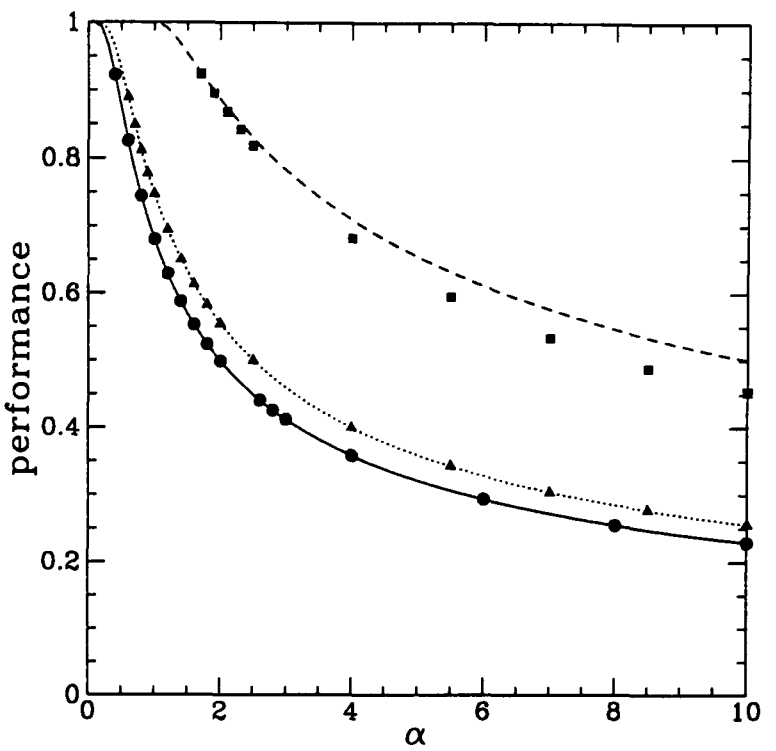


Figure 3.13: The performance overlaps obtained as a function of α using RS (lines) and RSB1 (points) for $m_t = 0.8$ (solid line, circles), $m_t = 0.85$ (dotted line, triangles) and $m_t = 0.99$ (dashed line, squares).

at high values of α.

3.4.3 Conclusions

We have confirmed the existence of replica symmetry breaking in the noise-optimal perceptron, and observed a continuous transition to replica symmetry breaking where the RS solution becomes locally unstable. This continuous transition is in line with other studies of RSB in perceptrons. The effects on the observables of RSB are only noticeable for high values of m_t and α. The effects on the distribution of local stabilities are more pronounced, leading to a better agreement between the theory and results from simulations.

3.5 Retrieval Performance for Optimised Neural Networks

We conclude this analysis with a brief review of the performance of the two optimised nets we have investigated compared to the Hopfield network. After all, it is a considerably easier task to construct the Hopfield network than to construct an optimised net, and we need to be sure the extra effort is worth while.

We first present the basins of attraction at zero operating temperature for the extremely dilute Hopfield network (Figure 3.14a, from [DGZ87]), and the extremely dilute maximally stable network (Figure 3.14b, from [G89]). In these figures, m^* is the final retrieval overlap, given by the stable fixed-points of the mapping $f(m)$ (1.8), and m_B is the boundary of the basin of attraction. We see that the retrieval behaviour of the two networks is quite different. The dilute Hopfield network displays wide retrieval throughout the region where retrieval takes place, but at the price of a relatively low final overlap. The dilute MSN, on the other hand, only displays wide retrieval for $\alpha < 0.42$, for higher values of α requiring higher and higher initial overlaps in order to recall the stored patterns; however, this recall is always perfect. The results presented here are both replica-symmetric, but the analysis of the preceding sections of this chapter leads us to believe that the same qualitative behaviour will be observed for the exact results.

We finally present the phase diagram of the noise-optimal network in the operating temperature-storage capacity phase space (Figure 3.15, from [WS90b]). Wide retrieval occurs in region I, and narrow retrieval in region III. In the small region II, the mapping $f(m)$ proves to have two stable fixed-points, one at a low value of m^* and one at $m^* \sim 1$. The network will thus display wide retrieval to the lower fixed-point and narrow retrieval to the upper one. Some form of wide retrieval is observed for $\alpha < 0.6$ at $T = 0$; at $\alpha \sim 0.6$ the lower fixed-point goes to 0 continuously.

The dotted line in the figure shows the boundary of the retrieval region for the MSN at the given operating temperature. For all values of T, the wide retrieval region for the noise-optimal network is significantly larger than for the MSN. Note that for every

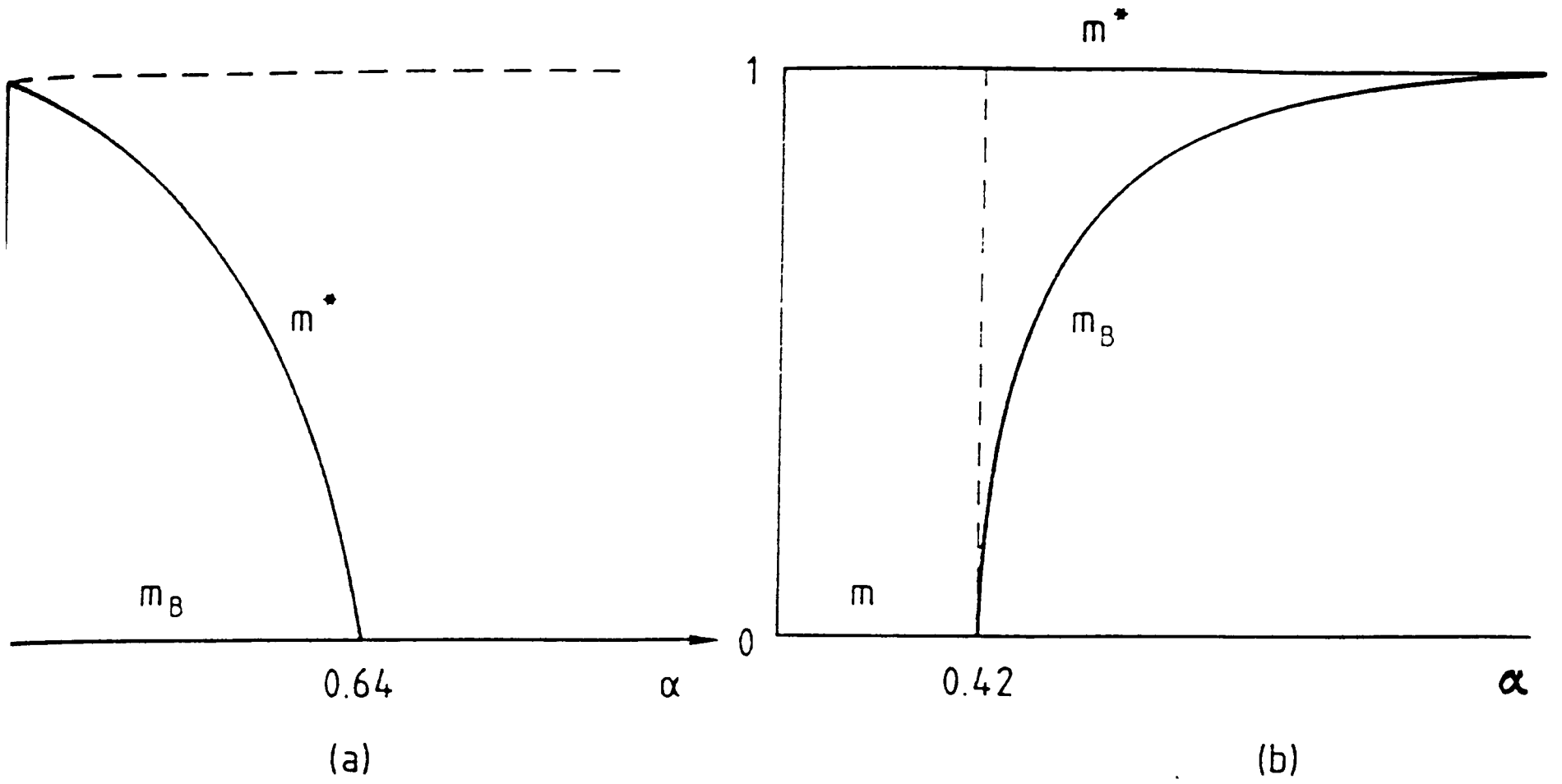


Figure 3.14: The basins of attraction for the Hopfield and Maximally stable networks

value of T, α, a different noise-optimal network exists; thus, when we are comparing the phase diagram for retrieval of the noise-optimal network with that for the MSN, we are not comparing one network with another, but rather the behaviour of an entire class of networks (the noise-optimal networks) with the behaviour of a single network.

In conclusion, this brief review has demonstrated that the noise-optimal network compares favourably with the MSN in terms of the size of the basins of attraction. The next topic of interest, therefore, is to attempt to find algorithms which allow us to realise the benefits of training with noise in an actual network.

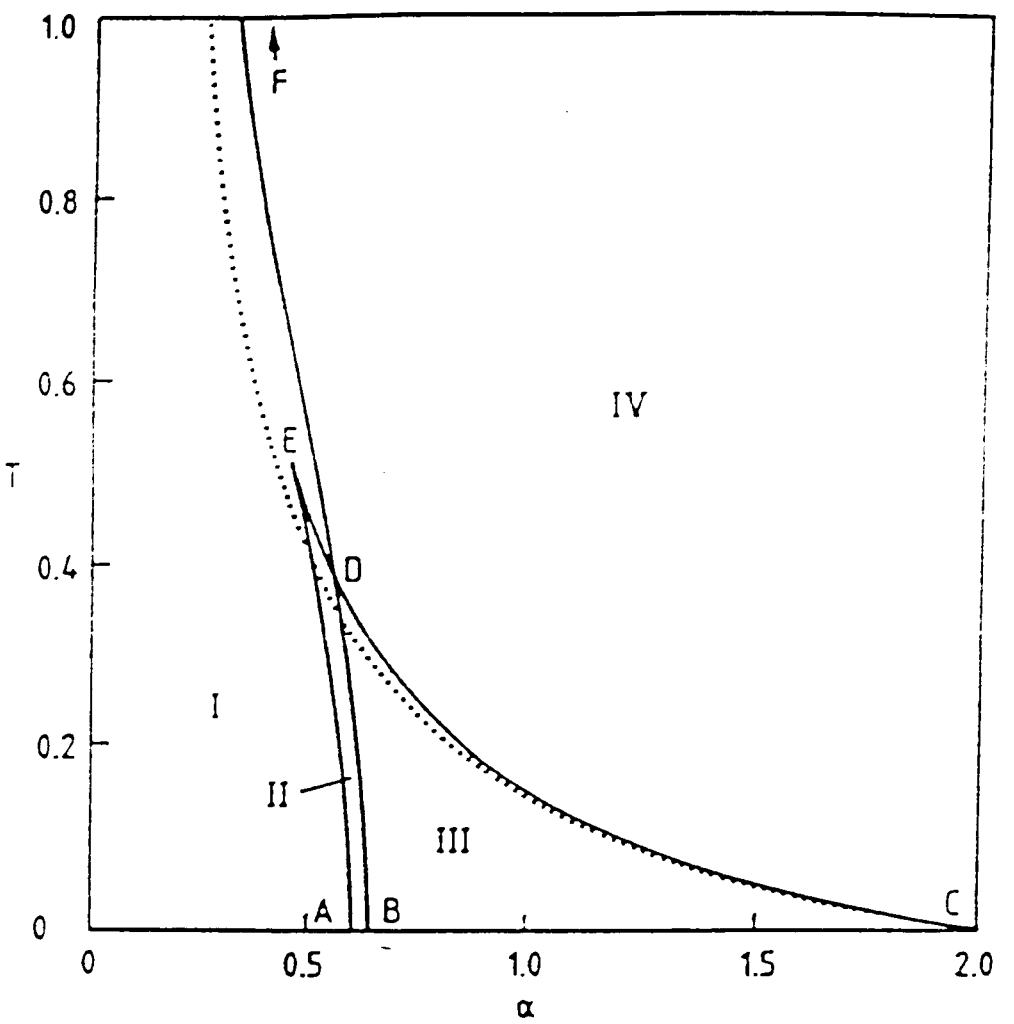


Figure 3.15: The phase diagram for the noise-optimal network

3.6 Training Algorithms for a Noise-Optimal Network

3.6.1 Introduction

Even for a system as simple as a perceptron, the problem of finding a training algorithm that is simple and effective is still the subject of vigorous investigation. To an extent, this is because different users may have different priorities, reflected in the choice of cost function as discussed earlier. Another reason, however, is that this field is still very empirical, especially for networks above saturation. A good review of the theory that exists at the moment is given in [WRB93]. Detailed dynamical studies have been performed on several different training algorithms ([O89], [HS90], [I95]), and there have been attempts to look at variations between algorithms in a systematic way ([GG91], [GPB92]), but there is, as yet, no theory of how to obtain the optimal algorithm; in a sense, the study of algorithms has its Hopfield calculations but no Gardner calculation.

In this section, we discuss five different algorithms to produce a noise-optimal net-

work. The first uses the prototype patterns, training by gradient descent on the annealed cost function 3.56. This produces a network whose performance is comparable to the calculated optimal performance of a network trained with noise. However, it is more realistic to train on the ensemble of noisy examples themselves, and this is what the remaining four algorithms are designed to do.

We study the case $\alpha = 0.5$. The highly dilute network made up of a set of perceptrons of maximal stability exhibits narrow retrieval at this loading [G89], with the minimum overlap necessary for retrieval $m_B = 0.67$. In the case of the noise-optimal network we theoretically find wide retrieval; it is the aim of this section to find out whether this is possible in practice. The parameter that governs whether or not "retrieval" (defined, somewhat loosely, as a final output overlap greater than the input overlap) takes place is $f(m) - m$, which we refer to as the "gain"; if this is positive retrieval will occur. Our results are presented in terms of this parameter. Figure 3.16 shows $f(m) - m$ for the noise-optimal network with $\alpha = 0.5$.

3.6.2 Training with Annealed Noise

Given the noiseless prototypes, it is a straightforward matter to train by gradient descent on the cost function (3.56). This training was carried out for networks of size $N \in \{200, 400, 800\}$ for a number of different initial J_{ij}'s. The minimised free energy corresponds to $-f_m(m)$.

Figure 3.16 shows $f(m) - m$ averaged over 10 runs at $N = 200$. Errorbars are suppressed for clarity. At high m_t the gradients are not sufficient for good learning; however, at low m_t this algorithm gives output overlaps close to the calculated maximum output overlap. It displays retrieval for $0 < m_t < 0.95$.

We have also calculated $\rho(\lambda)$ for the noiseless patterns, for $\alpha = 1.5, m_t = 0.9$. The results are displayed in Figure 3.11, and show good qualitative agreement with the RSB1 solution.

The annealed-noise cost function therefore provides an effective way of producing a

network that is nearly noise-optimal. However, one reason for considering training with noise in the first place was that we frequently only have a noisy training ensemble to train on, and the noiseless prototypes are not available. We would therefore like to develop training algorithms to use in this situation. Some such algorithms are described in the next section.

3.6.3 Training with Quenched Noise

We now turn to consider the case where we are presented with a training ensemble generated from a set of prototypes using a noisy rule. SW have shown that, except where the noise level is very low, it is impossible to have entirely error-free storage. We are thus attempting to find training algorithms for the perceptron above saturation.

In this case there are two error measures of interest. The *training error* is the error rate on the patterns in the training ensemble; the *generalisation error* is the error rate on patterns which are not in the training ensemble but are generated with the same noise level. The generalisation error ϵ_g is related simply to the one-step output overlap $f(m)$ by

$$f(m) = 1 - 2\epsilon_g. \tag{3.57}$$

Our learning strategy depends on the knowledge we assume to be available. If we are training a net as an associative memory, then the noisy examples that we have to train on will be presented one at a time and an appropriate algorithm treats each example on an equal footing, not knowing which prototype it is derived from or what the noise level is (if any). For a perceptron that is being trained by the method of training with noise to give the correct output to an input stimulus, however, we will know which output corresponds to which example and can thus construct a more sophisticated algorithm.

Our aim is to produce the lowest possible generalisation error, given our training set. This is a problem which has been discussed by [HS90] (under the name of the "proximity problem", though in fact the study is equivalent to training with noise). This study showed generalisation to be adversely affected by the phenomenon of "overfitting", where

an insistence on storing the members of the training set exactly leads to an increase in the generalisation error. Allowing a fraction of errors in the training set improves the generalisation ability, in a reflection of the sacrificial effect discussed earlier; however, in order to calculate the appropriate fraction of errors we must have prior knowledge about the training set.

In dealing in more detail with issues arising from overfitting, Watkin [W93] has distinguished between "Gibbs learning" and "optimal learning" In Gibbs learning we simply attempt to store as many members as possible of the training set without regard to the generalisation abilities of the network. In optimal learning we assume that the training set has been generated by application of some noisy rule. By reconstructing the prototypes and training on them with a rescaled noise parameter we attempt to produce a network that optimally implements this noisy rule. Optimal learning is therefore appropriate for the second situation mentioned above, that of training a perceptron with noise, while for an associative memory we will want to find a Gibbs-type algorithm.

Gibbs Learning

Our first aim is to find a Gibbs algorithm which reduces the training error, and therefore (indirectly) the generalisation error to near-optimal levels on a reasonable timescale. Three different algorithms were investigated. All three were implemented on a network with $N = 20, \alpha = 0.5, m_t \in \{0.1, 0.2, \ldots, 1\}$. A set of $P = \alpha N$ prototypes ξ_i^μ were generated and from each prototype a set of Q noisy examples were obtained. The specific number of training steps varied from algorithm to algorithm. All algorithms were started on the Hebb network for the training ensemble, as the best guess available. After training was completed, a further set of Q examples were generated from the prototypes to get the generalisation error. The algorithms were implemented for $Q = 50$ and $Q = 500$. Results presented are the average over 20 runs for $Q = 50$ and over 10 for $Q = 500$.

Before describing the algorithms we did use, we briefly discuss the use of simulated annealing with a Monte Carlo rule [M&a52] on a landscape $E \equiv \sum_{\mu\nu} \theta(-\lambda_{\mu\nu})$. It appears that this method is simply too slow to be practical. A preliminary attempt to use sim-

ulated annealing to find a marginally stable network for $N = 50, \alpha = 1.5$ had stored no more than 80% of the patterns correctly after 10000 timesteps; we therefore decided to use other algorithms for the, more difficult, case under consideration.

The Perceptron Algorithm

The perceptron algorithm can be stated as

$$J_j(t+1) = J_j(t) + \Delta J_j(t); \quad \Delta J_j(t) = \epsilon \sum_{\mu\nu} \theta[-\lambda^{\mu\nu}(t)] \xi^\mu \eta_j^{\mu\nu}. \tag{3.58}$$

It can be considered to be a gradient descent on the perceptron cost function (3.11). We take the version of the algorithm in which all J_j's are updated simultaneously and all examples are considered together; alternative formulations update the J_j's after each pattern.

The perceptron cost function is guaranteed to converge to a solution that stores all the patterns correctly if such a solution exists. However, above saturation there is no sacrificial effect of the kind we have assumed to be necessary for optimal storage; all unstable patterns are treated equally. We would therefore expect the training and generalisation errors to compare unfavourably with the optimal results. For the network above saturation with random patterns, [GG91] found that the perceptron cost function produced an error rate that was up to twice that of the GD cost function.

The obtained gain in this case is displayed in Figure 3.16. Best results were found for $\epsilon = \mathcal{O}(1/PQ)$. Training was stopped when the minimum energy obtained to date had not changed over the previous 1000 training cycles.

The Maxover Algorithm

The Maxover algorithm (a refined form of which is described in [W95]) is motivated by the concept that if the aim is to store as many of the patterns in the training ensemble as possible, we should train on the pattern that is *closest* to being stable. The algorithm

is thus defined as

$$J_j(t+1) = J_j(t) + \Delta J_j(t); \Delta J_j(t) = \epsilon \xi^{\mu m} \eta_j^{(\mu\nu)m}, \tag{3.59}$$

where $\xi^{\mu m}, \{\eta_j^{(\mu\nu)m}\}$ maximise

$$\lambda^{\mu\nu} \theta[\lambda^{\mu\nu}]. \tag{3.60}$$

The maxover algorithm, in which we train on the pattern of maximum negative stability, is named in analogy with the *minover* algorithm [KM87], in which we train on the pattern with minimum stability. If the system is capable of storing all the patterns, the maxover algorithm should store them; if not, it should store as many of them as possible. Results presented in [W95] for a random training set are extremely promising, even though the inspiration for the rule is empirical rather than theoretical.

The gain is displayed in Figure 3.16. The best value of ϵ was found to be $\epsilon = \mathcal{O}(1/P)$. Since training occurs on one pattern at a time, rather than on $\mathcal{O}(PQ)$ as for the perceptron algorithm, we trained over a longer timescale, stopping the training only when the minimum energy obtained to date had not changed over the previous 10,000 timesteps.

The Cooling Algorithm

The cooling algorithm corresponds to a gradient descent algorithm on the cost function

$$E = -\sum_{\mu\nu} g(\lambda^{\mu\nu}), \quad g(\lambda) = \text{erf}\,(\beta\lambda), \tag{3.61}$$

At high $T \equiv \beta^{-1}$ it trains on all patterns equally; at low T it will only train strongly on those patterns which are marginally stable or unstable. It thus interpolates between the Hebb rule and the Maxover algorithm just presented.

The choice of a cooling schedule is of necessity somewhat empirical; of various schedules tried, an acceptable compromise between training times and low generalisation error was provided by taking T to scale as the square root of the number of training steps taken

to date. The system was trained for 5000 steps, as this provided a timescale comparable to that allowed for the perceptron algorithm. The gain is displayed in Figure 3.16.

Optimal Learning

In the case where we know which example corresponds to which prototype, and have an idea of the noise level, it is possible to employ *optimal learning* ([W93], [W&a93]). In this case it is possible to make a guess at the prototypes; the best guess possible is that each prototype is simply a normalised sum of all the examples generated from it. We can then train on these prototypes with a rescaled noise parameter $\tilde{m} = m/\sqrt{1+1/\tilde{Q}}$, $\tilde{Q} \equiv Qm^2/(1-m^2)$, which reflects both the noise we know to be present and the uncertainty in our estimate of the prototypes. By gradient descent on this cost function we can maximise the expected output overlap for the network.

The optimal learning training algorithm is therefore simply gradient descent on the new prototypes with the new noise parameter, carried out in exactly the same way as for the annealed training. Since training only takes place on P, as opposed to PQ, patterns, this algorithm is much faster than any of the Gibbs algorithms presented above. We implemented optimal training for $Q = 50$ and $Q = 500$. In the high training overlap limit it was hard to obtain good results, given the small range of λ for which the cost function has a significant gradient; however, for $m_t \leq 0.9$ the algorithm was extremely successful.

3.6.4 Results

The gain, $f(m) - m$, for all three Gibbs algorithms is displayed in Figure 3.16a ($Q = 50$) and Figure 3.16b ($Q = 500$). As can be seen, there is a substantial improvement in performance when the number of examples increases.

For the perceptron algorithm, retrieval occurs for $m \geq 0.5$ ($Q = 50$) and $m \geq 0.2$ ($Q = 500$). For the cooling algorithm retrieval occurs for $m \geq 0.5$ at both values of Q. This is to be compared to the result for the perceptron of maximal stability, in which retrieval only occurs for $m \geq 0.67$. In both cases, the Maxover algorithm is not

satisfactory; the gain is negative for all values of m and retrieval does not occur.

Of the three Gibbs algorithms presented, the perceptron seems to be the best available. It is the fastest of the algorithms and produces results at least as good as any other. We have also found in studies on the $\alpha = 1.5$ case (not included here for reasons of space) that, despite the lack of a sacrificial effect, the perceptron algorithm again produces a generalisation ability that is comparable with that of any other algorithm. We therefore conclude that it is the algorithm of choice for the Gibbs learning case.

The performance of the optimal learning algorithm is extremely good in the $Q = 500$ case, and even in the $Q = 50$ case it provides an upper bound to the gain obtained by use of the Gibbs learning algorithms (as we would expect). It is therefore the algorithm of choice when it can be used.

3.6.5 Conclusions

We have investigated five algorithms to try and realise a noise-optimal network. Of these, the annealed-noise algorithm is the most reliable, but it can only be used when the prototypes themselves are known. In the case where preprocessing can be carried out on the data set the method of optimal learning can be applied successfully. However, if no preprocessing is possible, the most successful algorithm (surprisingly) is the perceptron algorithm, which gives the best output gain achieved with Gibbs algorithms, and has the additional virtues of speed, simplicity and robustness.

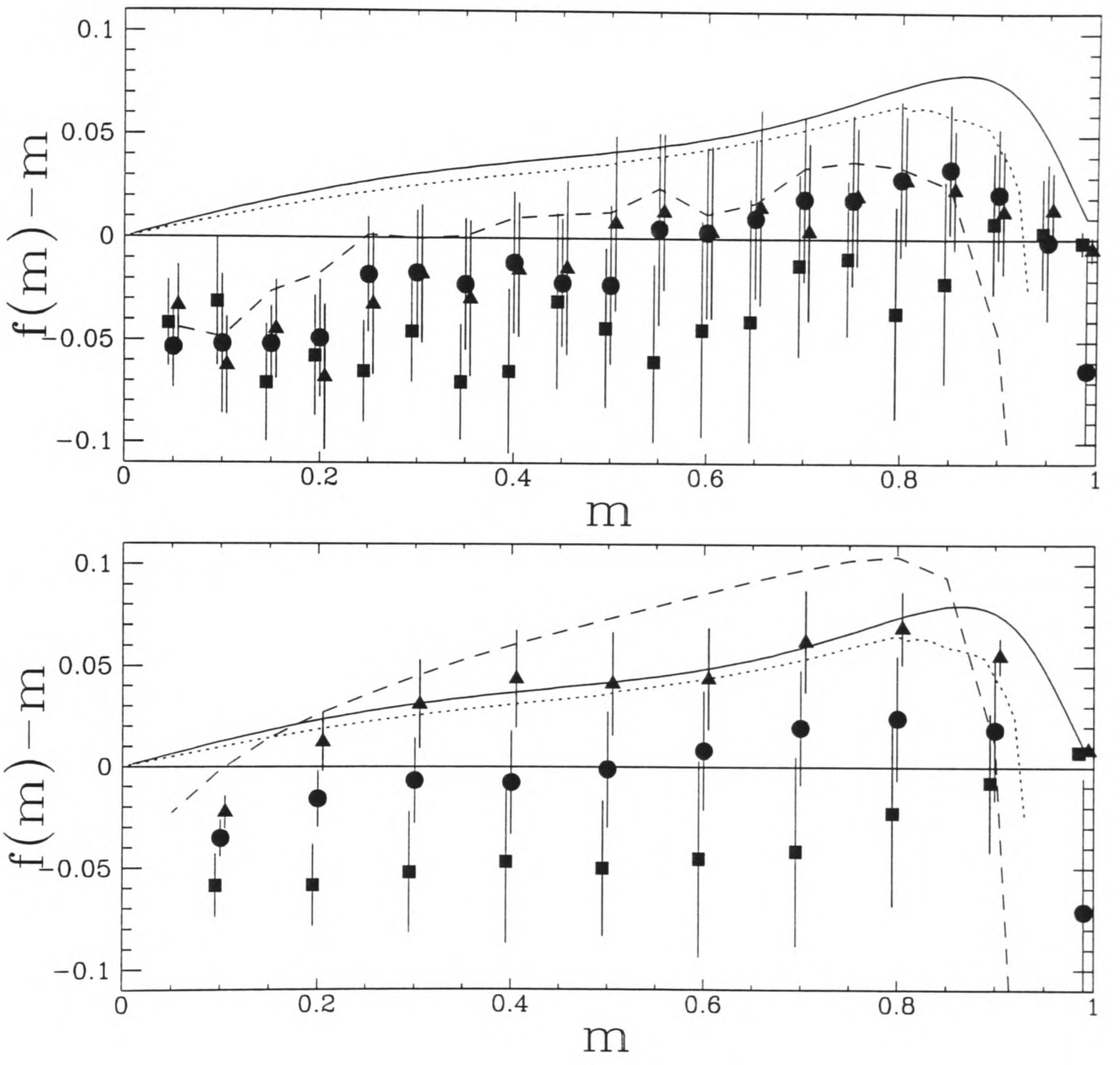


Figure 3.16: The gain, $f(m) - m$: calculated for the noise-optimal network (solid line), obtained by gradient descent on the annealed cost function (dotted line) and the optimal learning function (dashed line), and by simulations using the perceptron (triangles), maxover (squares) and cooling (circles) algorithms. The perceptron and maxover results are slightly offset to ensure clarity in the error bars. The upper figure is for $Q = 50$, the lower for $Q = 500$. The calculated gain and the gain obtained by training on the annealed cost function are reproduced in both figures for ease of comparison.

Part II

Chapter 4

Storing Sequences of Patterns - Two Special Cases

4.1 Introduction

The previous section has dealt with recall as a matter of reaching a stable, static configuration of the network. However, there are many kinds of memories which can only be considered as dynamic objects. The most obvious example is musical tunes, but any memory which evolves in time would fit this description. These memories cannot be treated by the equilibrium calculations of the previous sections; new methods are needed.

Two main approaches have emerged to this problem. One is to assume a distribution of transmission delays in the synaptic interactions, as was done by Kleinfeld [Kl86], Sompolinsky and Kanter [SK86] and Herz *et al* [HLH91], the last of whom have been able to derive many useful results, such as the storage capacity and the phase diagram of their Hebb-like model, from introducing a form of temporal symmetry. The other approach, which we follow here, is to preserve the concept of interactions being instantaneous and to see what kinds of behaviour can be obtained using only the asymmetry of the synaptic matrix. In this case, there will be a conflict between the need to produce a strong overlap with the pattern that is to be recalled and the need to move on to the next pattern in

the sequence. It is hoped that this conflict will produce interesting results.

The approach was pioneered for sparse-coding networks using parallel dynamics by Buhmann and Schulten [BS87] and Nakamura and Nishimori [NN91]. More recently, Coolen and Sherrington (CS) [CS92] have obtained cycling behaviour under both parallel and sequential dynamics in a network without sparse coding, and it is this model that we use as the basis for these investigations.

In this chapter, we perform two further investigations of networks based on the CS model, in both cases introducing a disruptive element to the network and observing the effects on the network's ability to process the sequences. In the first case we introduce correlations between the patterns in a sequence; this is justifiable biologically and enables us to examine how well simple prescriptions for sequence processing deal with the case where it is hard to distinguish between consecutive patterns. In the second case we investigate numerically the effects of noise caused by the network attempting to store an extensive number of sequences.

4.2 Storing Sequences of Correlated Patterns

In this section, we study the effects of introducing correlations between patterns to a network that features competition between pattern storage and sequence processing without using transmission delays. This network displays limit-cycle behaviour in a large region of the phase diagram. From the practical point of view, the most interesting and useful behaviour of the network is when it is in a stable limit-cycle with the individual pattern overlaps taking a wide range of values. For reasons to be given, we expect correlations to exist between patterns in a sequence in a biological network; we are therefore strongly motivated to see if their effects will be constructive or destructive on recall of the sequence as a whole.

The first of the networks that we study is very similar to that analysed (without correlations) by Coolen and Sherrington [CS92], henceforth referred to as CS. Here we show that the effect of the correlations, in combination with the competing aspects of the dynamics, is to tend to "smear out" the network state into one of equal overlap with all the patterns. The second network is more reminiscent of those in [BS87], [NN91]. Here, when we introduce an extra term into the synaptic matrix designed to suppress this symmetric mixture state, we discover that increasing the correlations will increase the range of values that the individual pattern overlaps take and improve the robustness of the sequence processing behaviour. This improvement is greatest for non-zero temperature and for a network intermediate between the two extremes of pure sequence processing and pure Hebbian pattern storage.

We will be looking at patterns whose total "magnetisation" is zero, but which are positively correlated with the patterns that come close to them in the stored sequence. We take the correlation to decrease with the separation of the two patterns in the sequence, and to depend only on that separation. This form of correlation is inspired from several different sources. It is intuitively appealing that the patterns in a remembered sequence will be correlated in a way that depends on their separation, if only because, since things in the real world change continuously, one stimulus will inevitably have some similarity to

the ones immediately before and after it. Secondly, experiments on monkeys performed by Miyashita *et al* [MC88a], [MC88b], [SM91a] have demonstrated that if they are presented with a set sequence of stimuli many times, then subsequently presenting them with a stimulus from the sequence will result in their recalling not just that stimulus, but its neighbours in the sequence. This phenomenon was modelled by Griniasty *et al* [GTA93] assuming that the stored patterns corresponding to the stimuli were uncorrelated but that the neural interactions modify themselves so as to recall the patterns that have become associated with the presented one. Here, we are inspired by this result to consider the case in which the stored patterns are correlated. Finally, many of the results derived in CS depended only on the matrix of correlations of the patterns being Toeplitz. This paper can therefore be regarded in part as a generalisation of this previous work.

This section is organised as follows. First, we define the form of the synaptic matrix and of the correlations. We then we investigate the effects of correlations on the synaptic matrix of CS, and show that the effect of increasing correlations is to decrease both the area of the phase diagram in which limit-cycle behaviour will take place and the ability of the network to distinguish between patterns. Finally we look at a slightly different synaptic matrix, and show that in this case increasing the correlations improves the quality of the limit-cycle behaviour in certain circumstances.

4.2.1 Construction of the Network and Correlations

Synaptic Matrix and Dynamic Laws

In this section we describe our model in more detail and derive the flow equations.

Our model is an Ising spin neural network of N spins $s_i \in \{-1, 1\}$, corresponding to neuron i being at rest or firing respectively. We wish to study a system that has learned a given set of patterns $\xi^\mu \in \{-1, 1\}^N$, $\mu = 1, \ldots, p$, via synapses

$$J_{ij} = \frac{1}{N} \sum_{\mu\rho=1}^{p} \xi_i^\mu A_{\mu\rho} \xi_j^\rho. \tag{4.1}$$

The quantities of interest are the macroscopic overlaps

$$q_\mu(\vec{s}) \equiv \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu s_i \quad (\mu = 1 \ldots p), \tag{4.2}$$

whose evolution we wish to study under parallel (synchronous) dynamics and sequential (asynchronous) dynamics. In both these cases we take $p \ll \sqrt{N}$ as $N \to \infty$.

We first derive the flow equations for a network operating under parallel dynamics, following [B91]. In this case, all of the neurons are updated simultaneously following the rule $P[s_i(t+1)] = \frac{1}{2}(1 + s_i(t+1)\tanh[\beta \sum_j J_{ij} s_j(t)])$. The probability of the network going from state $\vec{S}'$ to to state $\vec{S}$ is thus

$$P(\vec{S}' \to \vec{S}) = \frac{\exp\left[\beta(\sum_{ij} S_i J_{ij} S'_j + \sum_i S_i \theta_i)\right]}{\sum_{\vec{T}} \exp\left[\beta(\sum_{ij} T_i J_{ij} S'_j + \sum_i T_i \theta_i)\right]}, \tag{4.3}$$

where for mathematical convenience we have introduced the external fields $\theta_i = \sum_\mu \theta_\mu \xi_i^\mu$, which will later be taken to zero. Using the definitions in (4.1) and (4.2) we can rewrite the transition probability purely in terms of variables that have pattern rather than site indices, as follows:

$$P_{t+1}(\vec{q}) = \sum_{\vec{q}'} P_t(\vec{q}')\mathcal{N}(\vec{q}')\frac{\exp\left[\beta(\vec{q}\cdot\mathbf{A}\vec{q}' + \vec{q}'\cdot\vec{\theta})\right]}{\sum_{\vec{q}''}\mathcal{N}(\vec{q}'')\exp\left[\beta(\vec{q}''\cdot\mathbf{A}\vec{q}' + \vec{q}''\cdot\vec{\theta})\right]}. \tag{4.4}$$

Here the quantity $\mathcal{N}(\vec{q})$ measures the number of states S which have overlaps $\vec{q}(S) = \vec{q}$. The probability distribution of the overlaps $\vec{q}$ can thus be expressed in terms of the derivative with respect to $\vec{\theta}$ of a function which we call $Z(\vec{q})$:

$$Z(\vec{q}) = \sum_{\vec{q}'} \exp\left[N\beta(\vec{q}\cdot\mathbf{A}\vec{q}' + \vec{q}\cdot\vec{\theta})\right]\mathcal{N}(\vec{q}') \tag{4.5}$$

It remains to calculate Z, and we do this using the partition introduced by van Hemmen *et al* [H&a86], [HK90]. The set of all sites is divided into sets of those sites which have

identical realisations of the pattern bits:

$$(i \leq N) = \cup_\eta I_\eta \quad \text{where} \quad I_\eta = \{i : \vec{\xi}_i = \vec{\eta}\}. \tag{4.6}$$

We can thus write the overlaps $\vec{q}(\vec{S})$ as

$$\vec{q}(\vec{S}) = \frac{1}{N} \sum_\eta |I_{\vec{\eta}}| m_{\vec{\eta}}(\vec{S})\, \vec{\eta}, \tag{4.7}$$

where

$$m_{\vec{\eta}}(\vec{S}) = \frac{1}{|I_{\vec{\eta}}|} \sum_{i \in |I_{\vec{\eta}}|} S_i \tag{4.8}$$

and $|I_{\vec{\eta}}|$ denotes the number of sites in $I_{\vec{\eta}}$. Our expression for Z has now become

$$Z(\vec{q}) = \sum_{\vec{m}} \exp\left[N\beta \sum_{\vec{\eta}} m_{\vec{\eta}} \left(\vec{q}' \cdot \mathbf{A}\vec{\eta} + \vec{\xi} \cdot \vec{\theta}\right)\right] \mathcal{N}(\vec{m}), \tag{4.9}$$

and in the limit $N \to \infty$ we can substitute for $\mathcal{N}(m_{\vec{\eta}})$:

$$\mathcal{N}(m_{\vec{\eta}}) = \exp\left[-\frac{N}{2} \sum_{\vec{\eta}} \left(\left(\frac{|I_{\vec{\eta}}|}{N} + m_{\vec{\eta}}\right) \ln\left(\frac{|I_{\vec{\eta}}|}{N} + m_{\vec{\eta}}\right) + \left(\frac{|I_{\vec{\eta}}|}{N} - m_{\vec{\eta}}\right) \ln\left(\frac{|I_{\vec{\eta}}|}{N} - m_{\vec{\eta}}\right)\right)\right]. \tag{4.10}$$

Now it is possible to turn the m's into continuous variables as $N \to \infty$ and use the saddlepoint method to get the result

$$Z(\vec{q}) = \exp\left[N\beta \sum_{\vec{\eta}} m_{\vec{\eta}}\, \vec{\eta} \cdot \vec{\theta}\right] \tag{4.11}$$

with $m_{\vec{\eta}}$, the saddle-point value, given by

$$m_{\vec{\eta}} = P(\vec{\eta}) \tanh \beta \left[\vec{\eta} \cdot A\vec{q} + \vec{\theta} \cdot \vec{\eta}\right]. \tag{4.12}$$

On taking $dZ/d\theta$, this gives us the final result that

$$\langle F(\vec{q})\rangle_{P_{t+1}} = \langle F(\Gamma(\vec{q}))\rangle_{P_t} \tag{4.13}$$

where

$$\Gamma(\vec{q}) = \sum_{\vec{\eta}} P(\vec{\eta})\eta \tanh\beta\left[\vec{\eta}\cdot A\vec{q} + \vec{\theta}\cdot\vec{\eta}\right]. \tag{4.14}$$

and the probability $P(\vec{\eta})$ is

$$P(\vec{\eta}) \equiv \lim_{N\to\infty}\frac{|I_{\vec{\eta}}|}{N} = \lim_{N\to\infty}\frac{1}{N}\sum_i \delta[\vec{\xi}_i - \vec{\xi}], \quad \vec{\xi}_i \equiv (\xi_i^1 \ldots \xi_i^p). \tag{4.15}$$

On letting the field $\theta \to 0$ we obtain the simple result that the evolution in time of the q_μ's in the thermodynamic limit is governed by the set of coupled non-linear mappings:

$$\vec{q}_{t+1} = \langle\vec{\xi}\tanh[\beta\vec{\xi}\cdot\mathbf{A}\vec{q}_t]\rangle_{\vec{\xi}}, \tag{4.16}$$

where the average is defined as

$$\langle\Phi(\vec{\xi})\rangle_{\vec{\xi}} \equiv \sum_{\vec{\xi}\in\{-1,1\}^p} P(\vec{\xi})\Phi(\vec{\xi}). \tag{4.17}$$

Under sequential dynamics, the individual neurons are updated one at a time in a random order, according to the rule $P(s_i) = \frac{1}{2}(1 + s_i\tanh[\beta\sum_j J_{ij}s_j])$. If we take the duration of a single iteration to scale as $\frac{1}{N}$, then in the thermodynamic limit the behaviour of the macroscopic overlaps q_μ is governed by the set of coupled non-linear differential equations [H&a86],[CR88]:

$$\frac{d}{dt}\vec{q} = \langle\vec{\xi}\tanh[\beta\vec{\xi}\cdot\mathbf{A}\vec{q}]\rangle_{\vec{\xi}} - \vec{q}. \tag{4.18}$$

We merely state this result, as our studies will concentrate on parallel dynamics.

The fixed-points of either dynamics will satisfy

$$q_\mu = \langle \xi_\mu \tanh[\beta \vec{\xi} \cdot \mathbf{A}\vec{q}] \rangle_{\vec{\xi}}. \tag{4.19}$$

The forms of $\mathbf{A}$ under investigation here are

$$\begin{aligned} (i) \quad A_{\mu\rho} &\equiv \nu\delta_{\mu\rho} + (1-\nu)S_{\mu\rho} \\ (ii) \quad A_{\mu\rho} &\equiv \nu\delta_{\mu\rho} + \tfrac{1}{2}\sqrt{1-\nu^2}S_{\mu\rho} - \tfrac{1}{2}\sqrt{1-\nu^2}S^+_{\mu\rho} \\ \text{where} \quad S_{\mu\rho} &\equiv \delta_{\mu,\rho+1} \quad (\mu : \mathrm{mod}\, p) \end{aligned} \tag{4.20}$$

The parameter $\nu \in [0,1]$ allows us to interpolate smoothly between the simple Hopfield model ($\nu = 1$) and sequence processing models ($\nu = 0$). The forms of the coefficients of $\mathbf{S}$ and $\mathbf{S}^+$ are chosen to obtain a value for the critical temperature for the existence of non-trivial fixed-point solutions that is independent of the value of ν. The effect of the first form of $\mathbf{A}$ on a system in a pure state μ will be to move it towards the state $\mu+1$; the effect of the second form of $\mathbf{A}$ on the same system will be to move it towards a mixture of $\mu+1$ and the inverse of $\mu-1$. We therefore refer to the first $\mathbf{A}$ as "forward-propagating" and to the second as "double-propagating"; to distinguish them we call the latter $\mathbf{A}'$ hereafter.

For the forward-propagating case storing uncorrelated patterns and with parallel dynamics, CS obtained the phase diagram shown in Figure 1, with similar behaviour for other p values. It contains six distinct regions (note that our notation differs from theirs).

- For $T > 1$ there is a paramagnetic phase (P), where the trivial fixed-point will be the only fixed-point of the dynamics.

- At $T_{ps} = 1 \geq T \geq T_{sc'}(\nu)$ there is a phase (S) in which the symmetric fixed-point $\vec{q} = q^*(1,1,1,\ldots,1)$ is the only attractor of the dynamics.

- At $T_{sc'}(\nu) \geq T \geq T_{cr}(\nu)$ the system exhibits limit-cycle behaviour, whose period depends on the value of ν. There are two regions within this temperature range,

C and C′; within the region C′ the symmetric fixed-point solution is also stable, whereas within C it is unstable. Their boundary is $T_{c'c}(\nu)$.

- For $T < T_{cr}(\nu)$ the system is in a "retrieval" phase (R), in which there are stable non-symmetric fixed-points and recall of individual patterns is possible.

- There is also a corresponding phase R′ ($T < T_{cr'}$) in which the system displays limit-cycle behaviour with period p, independent of the value of ν.

Each of the critical temperature lines is symmetric under $\nu \rightarrow 1 - \nu$.

We might expect that the structure of the phase diagram for parallel dynamics with the matrix **A** and correlated patterns will be qualitatively the same as the structure of this phase diagram, but with altered parameters. In the case of the matrix **A′** it is more difficult to predict the detailed structure of the phase diagram *a priori*. We would expect the phases P, S, C, R to be observable. However, we would not expect to see an R′ phase, since this phase arises as a result of the symmetry between ν and $1 - \nu$ under parallel dynamics for the matrix **A**, and this symmetry does not exist for **A′**. These expectations are, in fact, borne out.

Pattern Distribution and Correlation

We take $\langle \xi^\mu \rangle = 0$, and define the correlation matrix $C_{\mu\nu} \equiv \langle \xi^\mu \xi^\nu \rangle$. The patterns are not spatially correlated within themselves or biased; the correlations only enter as a relationship between the bits of different patterns on the same site i. We note that $C_{\mu\nu}$ is symmetric by definition, and require it to have the following additional properties:

i) $C_{\rho\lambda} \doteq C_\Delta$, where $\Delta \equiv |\rho - \lambda|$ (**C** is translation-invariant in the space of patterns). **C** will therefore commute with both **A** and **A′**.

ii) $C_{\rho\lambda}$ is positive in all its entries.

iii) $C_{\rho\lambda}$ decreases monotonically from $C_{\rho\rho}\,(\equiv 1)$ to $C_{\rho,\rho+p/2}$, and thereafter (by the symmetry requirement) increases monotonically from $C_{\rho,\rho+p/2}$ to $C_{\rho,\rho-1}$.

When performing numerical simulations, and in order to have a single parameter characterisation of the whole distribution, we will assume that the probability distri-

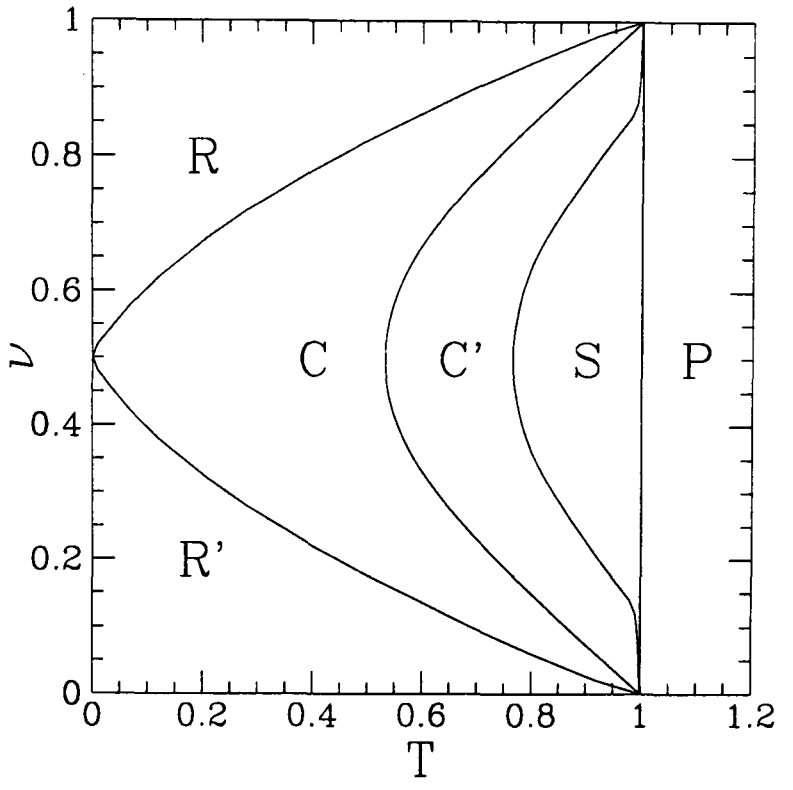


Figure 4.1: The phase diagram as determined by CS for $p = 10$. The labels refer to the phases described in the text.

bution $P(\vec{\xi})$ is of a form corresponding to the thermal distribution of a periodic p-site 1-dimensional nearest neighbour Ising ferromagnet,

$$P(\vec{\xi}) \sim e^{J(\sum_{\nu=1}^{p-1} \xi_\nu \xi_{\nu+1} + \xi_p \xi_1)}. \tag{4.21}$$

This gives a correlation matrix of the form

$$C_\Delta = \frac{{t_J}^\Delta + {t_J}^{p-\Delta}}{1 + {t_J}^p}, \tag{4.22}$$

with $t_J \equiv \tanh(J) \in [-1, 1]$, which satisfies the conditions above, provided that $J \geq 0$.

The parameter we will generally use as a measure of the correlations is $c \equiv \langle \xi_\mu \xi_{\mu+1} \rangle_{\vec{\xi}}$, the correlation between nearest neighbours.

Eigenvectors and Eigenvalues

As in CS, we define a set of vectors $\{|n\rangle\}$ which form an eigen-basis in p-dimensional space for the matrices **A**, **S** and **C**:

$$|n\rangle \equiv \left(\hat{e}_1^n, \ldots, \hat{e}_p^n\right), \quad n = 0, \ldots, p-1 \quad \hat{e}_\lambda^n \equiv \frac{1}{\sqrt{p}} e^{2\pi i n\lambda/p}. \tag{4.23}$$

These have the following properties:

$$\langle n|m\rangle = \delta_{nm}; \quad \mathbf{S}|n\rangle = e^{-2\pi i n/p}|n\rangle \equiv s_n|n\rangle, \quad \mathbf{A}|n\rangle \equiv a_n|n\rangle,$$
$$\mathbf{C}|n\rangle \equiv c_n|n\rangle, \quad c_n = \textstyle\sum_{\lambda=0}^{p-1} \cos(2\pi n\lambda/p) C_{1,\lambda+1}. \tag{4.24}$$

For both the **A**'s to be studied here, a_n is real only if $n = 0$ or (if p is even) $n = \frac{p}{2}$. With regard to **C**, we note that $c_n = c_{p-n}$, that c_n is always real, and that $\max_n c_n = c_0 = \sum_\lambda C_{\rho\lambda}$. When p is even, $\min_n C_n = c_{p/2} = \sum_\lambda (-1)^\lambda C_{\rho\lambda}$.

We can write the symmetric fixed-point in this basis as

$$\vec{q} \equiv \vec{q}^{+} = q^{+}|0\rangle. \tag{4.25}$$

If we take the correlations to be of the Ising ferromagnet type, the c_n's become:

$$c_n = \frac{(1 - t_J{}^p)(1 - t_J{}^2)}{(1 + t_J{}^p)(1 + t_J{}^2 - 2t_J \cos[\frac{2n\pi}{p}])} \tag{4.26}$$

In this case, it can be seen that the eigenvalues c_n also obey the condition $c_{n+1} < c_n$, so long as $t_J > 0$ and $n + 1 < p/2$. In the limit $t_J \to 0$, $c_n \to 1 \,\forall\, n$; in the limit $t_J \to 1$, $c_0 \to p$ and $c_n \to 0 \,\forall n \neq 0$.

Eigenvalues of any matrix with respect to the basis $\{|n\rangle\}$ are denoted by subscripts in the Roman alphabet; references to a particular pattern are denoted by Greek subscripts. We will also find it useful to define

$$\vec{\xi}|n\rangle \equiv x_n. \tag{4.27}$$

4.2.2 The Forward-Propagating A-matrix

In this section we take

$$A_{\mu\rho} \equiv \nu\delta_{\mu\rho} + (1-\nu)S_{\mu\rho}, \tag{4.28}$$

so

$$\mathbf{A}|n\rangle = a_n|n\rangle = [\nu + (1-\nu)e^{-2\pi in/p}]|n\rangle. \tag{4.29}$$

This is the case described for uncorrelated patterns by CS. For $p = 2$, this produces a symmetric synaptic matrix which may be analysed completely. We then perform as much analysis as possible on the $p > 2$ case, and follow this up with numerical simulations for $p = 10$.

The Toy Problem: the Symmetric Case $p = 2$

For $p = 2$, the probability distribution has the simple form

$$P[\xi^1, \xi^2] = \frac{1}{4}(1 + \xi^1\xi^2 c). \tag{4.30}$$

If we introduce the variables $z^{\pm} \equiv q_1 \pm q_2$, the dynamic equations decouple, yielding the fixed-point equations

$$z^+_{fp} = (1+c)\tanh[\beta z^+_{fp}], \quad z^-_{fp} = (1-c)\tanh[\beta(2\nu-1)z^-_{fp}]. \tag{4.31}$$

which in turn imply the critical temperatures below which non-zero values of $z^{\pm}$ are possible:

$$T_c(z^+) = 1 + c; \quad T_c(z^-) = (1-c)(2\nu-1), \tag{4.32}$$

We see that z^- will only be a non-zero fixed point if $\nu > 0.5$. However, if $\nu < 0.5$ and $T < T'_c(z^-) = (1-c)(1-2\nu)$, then under parallel dynamics z^- can oscillate between $\pm z^{-\prime}_{fp} = (1-c)\tanh[\beta(1-2\nu)z^{-\prime}_{fp}]$. We can thus identify four of the six phases from the general phase diagram:

- For $T > T_c(z^+) = (1+c)$ we are in the paramagnetic phase P.

- For $1 + c > T > (1 - c)(|2\nu - 1|)$, z^+ is non-zero and z^- is zero, and we are in the symmetric fixed-point phase S. There is no C phase.

- For $T < (1 - c)(|2\nu - 1|)$ we are in either R or R′, depending on whether ν is greater or less than 0.5 respectively.

The effect of increasing the correlation c for $p = 2$ is, therefore, to decrease the size of the regions P, R and R′. We would also expect to observe this at higher values of p. However, this toy model casts no light on the effect of increasing correlations on the relative sizes of regions S, C' and C.

Analytic Results for $p > 2$, Parallel Dynamics

We first attempt to locate the critical temperature T_{ps} for a transition from the paramagnetic phase to the symmetric fixed-point phase. An upper bound on this temperature is given by the critical temperature T_c for the existence of non-zero solutions of the fixed-point equation (4.19). Using the methods of CS, and the fact that $\max_n |a_n|$ and $\max_n c_n$ both occur at $n = 0$, we find that this temperature obeys

$$T \leq \frac{1}{2} \max_n \left[\langle n|\mathbf{C}|n\rangle + \langle n|\mathbf{A}^+\mathbf{C}\mathbf{A}|n\rangle \right] = \sum_{\Delta} C_{\Delta} = c_0. \tag{4.33}$$

The fact that this maximum occurs at $n = 0$ also indicates that the first type of non-trivial fixed point to become a solution of the dynamics is the symmetric fixed point, emphasising that the temperature we are locating here is indeed T_{ps}.

A lower bound on T_{ps} can be obtained by stability and bifurcation analysis of the trivial fixed-point. The condition for a fixed-point $\vec{q}$ to bifurcate under parallel dynamics is:

$$\det |1 - \beta \mathbf{\Gamma}(\vec{q})\mathbf{A}| = 0, \quad \text{where } \Gamma_{\rho\lambda} \equiv \langle \xi_\rho \xi_\lambda (1 - \tanh^2[\beta \vec{\xi}\mathbf{A} \cdot \vec{q}]) \rangle_\xi. \tag{4.34}$$

The condition for local stability of a fixed-point under parallel dynamics is

$$\max_{\vec{x}} \frac{\vec{x} \cdot \mathbf{A}^\dagger \mathbf{\Gamma}^2(\vec{q})\mathbf{A}\vec{x}}{\vec{x}^2} < T^2; \tag{4.35}$$

Both these equations are to be solved simultaneously with the fixed-point equation (4.19).

For the trivial fixed-point, $\mathbf{\Gamma}(\vec{q}) = \mathbf{\Gamma}(\vec{0}) = \mathbf{C}$. The bifurcation equation therefore becomes $\det|1 - \beta\mathbf{C}\mathbf{A}| = 0$, or

$$\exists \vec{y} : \mathbf{C}\mathbf{A}\vec{y} = T\vec{y}. \tag{4.36}$$

Because T is real, we require the left-hand side of this equation to be real. This means that $\vec{y} = |n\rangle$ where a_n is real. The only values of n that satisfy this are $n = 0$ and, if p is

even, $n = p/2$. These two solutions have associated with them the critical temperatures

$$\begin{aligned} T_c(\vec{q} = q|0\rangle) &= \sum_{\Delta} C_{\Delta} \equiv c_0; \\ T_c(\vec{q} = q|p/2\rangle) &= (2\nu - 1)\sum_{\Delta}(-1)^{\Delta} C_{\Delta} \equiv (2\nu - 1)c_{p/2} < c_0. \end{aligned} \tag{4.37}$$

On solving the stability equations for the trivial fixed-point, we obtain, using the fact that $\mathbf{\Gamma}(\vec{0}) = \mathbf{C}$, the condition:

$$\max_n |a_n|^2 c_n < T^2 \quad \Rightarrow \quad T_c(\vec{q} = \vec{0} \text{ becomes unstable}) = c_0. \tag{4.38}$$

We can conclude that at $T = c_0$ the trivial fixed-point becomes unstable and bifurcates continuously to the symmetric fixed-point. This fixes T_{ps}.

We now attempt to determine $T_{c'c}$. A stability analysis may be performed on the symmetric fixed-point in the same way as on the trivial fixed-point. In this case we obtain the following form for $\mathbf{\Gamma}$:

$$\Gamma_{\rho\sigma}(\vec{q}^{+}) = \langle \xi_\rho \xi_\sigma (1 - \tanh^2(\beta q^{+} \frac{1}{\sqrt{p}} \sum_{\nu} \xi_\nu)) \rangle_{\vec{\xi}}. \tag{4.39}$$

It is evident that $\Gamma_{\rho\sigma}(\vec{q}^{+})$ will depend only on $|\rho - \sigma|$. It is also possible to express $\mathbf{\Gamma}(\vec{q}^{+})$ in terms of the basis $|n\rangle$:

$$\mathbf{\Gamma}(\vec{q}^{+}) = \sum_{n} \gamma_n(\vec{q}^{+}) |n\rangle\langle n| \text{ with} \tag{4.40}$$

$$\gamma_n(\vec{q}^{+}) \equiv \langle |x_n|^2 (1 - \tanh^2(\beta q^{+} x_0)) \rangle_{\vec{\xi}} = \sum_{\lambda} \cos(\frac{2\pi n \lambda}{p}) \langle \xi_p \xi_\lambda (1 - \tanh^2(\beta q^{+} x_0)) \rangle_{\vec{\xi}}, \tag{4.41}$$

where $x_n \equiv \vec{\xi}|n\rangle$. Since $\mathbf{\Gamma}(\vec{q}^{+})$ is symmetric, all the γ_n's will be real, as (4.41) shows explicitly. It can also be seen that, except in the case where $\beta \to \infty$, we have $0 < \gamma_n < c_n \forall n$.

This gives the following condition for $\vec{q}^{+}$ to be stable:

$$\max_n \left[\nu^2 + (1 - \nu)^2 + 2\nu(1 - \nu)\cos(2\pi n/p)\right]^{1/2} \gamma_n < T; \tag{4.42}$$

This condition can be evaluated numerically, taking into account that the symmetric fixed-point cannot become unstable due to fluctuations in the direction $n = 0$. We discover that the maximising value of n under these conditions is $n = 1$, for the Ising-type correlations that we are using.

Bifurcation analysis for the symmetric fixed-point gives the following condition for a continuous bifurcation to take place:

$$\gamma_n \left[\nu + (1-\nu)e^{2\pi i n/p}\right] \langle n|y\rangle = T\langle n|y\rangle, \tag{4.43}$$

As before, any bifurcation must take place in a direction $|n\rangle$, where a_n is real. For the case of the symmetric fixed-point the bifurcation may not be in the direction $n = 0$. Therefore, if p is odd there will be no bifurcation. If p is even, a bifurcation can take place in the direction of the anti-symmetric fixed-point, $\vec{q} = \vec{q}^- = q^-|p/2\rangle$ at the temperature

$$T_b = \gamma_{p/2}(2\nu - 1). \tag{4.44}$$

However, numerical evaluation of this temperature shows it to be always below the temperature at which the symmetric fixed-point becomes unstable. This demonstrates that when this fixed-point becomes unstable it will not go continuously to another fixed-point solution but will instead go to a limit cycle. We therefore identify as $T_{c'c}$ the temperature at which the symmetric fixed-point becomes unstable, as given by the left-hand side of (4.42) with $n = 1$.

Finally, we attempt to locate T_{cr}. It is not possible to do this explicitly, but the analysis of CS suggests that, for p even, T_{cr} will be bounded from above by the temperature at which the pure antisymmetric state, $\vec{q}^-$, becomes a possible fixed-point of the dynamics. This temperature is given by $T_{\vec{q}^-} = (2\nu - 1)c_{p/2}$. Since a recall state must have a non-zero overlap with the antisymmetric state $|p/2\rangle$, it makes sense that this temperature should form an upper bound on the temperature below which recall states are found.

If $\nu < 0.5$ we would expect to find period-p limit cycles below $T = (1 - 2\nu)c_{p/2}$.

We summarise the results of this section in Figure 2a and 2b, showing respectively

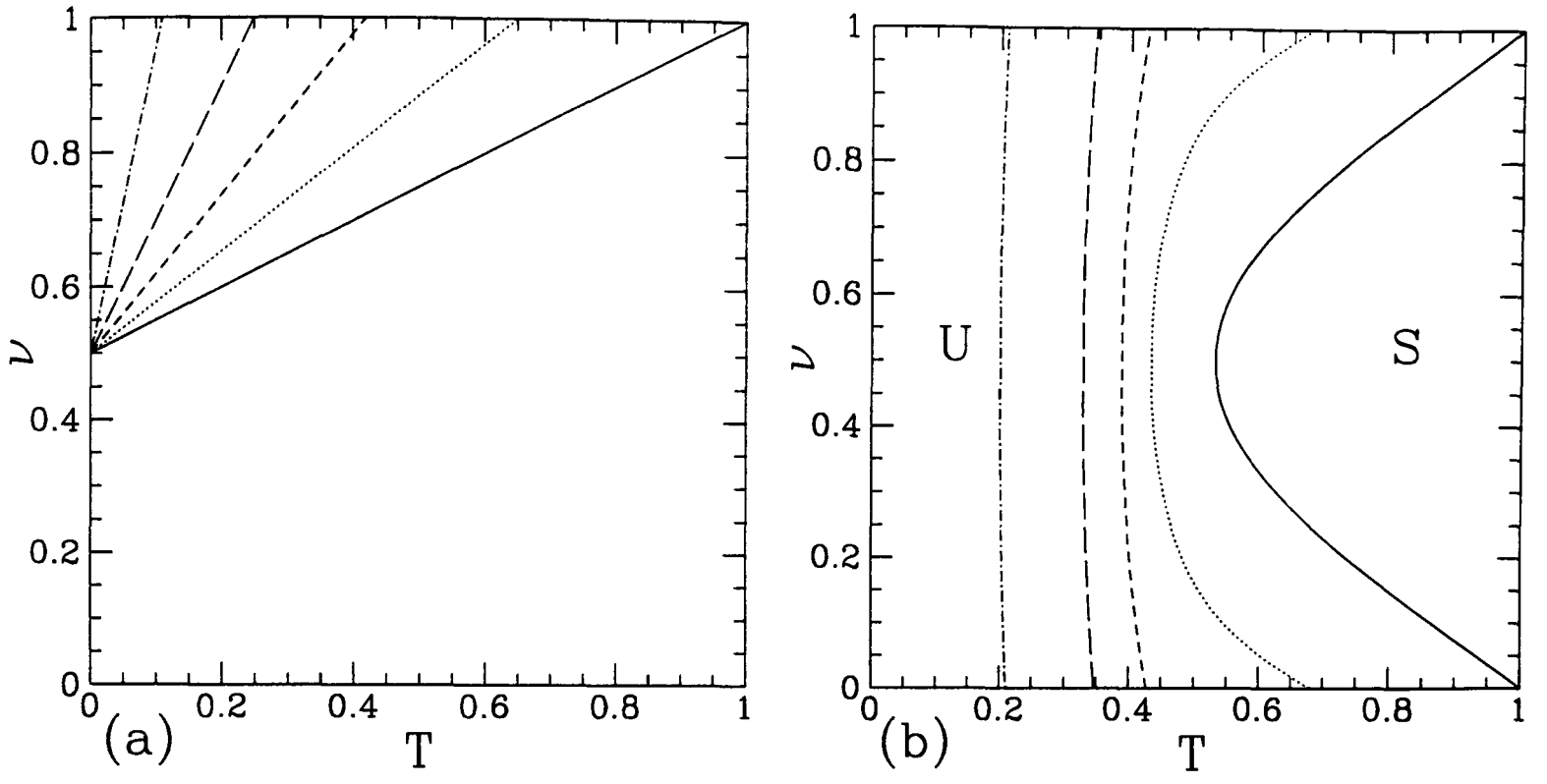


Figure 4.2: (a): The temperature $T_{\vec{q}^-}$ at which the pure anti-symmetric fixed-point $\vec{q} = q^-|p/2\rangle$ becomes a fixed-point for the dynamics, for $p = 10$ and nearest-neighbour correlations of $c = 0$ (solid line), 0.2 (dotted line), 0.4 (short-dashed line), 0.6 (long-dashed line), 0.8 (dot-dashed line). (b): The temperature $T_{c'c}$, dividing parameter space into the corresponding regions in which the symmetric fixed-point is stable (S) and unstable (U) under parallel dynamics.

$T_{\vec{q}^-}$ and $T_{c'c}$.

Numerical Results for $p = 10$, Parallel Dynamics

The dynamical laws described above, equation (4.16), were implemented numerically. These were not simulations on a finite-sized network but direct numerical iterations of the dynamics (4.16). Qualitatively, the phase diagram structure is as in Figure 1, but with altered parameters.

To simplify the discussion of the results we describe the system in terms of three variables, Q_s, Q_d and the period P, where

$$Q_s(t) \equiv \frac{1}{p}\sum_{\mu} q_\mu(t), \quad Q_d(t) \equiv |\vec{q}(t) - Q_s(t)\sqrt{p}|0\rangle|, \tag{4.45}$$

and Q_s, Q_d denote the asymptotic values of these variables, $Q_s(\infty)$ and $Q_d(\infty)$. Along with the stability calculation of (4.42), these are all we need to determine the area of the phase diagram that the system is in, as follows:

- In the paramagnetic region P, Q_s and Q_d are both zero.
- In the symmetric fixed-point region S, Q_s is non-zero and Q_d is zero.
- In the limit-cycle regions C and C′, Q_d and P are non-zero. These regions are distinguished by the stability of the symmetric solution (calculated analytically in (4.42)).
- In the retrieval region R, the system is at a fixed point with $Q_d \neq 0$.
- In the period-p limit-cycle region R′, Q_d is non-zero and the period $P = p$.

The system was started in a "pure state" ($q_\mu = 1, q_{\nu\neq\mu} = C_{\mu\nu}$) and run under the parallel dynamics (4.16). A small number of runs were undertaken from random starting positions. These seemed to confirm the result of CS that the asymptotic values of $|Q_s|$, Q_d are independent of the initial q_μ's, indicating an attractor basin covering all random and pure start states for each value of the control parameters (p, ν, T). As in CS, these implementations do not distinguish between the C and C′ regions, but yield

cyclical solutions in both cases from the above start states. The C and C′ regions are distinguishable by implementations started from a state close to the symmetric fixed-point.

In all cases, the quantity used to measure the correlations was c, the correlation between neighbouring patterns in a sequence.

The results can be summarised as follows.

- The value of T_{cr} decreases rapidly with increasing c. The boundary between R, R′ and C is marked not just by an abrupt change in the period P, as would be expected, but also by an abrupt drop in the asymptotic value of Q_d.

- Within the C, C′ phases, as c increases at constant (ν, T), the asymptotic value of Q_d in general decreases. At low temperatures, however, after entering the phase, Q_d shows a slight increase with increasing c before abruptly dropping again. Any such sharp change in Q_d does not always correspond to changes in the period. If $T <\sim 0.15$ then the period P changes slightly with increasing c, first increasing and then decreasing (for $\nu > 0.5$) or first decreasing and then increasing (for $\nu < 0.5$).

- Increasing c also decreases $T_{sc'}$. As T approaches $T_{sc'}$ from below, Q_d goes continuously to zero.

- Within the S phase, as T increases at constant (c, ν), Q_s goes continuously to zero.

- T_{ps} is confirmed as c_0.

Figure 3, showing asymptotic values of Q_d, P as a function of c for various values of T and $\nu = 0.2$, displays most of these features. There is a slight increase in Q_d as c is increased near $c \sim 0.3$, for $T = 0$. Inspection of the raw q_μ's shows that within this range the effect of increasing the correlation is to increase q_μ for the pattern with the second-highest overlap and decrease q_μ for the pattern with the second-lowest overlap. This will increase the standard deviation of the distribution of overlaps, which is exactly what Q_d measures. This increase in Q_d therefore does not correspond to the form of improved retrieval behaviour discussed in the Introduction.

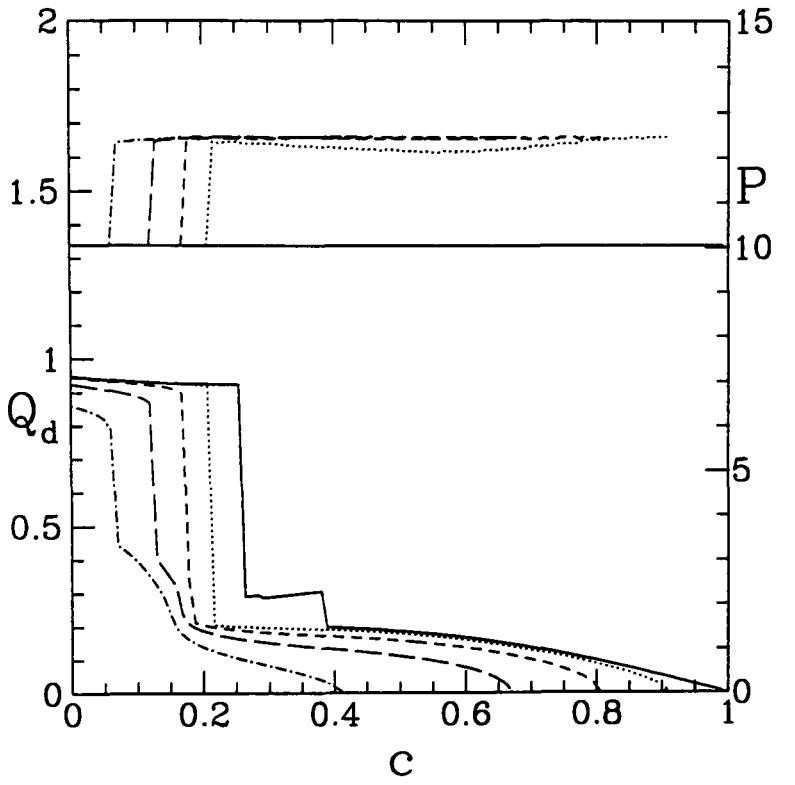


Figure 4.3: Q_d (lower lines, left-hand scale) and P (upper lines, right-hand scale), plotted against increasing c, for $p = 10$, $\nu = 0.2$, $T \in \{0$ (solid line), 0.1 (dotted line), 0.2 (short-dashed line), 0.3 (long-dashed line), 0.4 (dot-dashed line) $\}$. The system was started from a state of overlap 1 with one pattern and run under parallel dynamics. The sharp transition in Q_d and P marks the boundary of the R region.

We conclude this section by presenting the values of $T_{sc'}$, T_{cr} obtained by this numerical study (Figures 4a and 4b). These are to be compared with the figures obtained analytically, Figures 2a and 2b. We obtain close agreement at high values of c between the boundary of the S phase and the stability of the symmetric fixed-point, indicating that as correlations increase the size of the C′ region decreases rapidly. At lower values of c it is possible for limit-cycle behaviour to persist even when the symmetric fixed-point is stable.

The upper bound obtained analytically for existence of the retrieval phase R can be seen to be extremely loose. In fact, we find numerically that there is no R or R′ phase at all for $c > 0.336$.

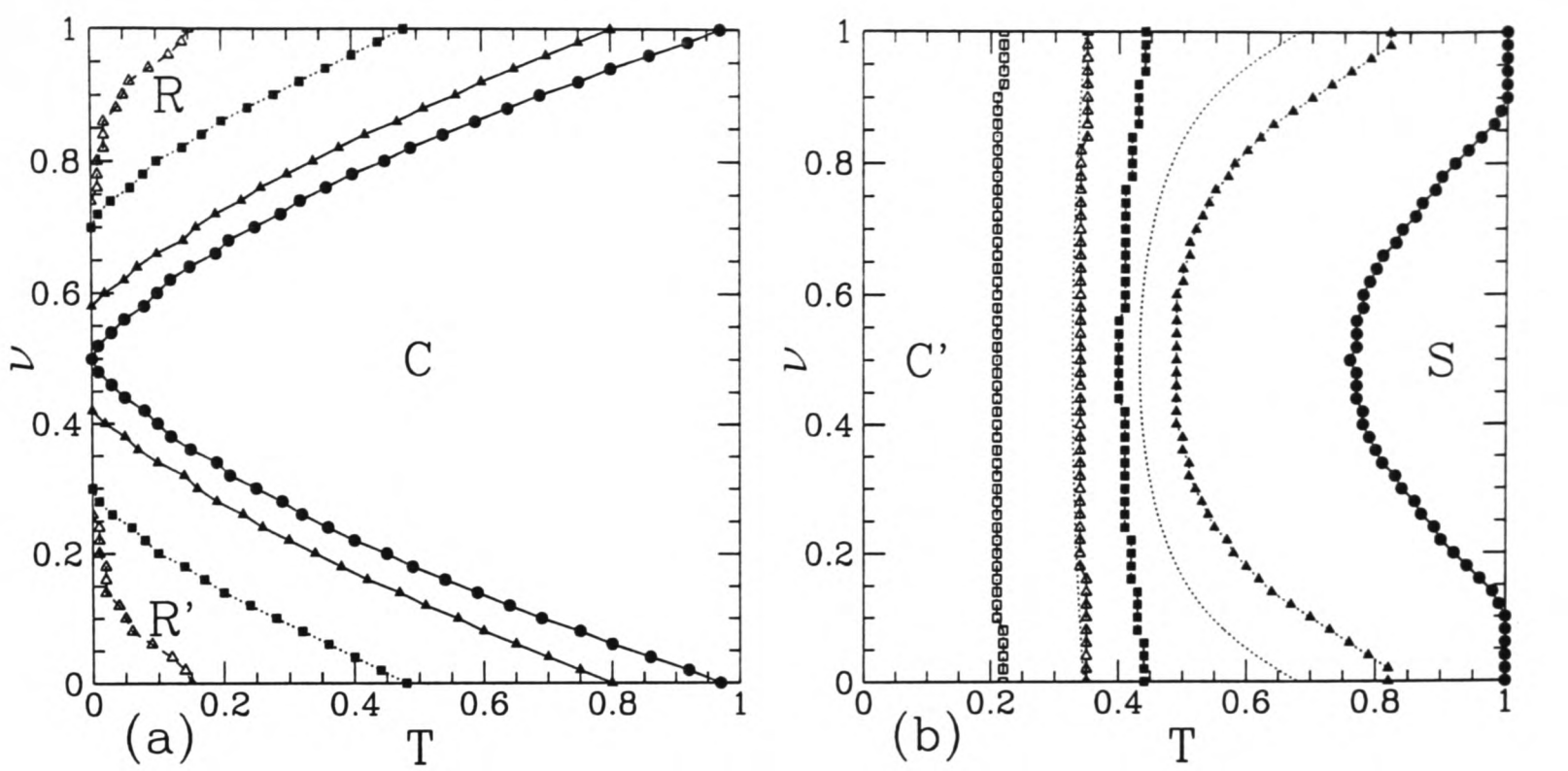


Figure 4.4: (a) T_{cr}, found numerically for parallel dynamics with $p = 10$ and $c = 0, 0.1, 0.2, 0.3$ (going from right to left). (b) $T_{sc'}$, found numerically for parallel dynamics with $p = 10$ and $c = 0, 0.2, 0.4, 0.6, 0.8$ (going from right to left). The dotted lines are the corresponding lines on which the symmetric fixed-point becomes unstable for the same values of c, reproduced from Figure 2b for ease of comparison.

Sequential Dynamics

As found by CS in the case of uncorrelated patterns, sequential dynamics can also lead to limit-cycle behaviour in the thermodynamic limit $N \to \infty$. However, numerical implementations of this system under sequential dynamics are far more expensive of computer time than implementations of parallel dynamics. We therefore restrict this discussion to mentioning a few brief points of interest.

The symmetry between ν and $1 - \nu$ that exists for parallel dynamics does not exist for sequential dynamics. We therefore do not expect an R′ region of the phase diagram to exist.

The stability properties of the trivial fixed-point are the same under both parallel and sequential dynamics.

The stability properties of the symmetric fixed-point, however, show an interesting difference. We might expect that as c increases the temperature at which the symmetric

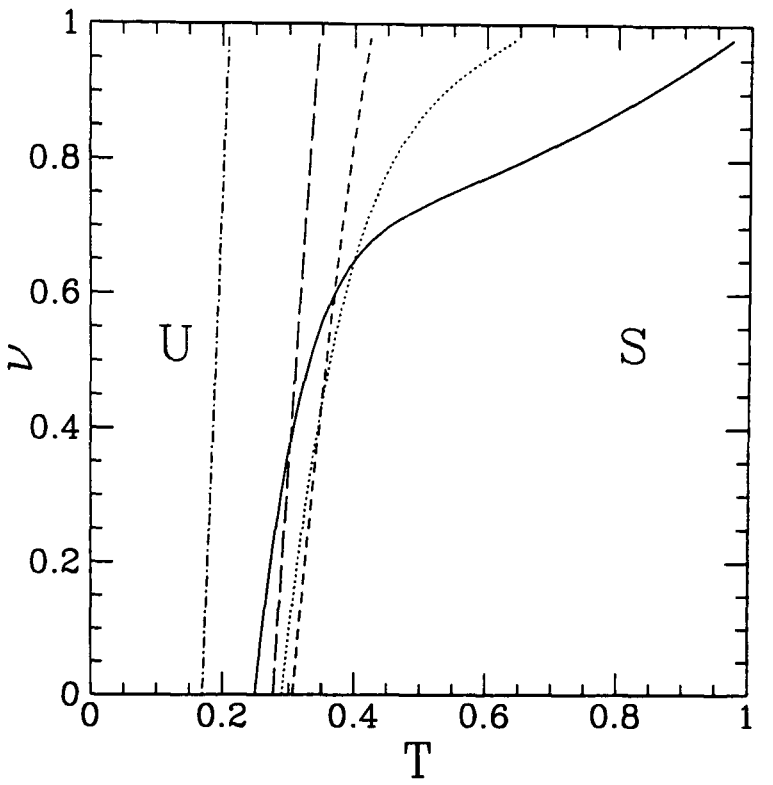


Figure 4.5: The regions in which the symmetric fixed-point is stable (S) and unstable (U) under sequential dynamics for $p = 10$, with $c = 0$ (solid line), 0.2 (dotted line), 0.4 (short-dashed line), 0.6 (long-dashed line), 0.8 (dot-dashed line).

fixed-point becomes unstable would decrease monotonically, as for parallel dynamics. This is the case for p odd. For p even and low values of ν, on the other hand, as c is increased this critical temperature will first increase and then decrease. So there exists a small range of temperatures for which increasing the correlation will bring the system from the S phase to the C phase and then back to the S phase again. This is displayed in Figure 5. Numerical implementations (starting from both the pure state and a state near the symmetric fixed-point) confirm this result.

4.2.3 The Double-Propagating A-Matrix

We now switch to look at the matrix $\mathbf{A}' \equiv \nu\mathbf{1} + \frac{1}{2}\sqrt{1-\nu^2}(\mathbf{S} - \mathbf{S}^\dagger)$, in the hope that the antisymmetric aspects of this matrix will reduce the strength of the symmetric fixed point in the case of correlated patterns. This matrix is very similar to that studied in [BS87], [NN91], [NNS90]. These papers, however, looked at the case of very sparse coding so that no neuron was firing in more than one pattern. Here we are taking non-sparse coding ($\frac{1}{N}\sum_i \xi_i^\mu = 0\,\forall\mu$) and a different form of correlations between patterns.

We expect the phases P, S, C and R to exist for $\mathbf{A}'$, as they existed for $\mathbf{A}$; however, we do not expect the boundaries to lie in the same positions, nor do we expect the detailed structure of the phase diagram necessarily to be the same. Since, for $\mathbf{A}'$, the $p = 2$ case gives the Hopfield network, it does not provide a useful toy model. Our ability to treat the model analytically is therefore confined to calculating the stability of the trivial and symmetric fixed-points and calculating the temperatures at which the symmetric and anti-symmetric fixed-points become solutions of the dynamics.

We confine our discussion to parallel dynamics for the reasons given in the previous section.

Analytic Results for $p > 2$, Parallel Dynamics

The eigenvalues of $\mathbf{A}'$ with respect to the basis $|n\rangle$ are

$$\mathbf{A}'|n\rangle = a'_n|n\rangle, \quad a'_n = \nu + i\sqrt{1-\nu^2}\sin(\frac{2\pi n}{p}). \tag{4.46}$$

We first attempt to discover the region in which the trivial fixed-point will be the only attractor of the dynamics, using stability and bifurcation analysis. As before, a bifurcation from the trivial fixed-point can only be in the direction of the symmetric or anti-symmetric fixed-point. The bifurcation temperatures are respectively

$$T_{bif} = \nu c_0,\ \nu c_{p/2}. \tag{4.47}$$

These provide a lower bound on the paramagnetic phase.

Next we look at the stability properties of the trivial fixed-point. Since $\mathbf{\Gamma}(\vec{0}) = \mathbf{C}$. for stability of the trivial fixed-point we require

$$\max_n |a'_n| c_n < T. \tag{4.48}$$

This leads to the following result:

- For $\nu < \nu_{crit}(t_J) = \frac{1-t_J}{\sqrt{1+{t_J}^2}}$, the condition for stability is

$$T > T_{stab} = \left[\frac{(1-{t_J}^p)(1-t_J^2)}{(1+{t_J}^p)(1+{t_J}^2-2t_J k)}\right]\left[\nu^2+(1-\nu^2)[1-k^2]\right]^{1/2}, \tag{4.49}$$

  where $k \equiv \cos(\frac{2\pi n}{p}) = \frac{2t_J}{(1-\nu^2)(1+{t_J}^2)}$, this giving the n that maximises the LHS of the inequality (4.48).

- For $\nu > \nu_{crit}$,the condition for stability is

$$T > c_0 \nu. \tag{4.50}$$

  and the maximising value of n is 0, corresponding to an instability in the direction of the symmetric fixed point.

In the uncorrelated case, $\nu_{crit} = 1$ and the trivial fixed-point becomes unstable at $T = 1$ for parallel dynamics. The maximising value of n will only be zero at $\nu = 1$. For any other value of ν the trivial fixed-point becomes unstable in a direction other than the direction of the symmetric fixed-point. In the correlated case, for $\nu > \nu_{crit}$ the trivial fixed-point will go continuously to the symmetric fixed-point as T is lowered. When $\nu < \nu_{crit}$ the trivial fixed-point becomes unstable in some other direction. The results of evaluating the above condition for $p = 10$ are shown in figure 6a.

We now look at the stability properties of the symmetric fixed-point. The condition

(4.35) for stability results in the final condition

$$\max_{n\neq 0}\left[\nu^2+(1-\nu^2)\sin^2(\frac{2\pi n}{p})\right]\gamma_n^2(\vec{q}^+)<T^2, \tag{4.51}$$

where

$$\gamma_n(\vec{q}^+)=\textstyle\sum_\lambda \cos(\frac{2\pi n\lambda}{p})\langle\xi_0\xi_\lambda(1-\tanh^2[\beta q^+\nu x_0])\rangle$$
$$q^+=\langle x_0\tanh[\beta\nu q^+x_0]\rangle \tag{4.52}$$

This too can be solved numerically and the results are displayed in Figure 6b. Interestingly, although the lines of stability for the symmetric and trivial fixed-points coincide for some (low) values of ν and c, the point at which they separate is not the point at which the trivial fixed-point becomes unstable in the direction of the symmetric fixed-point. This implies that there an area of the phase diagram in which the symmetric fixed-point is a stable solution of the dynamics but the trivial fixed-point will not become unstable in its direction, in contrast to the behaviour of the previous model.

Finally, we obtain an upper bound on the region in which recall behaviour is possible by calculating the temperature at which the anti-symmetric fixed-point becomes a possible fixed-point solution to the dynamics. This is found to be $\nu c_{p/2}$.

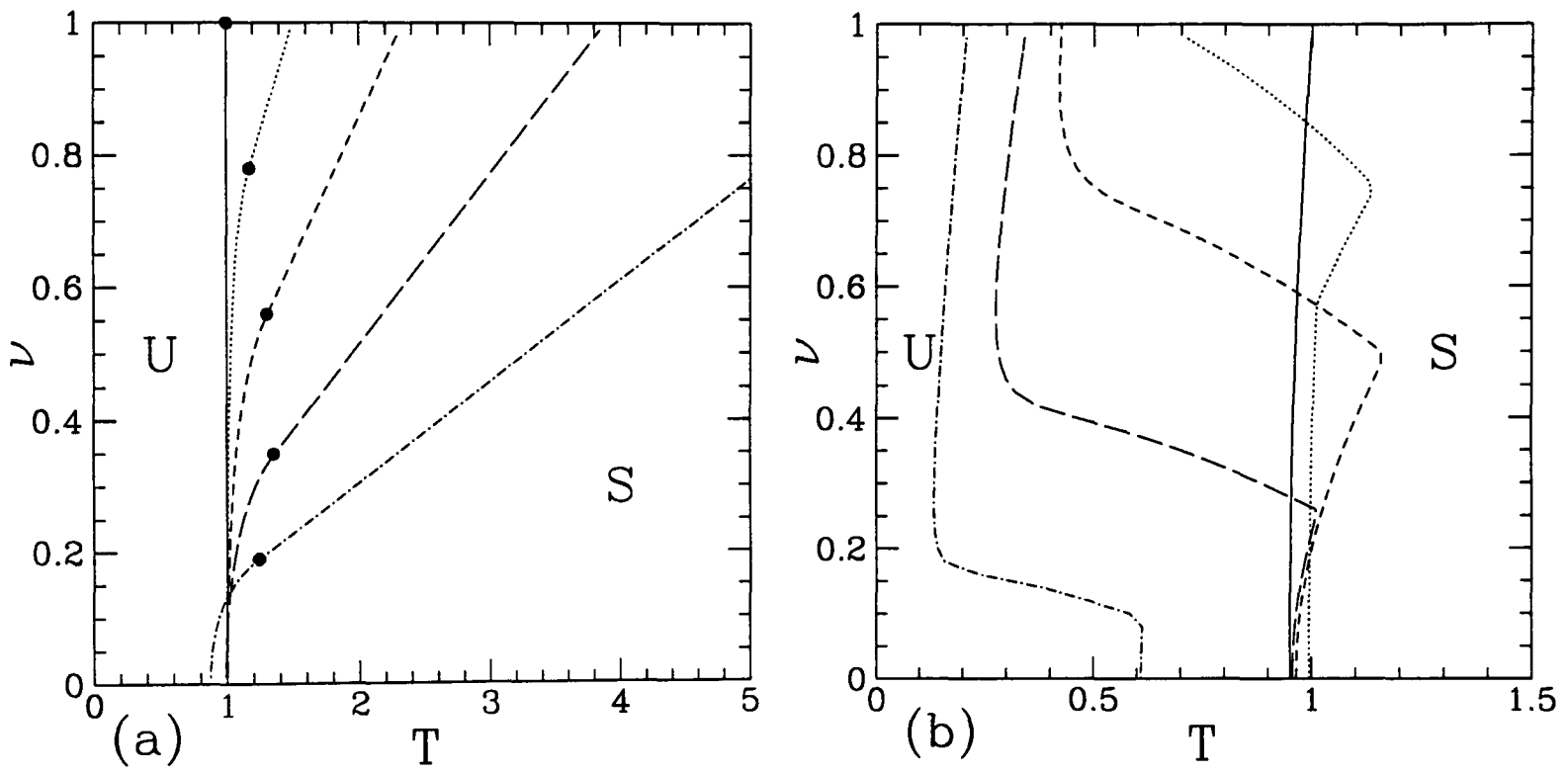


Figure 4.6: Regions where the trivial fixed-point (a) and the symmetric fixed-point (b) are stable (S) and unstable (U), found analytically for $p = 10$ under parallel dynamics, for $c = 0$ (solid line), 0.2 (dotted line), 0.4 (short-dashed line), 0.6 (long-dashed line), 0.8 (dot-dashed line). In (a), the dots mark the point (ν_{crit}) where the trivial fixed-point goes from becoming unstable in the direction of the symmetric fixed-point (above the dot) to becoming unstable in some other direction (below the dot).

Numerical Results for $p = 10$, Parallel Dynamics

We now present results obtained by numerical iteration of the macroscopic laws (4.16). As before, the behaviour of the network could be classified as R, C, S or P. As expected, we found no R′ region; there also proved to be no C′ region. The phase diagram is displayed below (Figure 8). However, in this case the behaviour in the C region was strikingly different from the behaviour in the previous section. We first describe this behaviour and then describe the phase diagram.

- At low T there were multiple stable asymptotic values for Q_d (shown in Figure 7a) when the system was started from random initial states. This contrasts with the previous case where the final values of Q_d, Q_s and P were independent of the initial states.

  As T increases, there is a decrease in the range of c for which this multiplicity of states exists, and a decrease in the number of states at any given c. Even at low T, as Figure 7a illustrates, this multiplicity of states does not prevent the emergence of the general trends, as follows.

- At fixed values of T and ν, increasing c no longer causes a monotonic decrease in Q_d. Instead, as c increases, Q_d initially increases too. At high values of T or low values of ν it peaks twice, with the second peak at a considerably higher value of Q_d than that obtained for $c = 0$, before decreasing (as we would expect) to 0. This behaviour is illustrated in Figure 7b, which also shows how this "two-hump" effect is more pronounced at high T. The "two-hump" effect is mirrored in the behaviour of the period P.

- The period displays an abrupt change in behaviour with increasing c. For $c < c_{crit}$, P is roughly constant. Above c_{crit}, P increases sharply and thereafter varies slightly with increasing c, first increasing and then decreasing.

  c_{crit} as defined here corresponds almost exactly to the value of c at which the minimal value of Q_d between the two humps first becomes 0. Around this value of c there is

an oscillation in the envelope of the q_μ's as a function of time. At values of c much higher or lower than this, the envelope of the q_μ's is relatively constant with time.

We see that there are two separate limit-cycle solutions to the dynamics, one of which exists at high c and one of which exists at low c. The amplitudes of these solutions depend on c and T; at high T the ranges of c within which these amplitudes are non-zero do not overlap, while at low T an overlap occurs, causing the oscillation in the envelope of the q_μ's with the corresponding beat frequency.

As stated in the Introduction, the most interesting behaviour of the network is a stable cycle with a large amplitude of oscillation. The increase of Q_d with c prompts us to investigate whether increasing the correlations genuinely increases the extent to which the network distinguishes between patterns in the sequence. As was pointed out in the previous section, Q_d can be increased by increasing the span of the distribution of q_μ's, defined as $(\max_\mu q_\mu - \min_\mu q_\mu)$, but it can also be increased by increasing the tendency of the q_μ's to "cluster" at their extreme values. If we look at the span as well as Q_d, we can say that for two distributions of q_μ with the same Q_d, the better performance is given by the system with a greater span. We do not display explicit results here for reasons of space, but the general findings are as follows.

For low c or low T, increasing c will not increase the span. If an increase in Q_d occurs it is therefore due to increased clustering. We also observe that at low c, Q_s is also small, implying that the q_μ's oscillate around 0 and that the system is not distinguishing between those patterns with large negative q_μ and those with large positive q_μ. For intermediate values of (ν, T, c), however, we find that an increase in Q_d is usually matched by an increase in the span. This is coupled with a non-zero value for Q_s. In practice, this means that we have limit-cycle behaviour in which one pattern has $q_\mu \sim 0.8$, another has $q_\mu \sim 0$, and the others are spread between them without too much clustering. In this region we can say that, paradoxically, the introduction of correlations genuinely causes the network to distinguish better between patterns.

We conclude by presenting the phase diagram for the network with $p = 10$. We do so in three parts. Figure 8a shows the border of the retrieval phase R as a function of

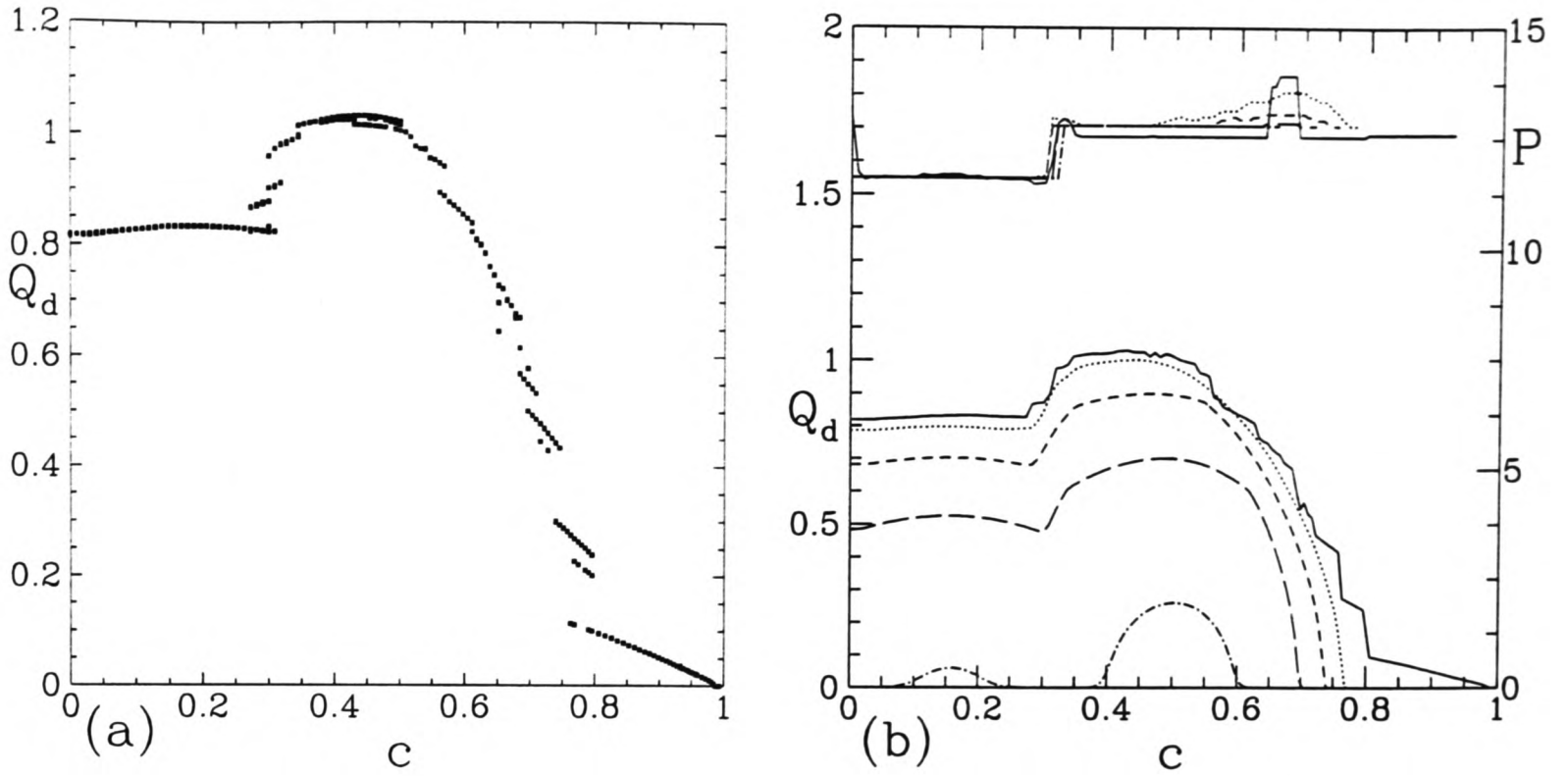


Figure 4.7: (a): Asymptotic values of Q_d for systems started from random states for $p = 10$, $\nu = 0.2$, $T = 0.01$. For $0.25 < c < 0.8$ there are, in general, multiple possible asymptotic values of Q_d. However, the tendency is for Q_d to increase with c until $c \sim 0.4$ and then decrease. (b) Q_d (lower curves) and P (upper curves) for $p = 10$, $\nu = 0.2$, $T = 0$ (solid line), 0.25 (dotted line), 0.5 (short-dashed line), 0.75 (long-dashed line), 1 (dot-dashed line). The curves for P are curtailed when $Q_d \to 0$ as then there is no periodic behaviour. The discontinuities in P for $T = 0$ are due to there being multiple asymptotic values of P; we were unable to find a way to ensure that the system ended up with one value rather than another. The curves displayed here were obtained by starting the system in the pure state; the results obtained by starting the system in random states display the same features.

c, ν for various values of T. As can be seen, for each value of T there is a ν at which c_{cr} increases abruptly. Above this ν we find there is no limit-cycle behaviour if the network is started in the pure state; this is a region of Hopfield–like behaviour.

Figure.8b shows the border of the limit-cycle region as a function of $(c. \nu)$ for various values of T. Figure 9 shows the region of stability of the symmetric fixed-point as a function of (ν, T) for various values of c, for ease of comparison with Figure 6b. The agreement between these two figures is good. Figure 10 is the overall phase diagram for the network with $c = 0.4, p = 10$.

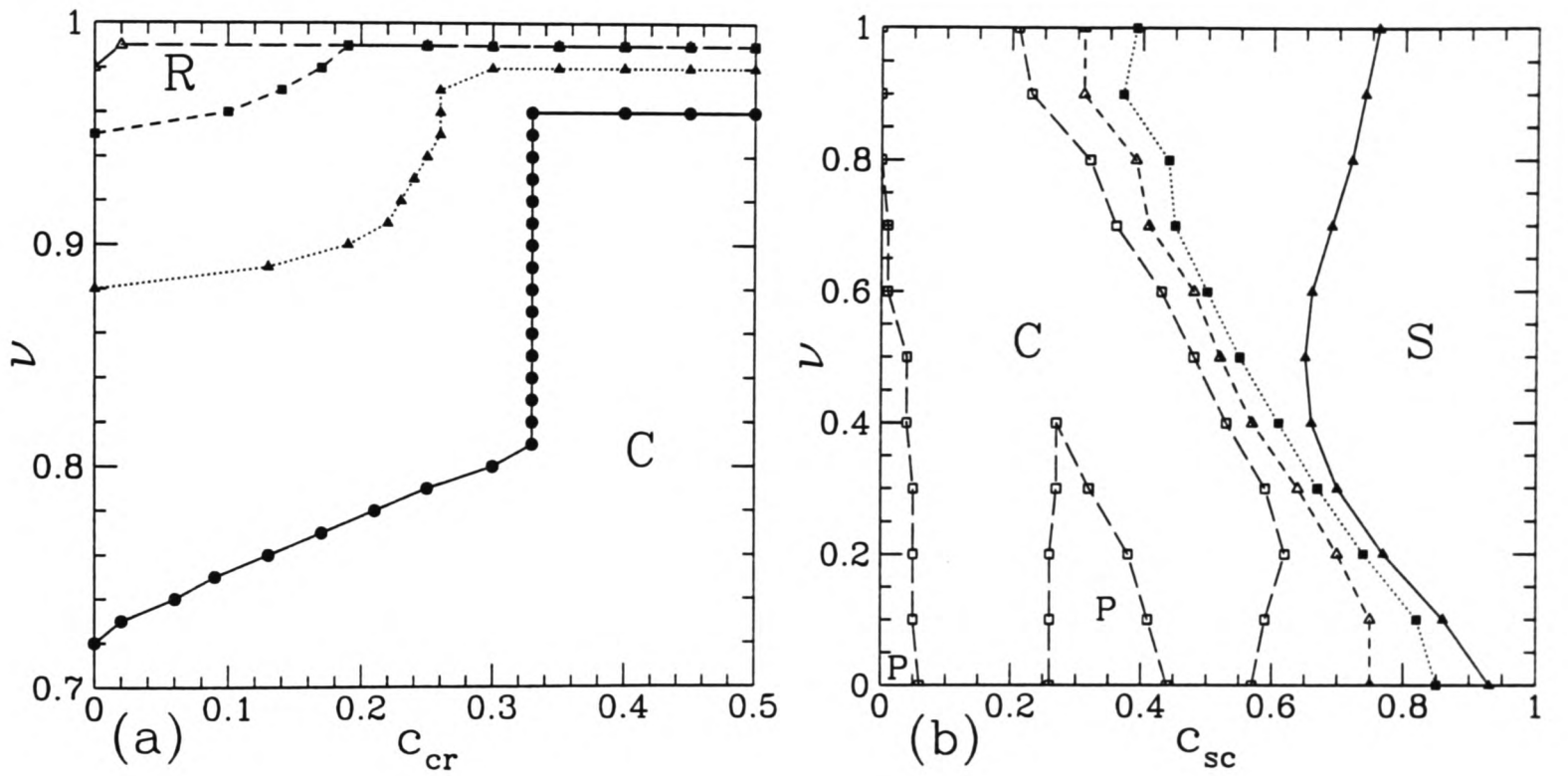


Figure 4.8: (a) The retrieval region R of the phase diagram for $p = 10$ and parallel dynamics. The lines of constant temperature divide the graph into the region where the pure state will go to a non-symmetric, Hopfield-like fixed point (R) and the region where it will go to a limit-cycle (C). The lines are at (going from C to R) $T = 0, 0.25, 0.5, 0.75$. (b) The limit-cycle region C for $p = 10$, shown as a function of (ν, c) for $T = 0.25, 0.5, 0.75, 1.0$. In the region S the system goes to the symmetric fixed-point; in the region P, for $T = 1$, it goes to the trivial fixed-point from a pure initial state. At $T = 0$ there is no S region: $c_{sc} = 1$.

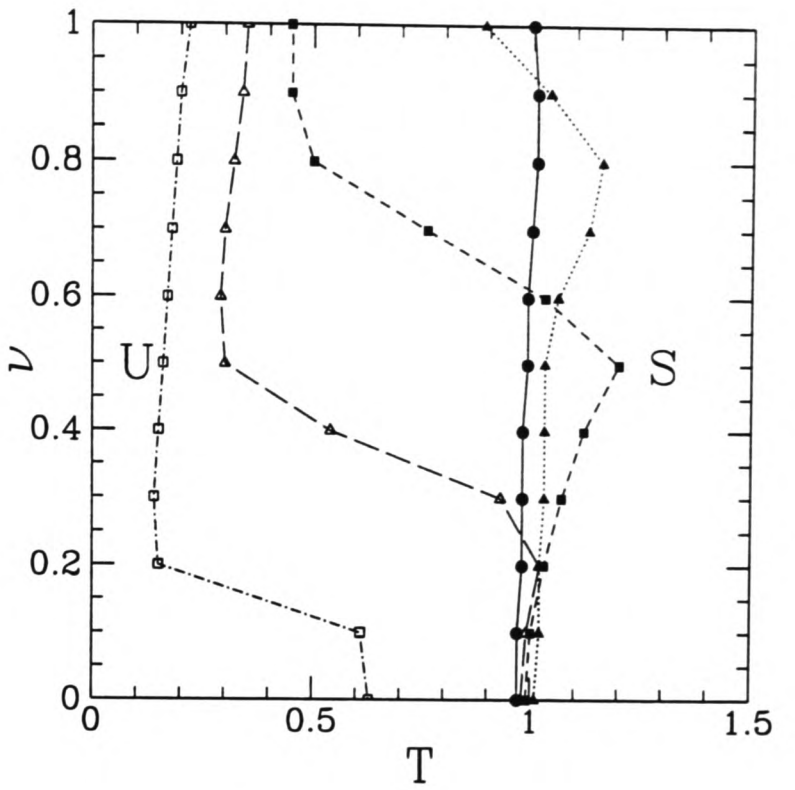


Figure 4.9: The regions in which the symmetric fixed-point is stable (S) and unstable (U), obtained numerically for $p = 10$ and shown as a function of (ν, T) for $c = 0$ (circles), 0.2 (full triangles), 0.4 (full squares), 0.6 (hollow triangles), 0.8 (hollow squares). This is to be compared to Figure 6b.

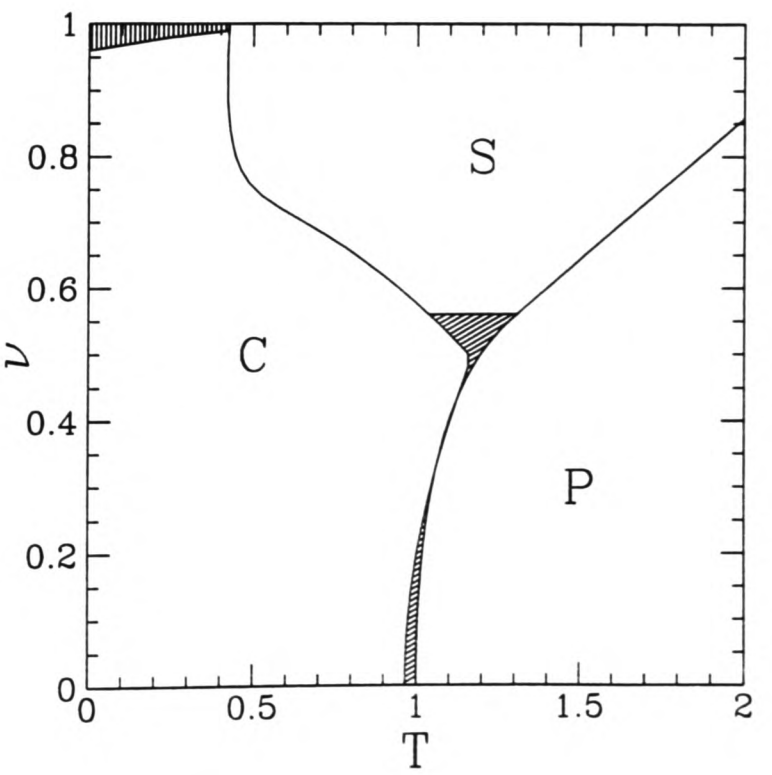


Figure 4.10: The phase diagram for the network for $p = 10, c = 0.4$, showing the retrieval region R (vertical shading, in the top left-hand corner of the diagram), the limit-cycle region C, the region of coexistence of the symmetric fixed-point and limit cycles C′ (diagonal shading, near $T = 1$), the symmetric fixed-point region S, and the paramagnetic region P.

4.2.4 Conclusions

For a certain form of synaptic matrix in a neural network designed to store sequences of patterns, it is possible to improve limit-cycle behaviour by introducing moderate correlations between the stored patterns. The behaviour is improved in that the difference between the largest and smallest overlap with any pattern is increased; the value of the largest overlap is increased, and the absolute value of the smallest overlap is moved closer to 0; and the system distinguishes adequately between consecutive patterns in the sequence. This behaviour depends strongly on the structure of the synaptic matrix, and in the other case investigated the introduction of correlations served only to increase the tendency of the network to go to a state in which it does not distinguish between patterns at all.

The above comments apply to parallel dynamics, although the small amount of research done into sequential dynamics suggests that broadly the same results will be obtained in this case. The research has also been restricted to investigation of binary neuron neural networks, and to patterns with overall magnetisation of zero. We are still, therefore, a long way from any kind of biological realism.

4.3 Storing Extensive Numbers of Sequences of Patterns

4.3.1 Introduction

In the previous section, we studied the disruption caused to a sequence of patterns by introducing correlations between the patterns. In this section we instead study numerically the effects of requiring the system to store an extensive number of patterns in the form of sequences.

To formalise this: we wish to study a system that has learned a given set of patterns $\xi^{\mu\nu} \in \{-1,1\}^N$. The patterns are grouped in sequences, each of length l. The first index of $\xi^{\mu\nu}$ refers to the sequence a pattern is a member of and the second index refers to its position in the sequence, so that, for example, $\xi^{1\,2}$ refers to the second pattern in sequence 1. We define p to be the total number of patterns, so that we will have $\frac{p}{l}$ sequences in total, and define the storage capacity $\alpha \equiv \frac{p}{N}$. The synaptic matrix is taken to be

$$J_{ij} = \frac{1}{N} \sum_{\mu} \left[\sum_{\nu\rho=1}^{l} \xi_i^{\mu\nu} A_{\nu\rho} \xi_j^{\mu\rho} \right]. \tag{4.53}$$

In other words, we have a Hebb-type sum of the synaptic matrices appropriate to store each sequence. We take $\mathbf{A}$ to be

$$A_{\mu\rho} \equiv S_{\mu\rho} \tag{4.54}$$

This is the extreme case ($\nu = 0$) of the matrix studied in section 4.2.2, though without correlations. Note that, for $l = 1$, $\mathbf{A}$ reduces to the Hebb rule.

The quantities of interest to us are the overlaps of the system state with each pattern, defined as usual:

$$q_{\mu\nu}(\vec{s}) \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu\nu} s_i. \tag{4.55}$$

The behaviour of interest to us will be referred to as "useful limit-cycle behaviour". By this we mean that if at a given time t the largest overlap of any pattern with the system

state is q_m with pattern $\xi^{\mu\nu}$, then at time $t+t'$, the largest overlap will be $q_m \pm \mathcal{O}(\frac{1}{\sqrt{N}})$ with pattern $\xi^{\mu,\nu+t'}(\nu + t' \bmod l)$. This is the equivalent, for sequences, of the standard recall criterion in the Hopfield model. Sequential dynamics only displays limit-cycle behaviour in the limit $N \to \infty$ [WSC95], so we concentrate here on parallel dynamics, which in this case are defined by

$$P(S_i[t+1]) = \tfrac{1}{2}(1 + S_i[t+1]\tanh(\beta \sum_{\mu\nu} \xi_i^{\mu\nu} q_{\mu\nu+1}[t])), \tag{4.56}$$

with all spins being updated simultaneously.

Our investigation will be entirely numerical. A system using a matrix $\mathbf{A}$ to store a single sequence of extensive length has been studied analytically [ZLPW94] using a Gaussian form for the intrinsic noise, and has been found to have a storage capacity of $\alpha_c = 0.28$. Though the use of Gaussian intrinsic noise is questionable, the method is directly applicable to any network of the form studied here storing sequences of length greater than 2, and yields the same value of α_c for all l.

4.3.2 Results

We modelled the system described above in networks of 2000 and 5000 neurons. Due to the separability of the dynamics (4.56) it was not necessary to store the entire synaptic matrix of the network; instead the spins were updated according to the rule 4.56. The network was started in the state $\xi^{1\,1}$ and run under these dynamics for up to 100 time-steps. If, by that time, limit-cycle behaviour had not broken down according to the criteria below, the run was marked as a success.

Limit-cycle behaviour existed if the largest overlap of the network at time t was with pattern $\xi^{1,t \bmod l}$. In order for an overlap to count as distinctively the largest overlap, it had to exceed the second-largest overlap by $\mathcal{O}\frac{1}{\sqrt{N}}$. Limit-cycle behaviour could break down in one of two ways, as follows:

- A breakdown of periodic behaviour within the sequence $\xi^{1\,\mu}$: the largest overlap was still with some pattern $\xi^{1\,\mu}$ but we no longer have $\mu = t \bmod l$. This will eventually

result in the system going to a symmetric fixed-point with respect to the sequence $\xi^{1\,\mu}$ in which the overlaps with all of the patterns in that sequence were the same.

- A breakdown of retrieval of the sequence $\xi^{1\,\mu}$: the largest overlap was with some pattern $\xi^{\nu\,\mu}$, $\nu \neq 1$.

These will be referred to hereafter as "Criteria A". We might expect that, above the storage capacity, the system would become a Hebb-rule network for the symmetric fixed-points of the various sequences; however, for the uncorrelated sequences under consideration this proved not to be the case.

Seven runs were made for different pattern realisations on a network of size $N = 2000$. Each run encompassed three sequence lengths ($l = 2, 5, 20$), forty different values of α ($\alpha = 0.01, 0.02, \ldots, 0.4$) and eleven values of T ($T = 0, 0.1, \ldots, 1$). One run was also implemented for a $N = 5000$ network, with sequence lengths ($l = 2, 5, 50$) and the same range of T, α as for the $N = 2000$ case. The following parameters were measured:

- Whether or not the sequence was successfully retrieved according to criteria A above.
- The form of the breakdown, if any.
- The final value of the overlap with the desired pattern.
- The final value of the largest overlap, no matter which pattern it was with.
- The final value of the largest overlap of the system with any pattern not in the sequence $\xi^{1\,\mu}$.
- The final value of the difference between the largest overlap with a pattern in the sequence $\xi^{1\,\mu}$ and the smallest overlap with a pattern in that sequence.
- The number of time-steps before limit-cycle behaviour broke down.

For $l \neq 2$ all the parameters basically told the same story. We therefore present the final value of the overlap with the desired pattern as our measure of recall success for $l = 5, 20, N = 2000$ (figure 4.3.2) and for $l = 5, 50, N = 5000$ (figure 4.3.2). For these

values of l, the storage capacity is $\alpha_c \sim 0.27$. A breakdown in sequence retrieval was always observed when the value of the desired overlap was less than 0.9. Breakdowns were almost always due initially to a transition to the symmetric fixed-point within the sequence, but once this transition had occurred the symmetric overlap itself deteriorated until the initial sequence was in no way distinguished from the other sequences. We do not present error bars, but these were calculated and behaved as usual – small in the regions away from the transition and large in the transition region.

Strikingly, the storage capacity found by these simulations is almost exactly twice the storage capacity of the Hopfield model. An explanation for this can be found in the nature of the synapses. The Hopfield model is, by definition, symmetric: $J_{ij} = J_{ji}$. However, for the matrix under consideration here, J_{ij} and J_{ji} are entirely uncorrelated if $l > 2$. The network is thus storing twice as much information, and this is reflected in the doubling of the storage capacity.

An interesting feature of the results is that for all runs at both values of N the zero-T storage capacity for $l = (20, 50)$ was greater than the zero-T storage capacity for $l = 5$. Although the difference in α_c between the two sequence lengths was frequently only 0.1 it was nonetheless a consistent result, which is not predicted by either the hand-waving argument of the previous paragraph or the calculations of [ZLPW94].

For $l = 2$ the system appears to have two critical values of α. The first, which has a value of ~ 0.22 at $T = 0$, is the value at which the overlap with the desired pattern goes from $\mathcal{O}(1)$ to $\mathcal{O}(0.1)$ as α increases. The second, which we were unable to obtain for $T < 0.3$, is the value of α above which the sequence is no longer successfully retrieved according to the criteria A. Both of these data are presented in figures 4.3.2 and 4.3.2. This persistence of a recall-like state, but with a very small overlap, is reminiscent of the remanent magnetisation effect which has also been observed in the contexts of the symmetric SK spinglass [Ki86] and the Hopfield model [AGS85]. The fact that it was only observed in this study for the $l = 2$ case is in accord with the observation that, in an asymmetric SK spinglass, the remanent magnetisation decreases as the symmetry of the interactions decreases [PRS91]. In our model, the $l = 2$ case is symmetric but in all other

cases the synaptic strengths J_{ij}, J_{ji} are uncorrelated, as has been noted; we are therefore not surprised not to observe remanent magnetisation for $l > 2$.

4.3.3 Conclusions

A neural network was trained with a one-step Hebb-like prescription to store an extensive number of sequences of patterns. If the number of patterns in each sequence was greater than 2, the storage capacity for successful recall of the sequence was approximately twice that of the Hopfield model at the same operating temperature. Sequences of length 2 displayed a storage capacity that was enhanced by a factor of ~ 1.4, and also caused the network to display a remanent magnetisation effect.

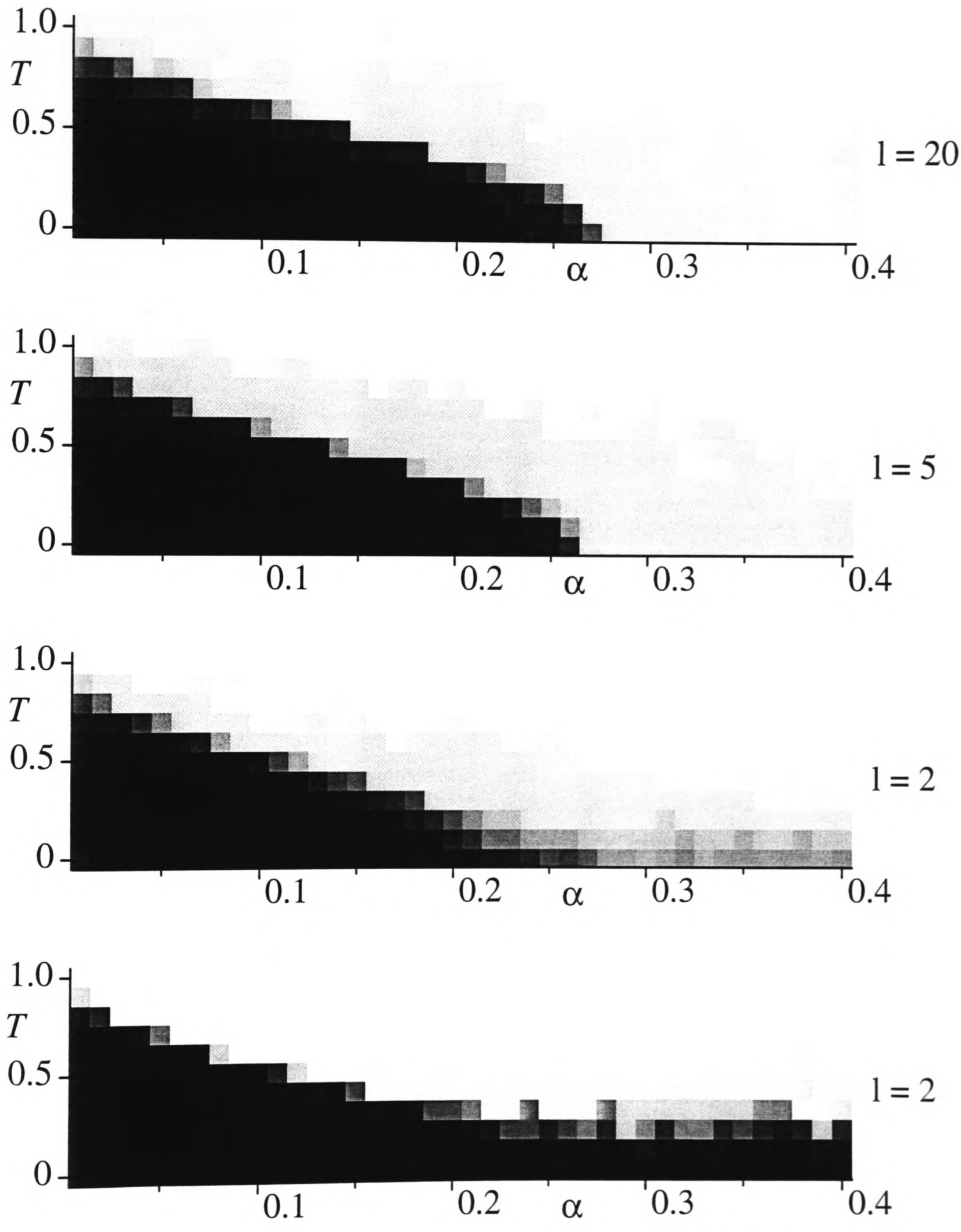


Figure 4.11: Top three diagrams: the overlap after 100 time-steps with the desired pattern, for $l = 20, 5, 2$. Darker boxes correspond to greater overlaps. Bottom diagram: The number of times that the criteria A for limit cycle behaviour are satisfied for $l = 2$. Results are averaged over seven runs.

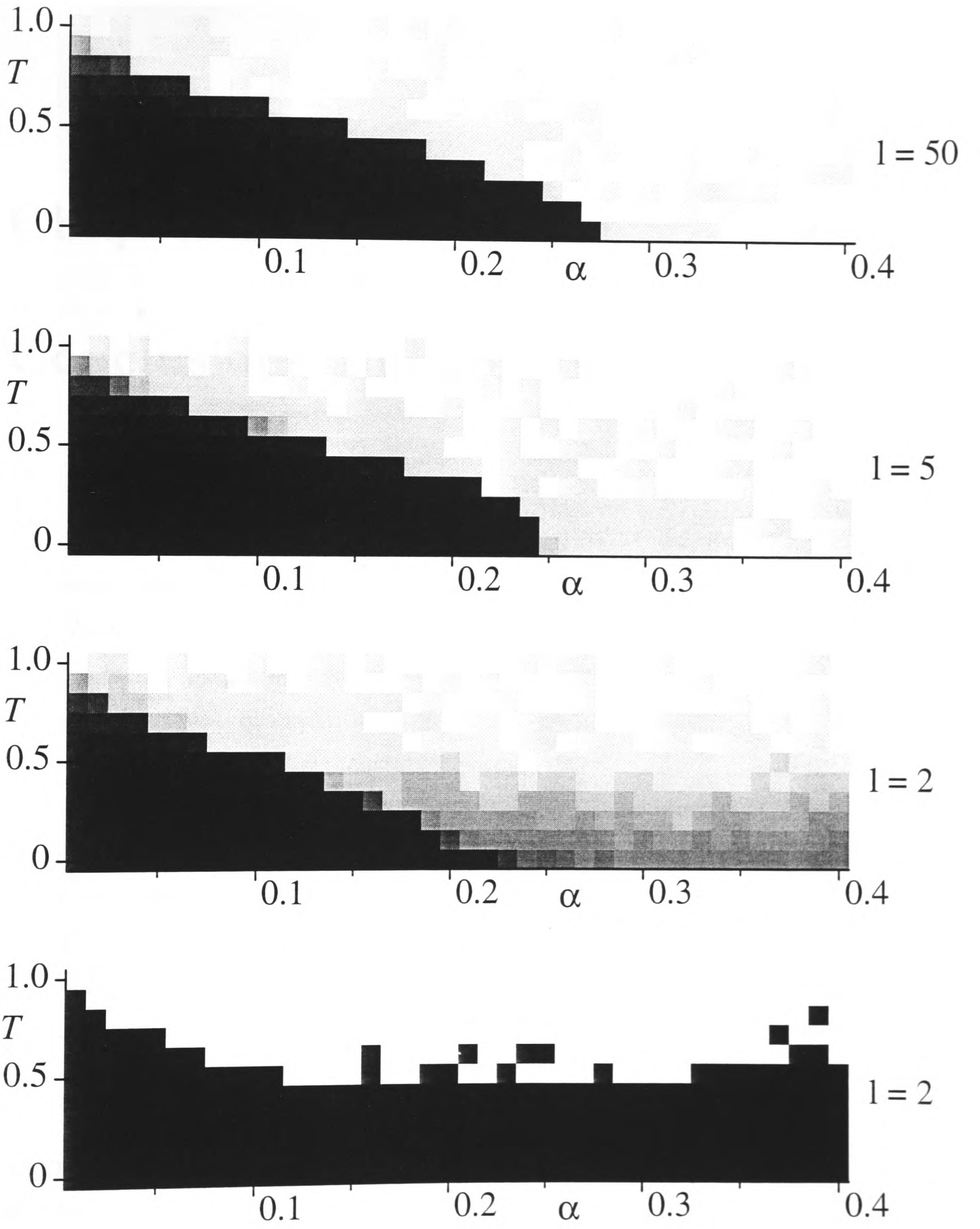


Figure 4.12: Top three diagrams: the overlap after 100 time-steps with the desired pattern, for $l = 50, 5, 2$. Darker boxes correspond to greater overlaps. Bottom diagram: Dark boxes indicate that the criteria A for limit cycle behaviour are satisfied for $l = 2$. Results are for one run only.

Chapter 5

Conclusions and Outlook

5.1 Conclusions

This section also serves as a summary of the original work in the thesis.

We have looked at two different forms of neural network models: first, attractor neural networks that optimise particular cost functions; second, attractor neural networks with separable interaction matrices which can be used (under parallel dynamics) to store sequences of patterns without the need for time delays.

In the case of the optimal neural networks we have studied the effects of replica symmetry breaking and attempted to find algorithms that will produce the optimal network where error-free storage is impossible.

For the Gardner-Derrida network we have shown that full RSB is necessary for an exact solution everywhere above saturation. We have shown further that, no matter what the cost function that is optimised, if the distribution of stabilities has a gap then the Parisi replica ansatz that has been made is unstable.

For the noise-optimal network we found a continuous transition to replica symmetry breaking at the AT line, as would be expected. The change to RSB1 improved the agreement between "experimental" and theoretical calculations of $\rho(\lambda)$ significantly. The effect on observables was smaller.

We have shown that for a training set generated with noise, the perceptron algorithm

is the best initial choice to implement the training. If additional information is available more sophisticated algorithms will be faster and give a smaller generalisation error.

In the case of the separable neural networks storing sequences of patterns, we have looked at the effects of correlations on a single-sequence network, and we have numerically investigated the storage capacity of a network storing an extensive number of patterns in such sequences.

When correlations were implemented along with a term in the interaction matrix designed to suppress those correlations, the competition between the two produced a rich range of behaviour, strongly reminiscent of biological behaviour although (due to the nature of the dynamics) not necessarily directly related to it. The most interesting finding was that, contrary to expectations, increasing the correlations and the operating temperature could improve the sequence-processing behaviour of the network.

Our investigation of the storage capacity of the network storing a large number of sequences of patterns demonstrated that such a network is capable of storing approximately twice as many patterns as the Hebb-rule network is.

5.2 Outlook

The study of the equilibrium statistical mechanics of neural networks has been a source of valuable insight into the behaviour of this kind of computers, and has thrown up many enlightening analogies to the processes that might actually be going on in the brain. The field is rapidly approaching the point, however, where care must be taken in the selection of problems to solve, so that those problems are the ones that are useful rather than merely possible.

On the other hand, the recent dynamical studies that have been performed promise to open the field up to many new possibilities, if the mathematical techniques can be developed. Although care must again be taken in the selection of problems, the dynamical approach is still sufficiently recent that it is impossible to judge its eventual scope or success.

Acknowledgements

My greatest thanks must go to my supervisor, David Sherrington, for his patience and persistence on what has not always been an easy ride for him.

I have made many good friends in Oxford whose support has been invaluable. It would be unfair to single out individuals; you know who you are.

Within Theoretical Physics, I would like to thank Ton Coolen and Michael Wong for being stimulating collaborators and good company.

Special thanks have to go to Martin Evans, Joe Burn and Daniel Fryer for help provided in the last months without which this thesis would never have been completed and submitted.

Funding for my first three years in Oxford was provided by the British Council, whom I gratefully acknowledge. Thereafter I have been funded partly by the Universities Superannuation Scheme and partly by my mother, to whom I am deeply grateful for this and many, many other reasons.

Finally, my thanks and love to Patricia, who makes it all worth while.

Bibliography

[A89] D J Amit 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)

[AGS85] D J Amit, H Gutfreund and H Sompolinsky 1985 *"Storing Infinite Numbers of Patterns in a Spin Glass Model of Neural Networks" Phys. Rev. Lett.* **55** 1930

D J Amit, H Gutfreund and H Sompolinsky 1987 *"Statistical Physics of Neural Networks near Saturation" Ann. Phys., NY* **173** 30

[AT78] J R L de Almeida and D J Thouless 1978 *"Stability of the SK Solution of a Spin Glass Model" J. Phys. A* **11** 983

[A&a] D J Amit, M R Evans, H Horner and K Y M Wong 1990 *"Retrieval Phase Diagrams for Attractor Neural Networks with Optimal Interactions" J. Phys. A: Math. Gen.* **23** 3361

[B78] A Blandin 1978 *"Theory versus Experiments in the Spin Glass Systems" J. Phys. (France)* **C6** 1499

[B91] O Bernier 1991 *"Stochastic Analysis of Synchronous Neural Networks with Asymmetric Weights" J. Phys. (France)* **C6** 1499

[B94] M Bouten 1994 *"Replica Symmetry Instability in Perceptron Models" J. Phys. A* **27** 6021

[BEKS90] M Bouten, A Engel, A Komoda and R Serneels 1990 *"Quenched versus Annealed Dilution in Neural Networks" J. Phys. A: Math. Gen.* **23** 4643

[BM78] A J Bray and M A Moore 1978 *"Replica Symmetry Breaking in Spin Glass Theory" Phys. Rev. Lett.* **41** 1068

[BS87] J Buhmann and K Schulten 1987 *"Noise-driven Temporal Association in Neural Networks" Europhys. Lett* **4** 1205

[C86] P Churchland 1986 *The Neurophilosophy of Mind* (MIT Press, Cambridge. Mass.)

[C95] C Campbell, private communication

[CAG86] A Crisanti, D J Amit and H Gutfreund 1986 *"Saturation Level of the Hopfield model for Neural Network" Europhys. Lett* **2** 337

[CR88] A C C Coolen and T W Ruijgrok 1988 *"Image Evolution in Hopfield Networks" Phys. Rev. A* **38** 4253

[CS92] A C C Coolen and D Sherrington 1992 *"Competition between Pattern Reconstruction and Sequence Processing in Non-Symmetric Neural Networks" J Phys A* **25** 5493

[CT92] L F Cugliandolo and M V Tsodyks 1992 *"Capacity of Networks with Correlated Attractors" J. Phys. A: Math. Gen.* **27** 741

[D92] E A Dorotheyev 1992 *"Stability of the One-Step Replica Symmetry Broken Phase in Neural Networks" J. Phys. A* **25** 5527

[DGD82] C de Dominicis, M Gabay and B Duplantier 1982 *"A Parisi Equation for Sompolinsky's Solution of the SK Model" J. Phys. A: Math. Gen.* **15** L47

[DGO81] C de Dominicis, M Gabay and H Orland 1981 *"Replica Derivation of Sompolinsky Free Energy Functional for Mean Field Spin Glasses" J. Physique. Lett* **42** L523

[DGZ87] B Derrida, E J Gardner and A Zippelius 1987 *"An Exactly Solvable Asymmetric Neural Network Model" Europhys. Lett* **4** 167

[DO87] S Diederich and M Opper 1987 *"Learning of Correlated Patterns in Spin Glass Networks by Local Learning Rules" Phys. Rev. Lett.* **58** 949

[EA75] S F Edwards and P W Anderson 1975 *"Theory of Spin Glasses" J. Phys. F: Met. Phys* **5** 965

[EES89] A Engel, H Englisch and A Schutte 1989 *"Improved Retrieval in Neural Networks with External Fields" Europhys. Lett* **8** 393

[ET93] R Erichsen Jr and W K Theumann 1993 *"Optimal Storage of a Neural Network Model: a Replica Symmetry Breaking Solution" J. Phys. A* **26** L61

[F88] B M Forrest 1988 *"Content-Addressability and Learning in Neural Networks" J. Phys. A: Math. Gen.* **21** 245

[FK88] J F Fontanari and R Koberle 1988 *"Information Processing in Synchronous Neural Networks" J. Phys. (France)* **49** 13

[G87] E J Gardner 1987 *"Maximum Storage Capacity in Neural Networks" Europhys. Lett* **4** 481

[G88] E J Gardner 1988 *"The Space of Interactions in Neural Network Models" J. Phys. A: Math. Gen.* **21** 257

[G89] E J Gardner 1989 *"Optimal Basins of Attraction in Randomly Sparse Neural Network Models" J. Phys. A: Math. Gen.* **22** 1969

[GD88] E Gardner and B Derrida 1988 *"Optimal Storage Properties of Neural Network Models" J. Phys. A* **21** 271

[GG91] M Griniasty and H Gutfreund 1991 *"Learning and Retrieval in Attractor Neural Networks above Saturation" J. Phys. A: Math. Gen.* **24** 715

[GPB92] M Gordon, P Peretto and D Berchier 1992 *"Learning Algorithms for Perceptrons from Statistical Physics" J. Phys. I (France)* **3** 377

[GTA93] M Griniasty, M V Tsodyks and D J Amit 1993 *"Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors" Neural Comput.* **5** 1

[H49] D O Hebb 1949 *The Organisation of Behaviour* (Wiley, NY)

[H82] J J Hopfield 1982 *"Neural Networks and Physical Systems with Emerging Collective Computational Abilities" Proc. Natl. Acad. Sci., USA* **79** 2554

[H91] N Hendrich 1991 *"Associative Memory in Damaged Neural Networks" J. Phys. A: Math. Gen.* **24** 2877

[HK90] J L van Hemmen and R Kuhn 1990 *"Collective Phenomena in Neural Networks"* in E Domany, J L van Hemmen and K Schulten (Eds), *Models of Neural Networks* (Berlin, Springer)

[HLH91] A V M Herz, Z Li and J L van Hemmen 1991 *"Statistical Mechanics of Temporal Association in Neural Networks with Transmission Delays" Phys. Rev. Lett.* **66** 1370

[HS90] D Hansel and H Sompolinsky 1990 *"Learning from Examples in a Single-Layer Neural Network" Europhys. Lett* **11** 687

[H&a86] J L van Hemmen, D Grensing, A Huber and R Kuhn 1986 *"Elementary Solution of Classical Spin-Glass Models" Z. Phys. B* **65** 53

[H&a89] H Horner, D Bormann, M Frick, H Kinzelback and A Schmidt 1989 *"Transients and Basins of Attraction in Neural Network Models" Z. Phys. B* **76** 581

[I95] J Imhoff 1995 *"A Polynomial Training Algorithm for Calculating Perceptrons of Optimal Stability" J. Phys. A: Math. Gen.* **28** 2173

[Ki86] W Kinzel 1986 *"Remanent Magnetisation of the Infinite-Range Ising Spin Glass" Phys. Rev. B* **33** 5086

[Kl86] D Kleinfeld 1986 *"Sequential State Generation by Model Neural Networks" Proc. Natl. Acad. Sci., USA* **83** 9469

[KA88] T B Kepler and L F Abbott 1988 *"Domains of Attraction in Neural Networks" J. Phys. (France)* **49** 1657

[KM87] W Krauth and M Mezard 1987 *"Learning Algorithms with Optimal Stability in Neural Networks" J. Phys. A: Math. Gen.* **20** L745

[KM89] W Krauth and M Mezard 1989 *"Storage Capacity of Memory Networks With Binary Couplings" J. Phys. (France)* **50** 3057

[MC88a] Y Miyashita and H S Chang 1988 *"Neuronal Correlate of Pictoral Short Term Memory in the Primate Temporal Cortex" Nature* **331** 68

[MC88b] Y Miyashita and H S Chang 1988 *"Neuronal Correlate of Visual Associative Long-Term Memory in the Primate Temporal Cortex" Nature* **335** 817

[MEZ93] P Majer, A Engel and A Zippelius 1993 *"Perceptrons Above Saturation" J. Phys. A* **26** 7405

[MP43] W S McCullough and W Pitts 1943 *"A Logical Calculus of the Ideas Immanent in Nervous Attractors" Bull. Math. Biophys* **5** 115

[MP88] M Minsky and S Pappert 1988 *Perceptrons* (Cambridge, MA: MIT Press).

[MPV87] M Mezard, G Parisi and M Virasoro 1987 *Spin-Glass Theory And Beyond* (Singapore: World Scientific) p 37

[MR91] B Muller and J Reinhardt 1991 *Neural Networks: An Introduction* (Berlin: Springer-Verlag)

[M&a52] N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller and E Teller 1952 *"Equation of State Calculations by Fast Computing Machines" J. Chem. Phys.* **21** 1087

[NN91] T Nakamura and H Nishimori 1991 *"Sequential Retrieval of Nonrandom Patterns in a Neural Network" J. Phys. A: Math. Gen.* **23** 4627

[NNS90] H Nishimori, T Nakamura and M Shiino 1989 *"Retrieval of Spatio-Temporal Sequence in Asynchronous Neural Network" Phys. Rev. A* **41** 3346

[O89] M Opper 1989 *"Learning in Neural Networks : Solvable Dynamics" Europhys. Lett* **8** 389

[P79] G Parisi 1979 *"Infinte Number of Order Parameters for Spin Glass" Phys. Rev. Lett.* **43** 1754

[P80a] G Parisi 1980 *"A Sequence of Approximated Solutions to the SK Model for Spin Glasses" J. Phys. A* **13** L115

[P80b] G Parisi 1980 *"The Order Parameter for Spin Glasses: a Function on the Interval 0-1" J. Phys. A* **13** 1101

[PCS93] R W Penney, A C C Coolen and D Sherrington 1993 *"Coupled Dynamics of Fast Spins and Slow Interactions in Neural Networks and Spin Systems" J. Phys. A: Math. Gen.* **26** 3601

[PGD85] L Personnaz, I Guyon and G Dreyfus 1985 *"Information Storage and Retrieval in Spin Glass Like Neural Networks" J. Phys. (Paris) Lett.* **46** L359

[PRS91] T Pfenning, H Rieger and M Schrekenberg 1991 *"Numerical Investigation of the Asymmetric SK Model with Deterministic Dynamics" J. Phys. I (France)* **1** 323

[PS94a] R W Penney and D Sherrington 1994 *"A Perceptron With A Skeletal Weight Space" J. Phys. A: Math. Gen.* **27** 23

[PS94b] R W Penney and D Sherrington 1994 *"Slow Interaction Dynamics In Spin Glass Models" J. Phys. A: Math. Gen.* **27** 4027

[SK75] D Sherrington and S Kirkpatrick 1975 *"Soluble Model of a Spin Glass" Phys. Rev. Lett.* **35** 1972

[SK86] H Sompolinsky and I Kanter 1986 *"Temporal Association in Asymmetric Neural Networks" Phys Rev Lett* **57** 2861

[SM91a] K Sakai and Y Miyashita 1991 *"Neural Organisation For The Long-Term Memory of Paired Associates" Nature* **354** 152

[T94] K Tokita 1994 *"The Replica Symmetry Breaking Solution of the Hopfield Model at Zero Temperature: Critical Storage Capacity and Frozen Field Distribution" J. Phys. A: Math. Gen.* **27** 4415

[W93] T L H Watkin *"Optimal Learning with a Neural Network" Europhys. Lett* **21** 871

[W95] A Wendemuth 1995 *"Learning the unlearnable"* (University of Oxford: Theoretical Physics Preprint)

[WH60] B Widrow and M E Hoff 1960 *WESCON Convention Report IV* (San Francisco: Western Periodicals Company)

[WRB93] T L H Watkin, A Rau and M Biehl 1993 *"The Statistical Mechanics of Learning a Rule" Rev. Mod. Phys.* **65** 490

[WRS92] K Y M Wong, A Rau and D Sherrington 1992 *"Weight Space Organisation of Optimised Neural Networks" Europhys. Lett.* **19** 559

[WS90a] K Y M Wong and D Sherrington 1990 *"Training Noise Adaptation in Attractor Neural Networks" J. Phys. A: Math. Gen.* **23** L175

[WS90b] K Y M Wong and D Sherrington 1990 *"Optimally Adapted Attractor Neural Networks in the Presence of Noise" J. Phys. A* **23** 4659

[WS93] K Y M Wong and D Sherrington 1993 *"Neural Networks Optimally Trained With Noisy Data" Phys. Rev. E* **47** 4465
K Y M Wong and D Sherrington 1994 Correction to above *Phys. Rev. E* **50** 1727

[WSC95] W Whyte, D Sherrington and A C C Coolen 1995 *"Competition between Pattern Recall and Sequence Processing in a Neural Network Storing Correlated Patterns" J. Phys. A: Math. Gen.* **28** 3421

[W&a93] T L H Watkin, K Y M Wong, A Rau and W Whyte 1993 *"Optimal Classification with Multilayer Neural Networks"* Unpublished.

[ZLPW94] F Zertuche, R Lopez-Pena, H Waelbroeck 1994 *"Recognition of Temporal Sequences of Patterns Using State-Dependent Synapses" J. Phys. A: Math. Gen.* **27** 5879