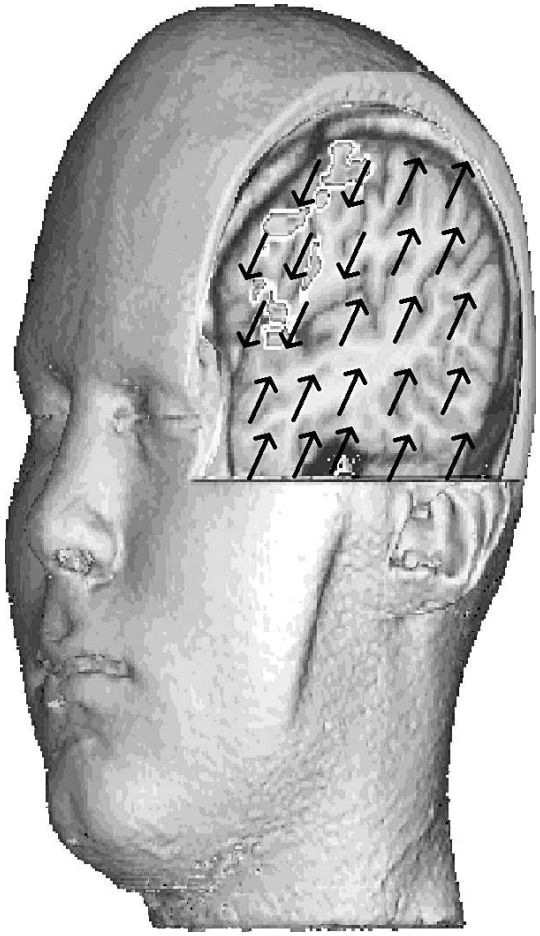
Universiteit Utrecht

May 1998, ITF-UU-98/02

Learning while Thinking

a Coupled System of Fast Neurons and Slow Synapses

Jan-Alexander Heimel



Supervisors: dr A.C.C. Coolen, King's College London
prof. dr Th. W. Ruijgrok, Universiteit Utrecht

Learning while Thinking

a Coupled System of Fast Neurons and Slow Synapses

thesis for graduation in theoretical physics
ITF-UU-98/02, May 1998, Jan-Alexander Heimel
supervisors:
dr A.C.C. Coolen, King's College London
prof. dr Th. W.Ruijgrok, Universiteit Utrecht

Preface

Isaac Asimov had Hari Seldon, the mathematical genius of his Foundation series, explain so often that the analysis of the behavior of a society of trillions of people is possible if only we know the essential properties of humans. To Seldon it is similar to the knowledge we have of a gas. Just knowing the essentials of the gas molecules, hard balls that attract each other when near, we can derive an equation of state, linking macroscopic observables as number, temperature, volume and pressure of the gas. Whether this is true remains to be seen, but this positive attitude that assumes that the properties of a system on a large scale can be derived from general knowledge of its constituents, is at the heart of physics and in the heart of many physicists. One problem, however, still defies that belief.

Almost a century after the establishment of the neuron as the elementary processing unit of our brain, we still have no clue as to how the essential properties of the neurons lead to all the marvelous things we experience every moment of our lives. The search for an explanation was not started by physicists. The theoretical physicists have joined in this quest only very recently. But what an unexpected perspective they have and what a toy model fetishists they are. The Ising-model, previously used to link the macroscopic thermal properties of, for instance, a horse-shoe magnet to the properties of the atoms it is built of, is recycled to model some of the features of the human brain.

The current level of the research is well illustrated by the fact that one of the most investigated macroscopic properties of the model used by physicists is the 'magnetization' of the neural network. In advanced models one can recognize some recognition, but we are far away from identifying or understanding speaking, planning or consciousness as emergent qualities of such networks. In this thesis these latter properties are not even mentioned outside the preface, the only thing that is done is the calculation of the average states of single neurons and the average of the states of the neurons of the entire network for a particular model. The relevance of this thesis to questions about the brain is not easy to see, but to use as excuse something Emerson Pugh said

> If the human brain were so simple, that we could understand it,
> We would be so simple that we couldn't.

At this point I want to thank Ton Coolen for the opportunity to study under his supervision in London in the recent past and in the near future. Also I would like to thank prof. Ruijgrok, who had the difficult task of supervising a research that is not his own. I thank both for the many hours they invested in my better understanding of neural networks. Furthermore I would like to thank Fieke for reviewing my thesis and for her continuing support.

Contents

Chapter 1

Introduction

In this thesis, a neural network is investigated where the states and the connections of neurons can change almost simultaneously. If we use every day terminology, we can say that the changing of the neuron states represents thinking, whereas changing connections means learning. In the neural network, as in the real world, thinking and learning are intertwined: thinking makes one learn, and learning alters ones thinking. In the particular neural we will study, we make the assumptions that learning is a much slower process than thinking. Under this assumption we can look how several parameters of the neural network influence the behavior. This will not be done before the end of chapter two. First we need to define our model.

Overview of chapter one

In this introductory chapter, we will first try to define a neural network and introduce a model for the behavior of the biological neurons. Later, the model is modified in order to make the stationary state of the network a state of 'thermal' equilibrium, which can be analyzed with tools from statistical mechanics. The drawback of the modification is that it will restrict us to considering only very artificial neural nets.

About halfway in the chapter, learning is added to the model. The particular way in which this is done here stems from 1987, but more recently (1993) people realised that the system was fit for a far better mathematical analysis. This particular neural network model stands out in the sense that is one of the very few models that can be analyzed where learning and thinking are fundamentally intertwined. In contrast to the more common models that first learn and then think, the neural network model introduced in this chapter is indeed *learning while thinking*.

Near the end of this chapter a very different neural net model is described. It is also a model that learns while it thinks and it succeeds in providing clues to some neurophysiological questions. This model is the main motivation for considering the particular extension of the 1993 model considered in this thesis.

Overview of the next chapters

In chapter 2 convenient variables are introduced that describe the behavior of the network when the number of neurons is extremely large. In the same chapter we will derive relations that these variables, the *order parameters* will satisfy. In the derivation of these variables some aspects of the behavior of the neural network model will surface. In the third chapter an analysis of the possible values of the order parameters is given. The results are presented in the form of various phase diagrams. The fourth chapter contains the results of a small number of numerical experiments that I have performed to confirm the predictions made in the earlier chapters. Finally, in the fifth and last chapter I will try to summarize the conclusions we can draw about the particular neural net model of the thesis.

1.1 Neural Networks

The name *neural networks* is used for a vast range of systems. It is difficult to give a definition of a neural network. A first attempt is that all neural networks consist of a set units, called *neurons*, that exchange information about their states with one another and can decide on the basis of this information to adjust their states. This is a definition so vague, that human society as well as a non-ideal gas are included. Obviously this cannot be a workable definition of a neural network, but before we can improve the definition, we must acknowledge the difference between the implementation of the network and the mathematical model describing its behavior. For the implementation, one must think of, for instance, the gray matter inside our skull or a silicon chip.

A less ambitious attempt to define 'neural networks' is just to define the mathematical models that we call neural networks. These models consist of a set of N variables, $\{S_1, \ldots, S_N\}$, called neurons. The neuron variables take values in $\mathbb{R}$. In some models the values of the variables range over the entire real line, in other models they can only be one or zero, or, as is the case in the model we will consider, just minus one and plus one.

The neuron variables are time-dependent and the state of a certain neuron at a certain time is a function of the states of the neurons at an earlier time. This function possibly includes a stochastic element and may be dependent on the history of the network. On one hand, this definition is general enough to include all neural networks that I have encountered and on the other hand, the one-dimensionality of the neuron state is restrictive enough to rule out human society and a gas.

The last definition of the words 'neural network' does not give any idea about the way neurons interact in general. To give an idea of the sort of interaction one must think of, we look at the mother —in more than one sense— of all neural networks, the human brain.

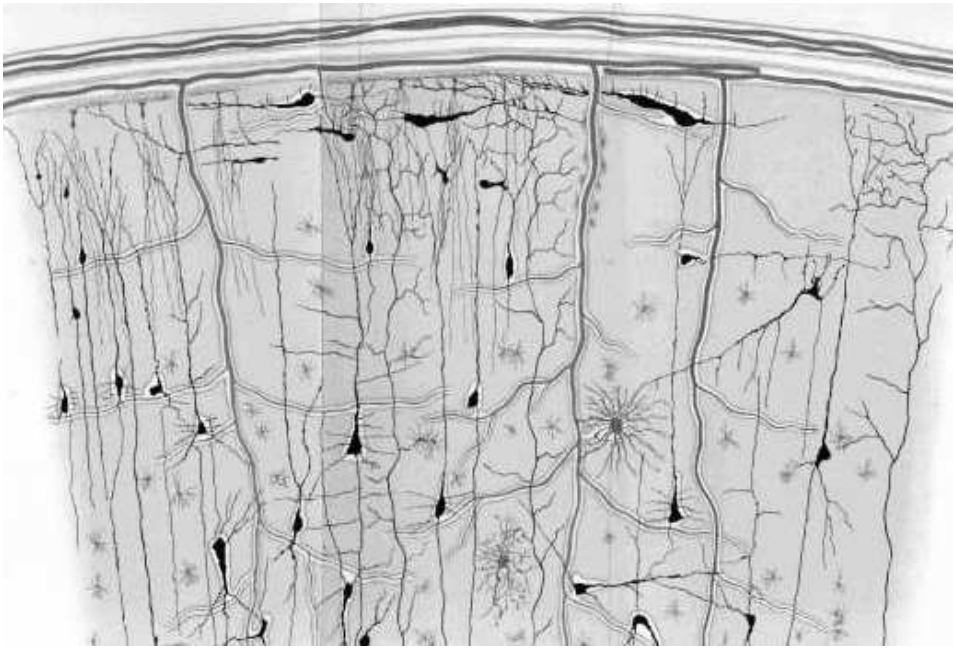
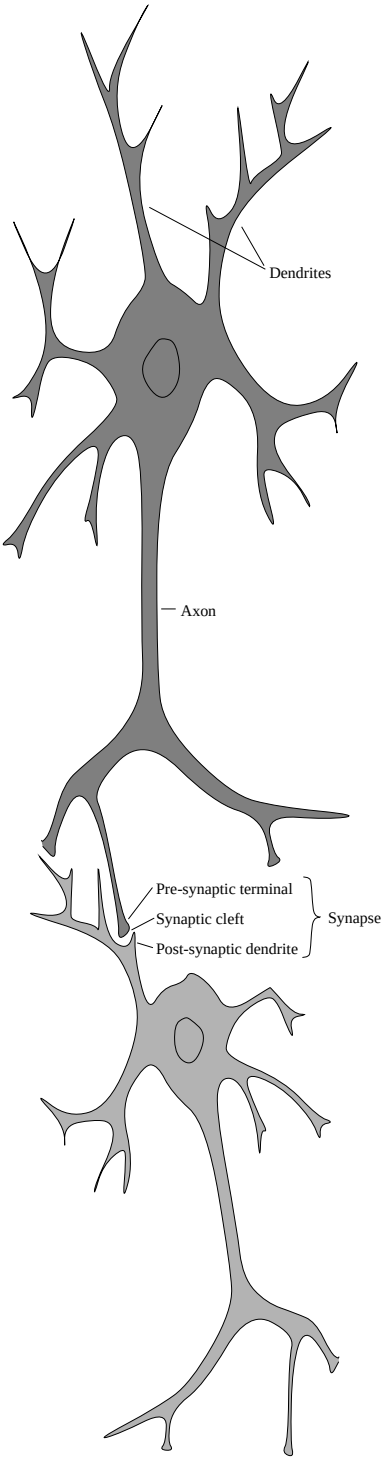


Figure 1.1: Golgi stained neurons in the human cortex, founded on plates by Ramon y Cajal, Retzius and Andriezen.

1.1.1 The Brain

The length of this subsection does justice neither to the complexity nor to the capacity of the human brain. Libraries can be filled with information we have about the brain, and truly many more libraries could be filled with knowledge we lack at present. A thesis for graduation in theoretical physics is not the proper place to give more than a glimpse of the knowledge we have about the working of the brain. I will try to tell just enough to give the reader an idea in which sense the neuron cells in our head are the implementation of a neural network as defined above.



The first to recognize a complex network of neuron cells in the brain was the Spaniard Ramon y Cajal. When at the turn of the century he stained brain tissue using a newly discovered coloring method, he was able to draw pictures such as figure (1.1). Even though only a fraction of all neurons in a sample is colored by this Golgi staining, one can clearly identify individual neurons and some of the tentacles connecting them. In the brain there are neurons of many kinds, but they all share the same basic structure and mode of operation. A neuron can be in two states: firing or non-firing. When a neuron fires, a signal called the *action potential* travels from the *cell body* (see the figure on the right) along its main *axon* into the all axon-branches until the signal reaches the *synapses*. At the synapses the signal of one neuron is transmitted to the *dendrite* of another. From there the signal travels through the dendritic tree to reach the cell body. There all signals (which basically are potential differences) from other neurons coming into this second neuron add up. If the total signal is larger than a certain threshold value, this neuron will fire, otherwise it stays silent. The synapses can be excitatory or inhibitory, the first creating positive signals stimulating the recipient neuron to fire, the latter creating negative signals discouraging the firing. One neuron may have several synapses linked to a certain other neuron. In this way the pre-synaptic neuron can have a larger influence on the post-synaptic neuron than is possible by having just one synaptic junction. In addition the efficacy of synapses may vary, creating smaller or larger signals in the dendrites.

We will now formalize this very brief discussion of the behavior of biological neurons. At time t, the state of the neuron labeled $i \in \{1, \ldots, N\}$ which is either firing or not, is represented by the function $s_i(t) \in \{0, 1\}$[1]. The sum of the efficacies of the synapses linking pre-synaptic neuron j to post-synaptic neuron i is described by the weight W_{ij}. If W_{ij} is positive, firing by neuron j excites neuron i, if W_{ij} is negative neuron j inhibits the firing of neuron i. Finally the threshold of a neuron i for firing is given by the variable ϑ_i.

A simple model for the biological neural network is then given by randomly selecting at each time step a neuron i and updating it according to the rule

$$s_i(t+1) = \begin{cases} 1 & \text{if } \sum_j W_{ij} s_j(t) \geq \vartheta_i \\ 0 & \text{if } \sum_j W_{ij} s_j(t) < \vartheta_i \end{cases} \tag{1.1}$$

[1] Actually the $s_i(t) = 1$ is used not to represent the fact that neuron i is firing at time t, but that it is firing at a high frequency, i.e. around time t, many action potentials are sent one after the other down the axon.

1.2 Mathematical Formulation

In the previous section we have spoken about some properties of neurons and neural networks. We have introduced a simple mathematical model of a neuron (1.1), but we have made no attempt whatsoever to mathematically analyze the behavior of such units. In this section we make a start by looking at a neural network consisting of neurons like (1.1) with fixed weights. To simplify the notation (and, for reasons that will become clear later on, to exploit the similarity between neural networks and systems called spin glasses) we introduce the symmetrical neuron variables $\sigma_i = \pm 1$ instead of s_i:

$$\sigma_i = 2s_i - 1. \tag{1.2}$$

At the same time we transform the threshold ϑ_i and weights W_{ij} into a variable θ_i, which we call the *external field* and new weights J_{ij}, respectively

$$J_{ij} = \frac{1}{2}W_{ij}, \qquad \theta_i = -\vartheta_i + \frac{1}{2}\sum_j W_{ij}. \tag{1.3}$$

In the new set of variables, if neuron i is selected for update at time t, the update rule reads:

$$\sigma_i(t+1) = \operatorname{sign} h_i(t), \qquad h_i(t) = \sum_j J_{ij}\sigma_j(t) + \theta_i. \tag{1.4}$$

where the sign is either plus or minus one. The variable h_i is the input to neuron i, also called the *local field*. The process of randomly selecting and updating one neuron at the time with the above rule is called *Glauber dynamics*.

The model is not perfect, nor are real life neurons. They are all too human, they make mistakes and are communication with varying success. In the application of neural networks to pattern recognition problems, this turns out to be not a bug, but a feature. Therefore we would like to include the element of neural failure in our model. We can do this by adding a random variable η to the local field of a neuron at each time step, i.e.

$$\sigma_i(t+1) = \operatorname{sign}\left[h_i(t) + \eta_i(t)\right], \tag{1.5}$$

where the noise $\eta_i(t)$ is chosen at each update and for each neuron independently from a certain distribution $P(\eta)$. Even if we know the precise state of the network at time t and we know that neuron i will be updated, the outcome will be a stochastic variable. If the noise is larger than minus the local field, the state of the neuron after update will be one, if the noise is smaller, the state will become minus one. The probabilities of the possible values of neuron i at time $t+1$ are

$$p(\sigma_i(t+1) = +1) = \int_{-h_i(t)}^{\infty} d\eta\, P(\eta), \qquad p(\sigma_i(t+1) = -1) = \int_{-\infty}^{-h_i(t)} d\eta\, P(\eta). \tag{1.6}$$

We will come back to the question of which distribution to choose for the noise, but for now we only assume that the distribution is continuous, symmetric around zero and everywhere positive. If the primitive $f(\eta)$ of $P(\eta)$ is chosen such that $f(0) = 0$, then $f \in C^1$, $f(-x) = -f(x)$ and $f'(x) > 0$. We can write the two probabilities in a single expression

$$p(\sigma_i(t+1) = \pm 1) = \frac{1}{2} \pm \int_{-h_i(t)}^{0} d\eta\, P(\eta) = \frac{1}{2} + \sigma_i(t+1) f(h_i(t)). \tag{1.7}$$

If we have the configuration $\vec{\sigma} = (\sigma_1, \ldots, \sigma_N)$ of the N neurons, then the configuration $F_i\vec{\sigma}$ is the configuration $\vec{\sigma}$ with neuron i flipped, i.e. $F_i\vec{\sigma} = (\sigma_1, \ldots, \sigma_{i-1}, -\sigma_i, \sigma_{i+1}, \ldots, \sigma_N)$. The chance, given a selection of the neuron i for updating, of the state of the neurons to change from $\vec{\sigma}$ to $F_i\vec{\sigma}$ is

$$W_i(\vec{\sigma}) = \frac{1}{2} - \sigma_i f(h_i(\vec{\sigma})). \tag{1.8}$$

We might start a neural network with a situation where we precisely know the states of all neurons. After a single time step we have lost this certainty and we can only speak about $p_t(\vec{\sigma})$, the probability of the neural network to be in state $\vec{\sigma}$ at time t. The time evolution of this probability can be written in the form

$$p_{t+1}(\vec{\sigma}) - p_t(\vec{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\left[W_i(F_i\vec{\sigma})p_t(F_i\vec{\sigma}) - W_i(\vec{\sigma})p_t(\vec{\sigma})\right]. \tag{1.9}$$

As there is no explicit time reference and the state of the system at time $t+1$ is not dependent on the states at times earlier than t, the equation above defines a Markov process. At this point we make one of the fundamental choices for the line of this thesis as we decide to look only at the probability distribution that is left invariant under this time evolution, the stationary distribution $p(\vec{\sigma})$ satisfying

$$\forall\vec{\sigma}: \qquad \sum_{i=1}^{N}\left[W_i(F_i\vec{\sigma})p(F_i\vec{\sigma}) - W_i(\vec{\sigma})p(\vec{\sigma})\right] = 0. \tag{1.10}$$

The analysis of the dynamics of the above equation is very interesting, but it is a course we do not want to pursue here.

1.2.1 Stationary state

Before looking at explicit expressions for the stationary state of the Markov process, we first want to prove the existence of a stationary state and the convergence of the system towards it. This part of the theory of Markov processes has been explored very thoroughly. With the help of a theorem of the spectral theory of stochastic matrices, it is very easy to prove both existence and convergence for a system with noise, as we will now show.

Let W be the matrix with coefficients $W(\vec{\sigma}|\vec{\sigma}')$ defined by the transition probabilities from state $\vec{\sigma}'$ to state $\vec{\sigma}$. The Markov process (1.9) is then expressed as

$$p_{t+1}(\vec{\sigma}) = \sum_{\vec{\sigma}'} W(\vec{\sigma}|\vec{\sigma}')p_t(\vec{\sigma}'). \tag{1.11}$$

In the Markov process, the total probability is of course conserved. This is reflected in the following property of W:

$$\sum_{\vec{\sigma}} W(\vec{\sigma}|\vec{\sigma}') = 1 \qquad \text{for all}\,\vec{\sigma}'. \tag{1.12}$$

Non-negative matrices obeying this relation are called *stochastic matrices* (actually mathematicians use the term for the transposes of these matrices). The above relation also says that W has 1 as an eigenvalue with the vector $(1,1,\ldots,1)$ as the corresponding left eigenvector. No eigenvalue λ of W will exceed modulus one as we can easily see: suppose $\vec{x}$ is a left eigenvector of W. Without loss of generality we can assume $|x_1| \geq |x_i|$ for all $i \in \{1,2,\ldots,2^N\}$. We have

$$|\lambda x_1| = |\sum_i x_i W_{i1}| \leq \sum_i |x_i||W_{i1}| \leq |x_1|\sum_i W_{i1} = |x_1| \tag{1.13}$$

$$\Rightarrow \qquad |\lambda| \leq 1.$$

Given a state, there are always states that cannot be reached in one time step, even in the presence of noise. This is due to the fact that states can differ in more than one spin. However for a system with noise, there will be a non-zero probability of the system being in any of the 2^N states after a minimum of N time steps. The process (1.9) can therefore be written as

$$p_{t+N}(\vec{\sigma}) = \sum_{\vec{\sigma}'} W^N(\vec{\sigma}|\vec{\sigma}')p_t(\vec{\sigma}'), \tag{1.14}$$

with a matrix W^N with only positive coefficients (such matrices are called *positive*).

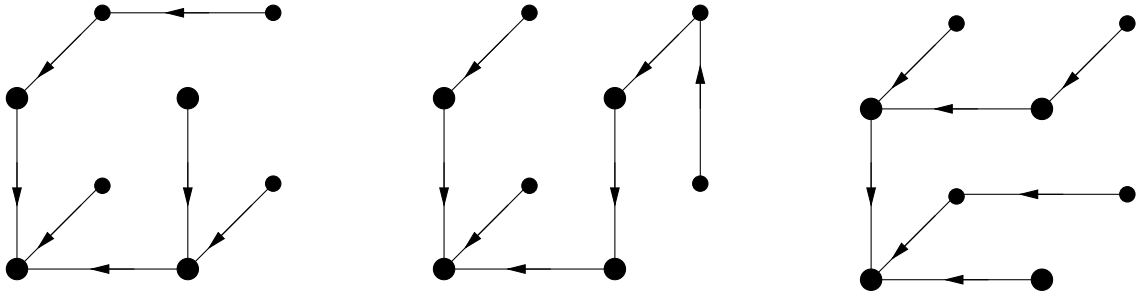


Figure 1.2: Some spanning in-trees of the cube (the state space of a three neuron network).

Perron (1907) has proved that a positive matrix always has an eigenvalue λ_1 which is real and positive and of multiplicity one and which exceeds in modulus all other eigenvalues (see e.g. [12]). Note that W itself is only non-negative, and therefore Perron's theorem does not apply to it. However, the eigenvalue 1 that we found for W will be present in W^N and has to be this Perron-eigenvalue of W^N. Furthermore, Perron proved that to this eigenvalue corresponds an eigenvector p^1 with positive coefficients $p_i^1 > 0$. This vector is, when normalized such that $\sum_i p_i^1 = 1$, the unique stationary probability distribution of the N-step Markov process (1.14). The stationary condition for the original Markov process certainly has one solution and if there would have been more than one, then these solutions would also have been eigenvectors of W^N with eigenvalue 1. As there exists only one eigenvector of W^N with eigenvalue 1, the Perron eigenvector of W^N and the eigenvector of W^1 with eigenvalue 1 are one and the same. The argument also guarantees that there are no other eigenvalues of W with modulus one.

The convergence is proven very easily. The matrix W has $n \equiv 2^N$ eigenvalues $\lambda_1, \ldots, \lambda_n$ with λ_1 the Perron eigenvalue 1. Let $p^1, \ldots, p^n$ be linearly independent corresponding eigenvectors and let $\sum_i p_i^1 = 1$. Any initial distribution p can be decomposed into this basis: $p = a^1 p^1 + \cdots + a^n p^n$. Then it follows from $|\lambda_j| < 1$ for $j \neq 1$ that

$$\lim_{m\to\infty} W^m p = \lim_{m\to\infty} \sum_{j=1}^{n} a^j \lambda_j^m p^j = a^1 p^1. \tag{1.15}$$

Because multiplication with the matrix W conserves the sum of the coefficients of a vector (conservation of probability), we can conclude that $a^1 = 1$ (or equivalently that $\sum_i p_i^j = 0$ for any j).

The work was done for us by Perron. There are many other proofs of existence, uniqueness and convergence (see e.g. [18]).

The stationary equation (1.10) is, in combination with the condition that the probability must add up to one, just a linear problem similar to

$$[W - \mathbb{I} + A]\,\vec{p} = \vec{u}, \quad \vec{u} = (1, \ldots, 1)^T, \tag{1.16}$$

where A is a matrix completely filled with 1, i.e. $A_{ij} = 1$ for all i, j. Because we have already proven the existence and uniqueness of a solution, we are confident to give the formal solution:

$$\vec{p} = [W - \mathbb{I} + A]^{-1}\,\vec{u}. \tag{1.17}$$

This expression is the basis for the high noise expansion of Coolen and Sherrington [8]. An explicit calculation of the inverse in (1.17) can be avoided by using Cramer's rule, but then we still have to calculate the determinants of N $2^N \times 2^N$-matrices.

A different approach is the following. Consider a directed graph G with 2^N vertices representing the states of the network and oriented edges between states for which a direct transition is possible. A directed edge from vertex $\vec{\sigma}$ to $\vec{\sigma}'$ is written as $(\vec{\sigma}, \vec{\sigma}')$. A *spanning in-tree* $G_{\vec{\sigma}}$ of G is a connected sub-graph of G without any cycles consisting of all vertices of G and all of its edges pointing towards

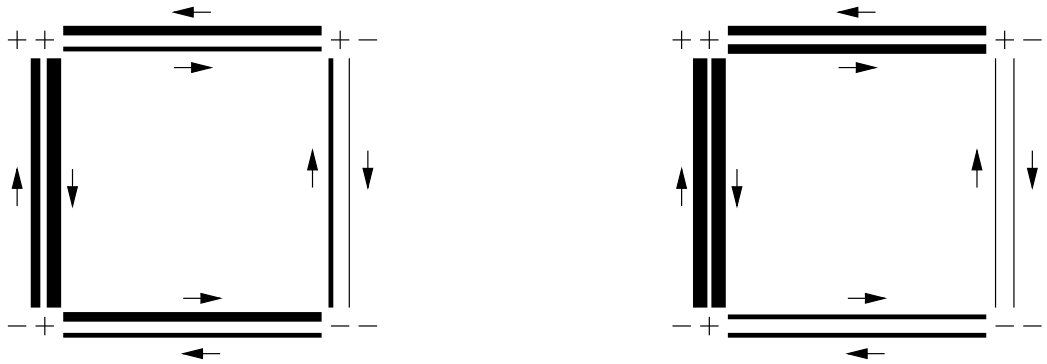


Figure 1.3: Shown are the transitions between the states of stationary probability distribution of a two neuron system. On the left, for each state, the total outward flux equals the total inward flux. This is true by definition for any stationary state. The stationary state on the right satisfies detailed balance. The flux from one state to another is identical in value to the reverse flux.

vertex $\vec{\sigma}$, see figure 1.2 for examples of spanning in-trees. King and Altman [20] have shown that the formal solution of (1.17) can be written as a sum over contributions coming from all spanning in-trees of G:

$$p(\vec{\sigma}) \propto \sum_{m} \prod_{(\vec{\sigma}',\vec{\sigma}'')\in G_{\vec{\sigma}}(m)} W(\vec{\sigma}'|\vec{\sigma}''), \tag{1.18}$$

where m labels all possible different spanning in-trees $G_{\vec{\sigma}}$. For the graph defined by the neural network is an N-dimensional hypercube, the number n_g of such spanning in-trees can be shown to be [29]

$$n_g = 2^{(2^N-N-1)} \prod_{i=1}^{N} i^{\binom{N}{i}}. \tag{1.19}$$

This gives some idea about the direct practical use of the King and Altman expression for neural networks. For a specific choice of the noise, Mogi [29] derived a marginally more explicit expression for the stationary distribution. Thus far nobody has found a satisfactory expression for the stationary probability distribution of a general neural non-symmetric network obeying (1.4).

1.2.2 Detailed balance

The solution given in the previous section could be called an explicit solution, but that is all. It is just impossible to do any calculations with the expression. There is not a more useful expression for the solution of the general stationary state condition (1.10) available and so to be able to do anything practical we have to restrict our problem.

Let us assume that the stationary probability distribution does not only satisfy (1.10), but also the much stronger condition:

$$\forall \vec{\sigma}: \qquad W_i(F_i\vec{\sigma})p_t(F_i\vec{\sigma}) - W_i(\vec{\sigma})p_t(\vec{\sigma}) = 0. \tag{1.20}$$

A distribution satisfying this condition is said to obey *detailed balance*. For any stationary distribution the probability of coming into a certain state, equals the probability of coming out of that certain state (this is just stating the definition of a stationary distribution in words). When a system is in detailed balance, then the probability of changing from one particular state to another particular state is identical to the chance of making the reverse transition. The explanation might become clearer after a look at figure 1.3. Being in detailed balance implies being in a stationary state.

The condition of detailed balance enables us to express the probability of being in any state relative to the probability of being in a certain reference state in a much easier way than is possible

without detailed balance. For instance:

$$\frac{p(F_iF_j\vec{\sigma})}{p(\vec{\sigma})} = \frac{p(F_iF_j\vec{\sigma})}{p(F_j\vec{\sigma})}\frac{p(F_j\vec{\sigma})}{p(\vec{\sigma})} = \frac{W_i(F_j\vec{\sigma})}{W_i(F_iF_j\vec{\sigma})}\frac{W_j(\vec{\sigma})}{W_j(F_j\vec{\sigma})}. \tag{1.21}$$

To be consistent, we want of course that a calculation of $p(F_iF_j\vec{\sigma})$ yields a result equal to the calculation of $p(F_jF_i\vec{\sigma})$. This condition is expressed in:

$$\frac{1-2f\left(\sigma_i h_i(F_j\vec{\sigma})\right)}{1+2f\left(\sigma_i h_i(F_iF_j\vec{\sigma})\right)}\frac{1-2f\left(\sigma_j h_j(\vec{\sigma})\right)}{1+2f\left(\sigma_j h_j(F_j\vec{\sigma})\right)} = \frac{1-2f\left(\sigma_j h_j(F_i\vec{\sigma})\right)}{1+2f\left(\sigma_j h_j(F_jF_i\vec{\sigma})\right)}\frac{1-2f\left(\sigma_i h_i(\vec{\sigma})\right)}{1+2f\left(\sigma_i h_i(F_i\vec{\sigma})\right)}, \tag{1.22}$$

where f is the primitive of the noise distribution chosen in (1.7). The inclusion of σ_i in the function $f(\sigma_i h_i)$ is possible, because f is odd by the assumption of the symmetry of the noise distribution. The consistency condition can be substantially simplified if we only consider systems without self-interacting neurons, i.e.

$$J_{ii} = 0 \qquad \text{for all } i. \tag{1.23}$$

Although this restricts the networks we can consider, we do not shed many tears for giving up the possibility of self-interaction. I do not know if self-interacting neurons were ever spotted in biological neural networks.

The important consequence is that $h_i(F_i\vec{\sigma}) = h_i(\vec{\sigma})$. Introduce the function $g(x)$ by:

$$g(x) = \frac{1-2f(x)}{1+2f(x)}. \tag{1.24}$$

Properties of $g(x)$ follow from the properties of $f(x)$. The function itself is positive with $g(0) = 1$, while its derivative, $g'(x) = -4f'(x)(1+2f(x))^{-2}$ is negative, because we have chosen a non-zero probability for any value of the noise, resulting in $f'(x) > 0$. The consistency equation written in terms of g is:

$$\begin{aligned} &g(\sigma_i \sum_k J_{ik}\sigma_k - 2\sigma_i J_{ij}\sigma_j + \sigma_i\theta_i)g(\sigma_j \sum_k J_{jk}\sigma_k + \sigma_j\theta_j) = \\ &\quad g(\sigma_j \sum_k J_{jk}\sigma_k - 2\sigma_j J_{ji}\sigma_i + \sigma_j\theta_j)g(\sigma_i \sum_k J_{ik}\sigma_k + \sigma_i\theta_i). \end{aligned} \tag{1.25}$$

Consider a network where all external fields and weights are zero except for J_{ij} and J_{ji}. Because the function g is strictly monotone, g is bijective and it follows that the two remaining weights are identical. A general network without self-interaction thus needs symmetric weights in order to reach detailed balance.

$$J_{ij} = J_{ji} \qquad \text{for all } i, j. \tag{1.26}$$

A very harsh condition which is certainly not satisfied in the neural networks that evolution has generated. Most artificial neural networks are feed-forward and they do not fall into the category of networks obeying this condition. All in all, it is a very high price to pay for making neural networks analytically tractable.

So far we have not chosen a particular noise distribution. But assuming detailed balance has, together with the conditions on the weights, fixed a noise distribution up to one parameter. We can see this if we look at a network where $J_{ik} = 0$ for all $k \neq j$. If in addition the external fields are zero, then the remaining consistency equation has the form (remember that $g(0) = 1$)

$$g(x)g(y) = g(x+y). \tag{1.27}$$

From this follows (remember $g \in C^1$),

$$g'(0)g(y) = g'(y) \qquad \Rightarrow \qquad g(y) = \exp g'(0)y. \tag{1.28}$$

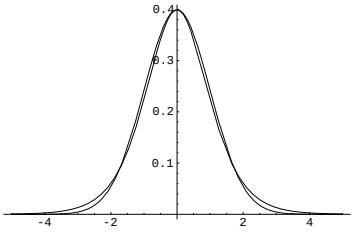


Figure 1.4: A normal distribution and the distribution dictated by detailed balance. Which is which?

We have the freedom to choose a remaining parameter $\beta \equiv -\frac{1}{2}g'(0) > 0$. Tracking back our steps from noise distribution, to f to g, we find a distribution:

$$g(x) = \exp -2\beta x \quad \Rightarrow \quad f(x) = \frac{1}{2}\tanh \beta x \quad \Rightarrow \quad P(\eta) = \frac{1}{2}\beta \left(1 - \tanh^2 \beta\eta\right). \tag{1.29}$$

Sometimes we talk about a 'neuron temperature' variable $T = \beta^{-1}$ instead of β itself. If this distribution of the noise was chosen from the onset, it can be shown [7] that there are some systems with self-interacting neurons that obey detailed balance in the stationary state, but detailed balance will not hold under small perturbations of these special weight configurations. In general, neural networks with direct self-interactions will not obey detailed balance. In trying to model nature's neurons, it would have been more logical to take a normal distribution of the noise. The stochastic behavior of biological neurons is probably the result of many small variations on a submicroscopic scale. Although we gave up modeling biological neurons by assuming detailed balance, we expect that taking the distribution dictated by detailed balance instead of a normal distribution will not fundamentally alter the properties of the system with exception of detailed balance. I do not have any convincing arguments that this indeed is the case, although figure 1.4 suggests that the two noise distributions are very similar and that on a short time scale and for small systems one will not be able to determine which distribution generated the noise.

Once we have chosen this distribution and restricted the systems we look at to symmetrically interacting neurons without self-interaction, it is easy to prove that the following probability distribution obeys detailed balance and thus is the stationary distribution sought for:

$$p(\vec{\sigma}) = Z_\beta^{-1} \exp \beta \left[\sum_{i<j} \sigma_i J_{ij} \sigma_j + \sum_i \sigma_i \theta_i\right] \tag{1.30}$$

The Z_β here is a just a normalization constant:

$$Z_\beta = \sum_\sigma \exp \left[\beta \sum_{i<j} J_{ij}\sigma_i\sigma_j + \beta \sum_i \theta_i \sigma_i\right] \tag{1.31}$$

If we define an energy for the neural network by

$$H = -\sum_{i<j} \sigma_i J_{ij} \sigma_j - \sum_i \sigma_i \theta_i, \tag{1.32}$$

we recognize the Gibbs-Boltzmann distribution $p(\vec{\sigma}) \propto \exp -\beta H(\vec{\sigma})$.

Note that in order to prove the convergence to this equilibrium distribution we have used to after N randomly chosen updates, we have a non-zero probability to be in any of the 2^N states of the network. If the number of neurons N is infinite, the so called *thermodynamic limit*, then for starting a specific state there are states that cannot be reached in a finite number of steps. In this case we cannot be sure that the system converges to a stationary state independently of the

initial state or distribution. The states of a system might be confined to an *ergodic component* of the state space dependent on the initial state. The Gibbs-Boltzmann distribution will be equal to the state distribution of a large number of trial runs of the network, but it will not be equal to the distribution of states encountered over a long time during a single run.

1.3 Spins or Neurons

Thus far, we only have spoken about neurons and neural networks. Although physicists for centuries considered them necessary for study, it was not before 1982 that they considered it necessary to study them. This does not mean that expressions similar to the ones in the previous section had not been studied before by physicists. In fact the previous decennia had been the highday of the research of systems with energies like $\sum_{i<j} \sigma_i J_{ij} \sigma_j$.

Models, known under the general term Ising models, consisting of connected units which are either minus or plus one, or *up* and *down* were used to calculate the thermodynamic properties of all sorts of system with names ending on the word -magnet. The units were models for the magnetic moments of the atoms[2] in for instance a ferromagnet and therefore the name *spin* was given to them. The interaction energy of these systems is precisely of the same form as the energy we used for the neural network, $\sum_{i<j} \sigma_i J_{ij} \sigma_j$. In the systems considered for modeling ferromagnetism, the coupling J_{ij} had two values, J for nearest neighbors and 0 for other pairs of spins. For the thermodynamical properties and, maybe even more important, the analytic tractability of the problem the dimension in the model was the crucial factor.

Around 1960 it became clear that some systems of noble metals with metal impurities exhibit unusual behavior in the low temperature specific heat. Examples of such systems are AuFe, CuMn and AgMn. The problem of measuring and explaining the thermodynamical properties of these alloys received very much attention during the next quarter of a century. For an excellent review of the theoretical work done during this period, read the book 'Spin Glasses' by Fisher and Hertz [11]. The unusual behavior of these doped metals is explained by the presence of a disordered interaction between the magnetic moments of the atoms. The name *spin glass* given to this type of systems expresses this knowledge.

In 1975 the model of Sherrington and Kirkpatrick [39] used an infinite dimensional Ising spin model with a random distribution for the value of the couplings to explain some of the typical spin glass phenomena. For the mathematical analysis of their model they devised the so-called *replica trick*. This article initiated an area of research called *replica theory*. The replica theory not only provided answers to questions in the field of spin glasses, but also provided answers for the problems with the analysis of neural networks. Replica theory was used, for example, to calculate the maximum number of stable patterns a certain pattern recognizing neural network can harbor (see section 1.4.3 on the Hopfield model) . In the next chapter we will use replica theory extensively. Because most of the replica theory and applications were developed for use in the theory of spin glasses in the context of Ising models, we will often borrow notation and nomenclature from the spin glass models. We write as often the word *spins* as we use *neurons*, or we use the word *couplings* when we discuss the *weights* or synaptic strengths. We will never talk about *thresholds*, but always use the term *external field*. As the reader has noticed, we already have used the physical term *temperature* for the amount of errors neurons make in firing or in learning, although we did not immerse the neural network in a heat bath. And we used the physical concept of *energy*. All this is the heritage of spin glass theory.

The model we construct and analyze in this thesis is inspired by neural networks and not on spin glasses. However, as we further develop the model in this chapter, we will see that the model

[2]The (more) correct quantum mechanical model used for interacting spins represents the magnetic moment of unit i by a linear combination of the three Pauli-spin matrices, coded into the spin-vector $\vec{\sigma}_i$. The Hamiltonian of an array of spins is the operator $H = \sum_{i<j} J_{ij} \vec{\sigma}_i \cdot \vec{\sigma}_j + \mu_0 \vec{\sigma} \cdot \vec{B}$, where μ_0 is the magnetic moment and $\vec{B}$ the magnetic field. In the Ising-model the $\vec{\sigma}$ are taken to be unit vectors parallel or anti-parallel to the field, which can vary in strength and sign but not in direction. The Ising-spins are thus reduced to being either up or down, $+1$ or -1.

also has an interpretation within the context of spin glasses. I like to think of this thesis as an analysis of a particular neural system, but after all, didn't Juliet already say

> What's in a name? that which we call a rose
> By any other name would smell as sweet

1.4 Some Example Systems

In a physicist's paper about neural networks, the name of Hopfield cannot be left out. In 1982, J.J. Hopfield published a landmark paper [15] in which he introduced a model of a neural network with symmetric interactions able to store and retrieve a number of patterns. The spin-neuron duality is shown clearly in the way we construct the Hopfield model here below. We will start out with a simple model for the physical ferromagnet consisting of spins. We modify the couplings between the spins of the ferromagnet to set a ground state to our liking. The resulting system is a *Mattis* magnet, named after its first constructor [25]. The addition of the weights of several Mattis magnets yields the Hopfield model, as we will see in a moment.

1.4.1 Ferromagnet

One of the most simple, perhaps *the* most simple from an analytical point of view, Ising-system is the thermodynamic limit ($N \to \infty$) of the fully connected ferromagnet. All neurons/spins are connected to each other with a uniform bond

$$J_{ij} = \frac{1}{N}, \quad \text{for all } i < j. \tag{1.33}$$

The thermodynamic limit of this bond configuration is also called the *infinite-dimensional* ferromagnet as it can be seen as a hypercube with nearest neighbor interactions in a N-dimensional space.

This system has only two ground states: all spins up, and all spins down. This becomes clear when looking at the energy of the system,

$$H = -\frac{1}{N}\sum_{i<j}\sigma_i\sigma_j. \tag{1.34}$$

If the system is embedded in a heat bath of zero temperature, the system will converge towards one of these two ground states. There are no other states that are even stable under a single flip. Which ground state will be the limit state, depends on the initial magnetization and for a finite system also on the choice of neurons to update. In the thermodynamic limit, the limit state will always have the sign of the initial magnetization (if there was any). The time average of the magnetization becomes:

$$m = \frac{1}{N}\widehat{\sum_i \sigma_i} \equiv \lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{t}\int_{t_0}^{t} dt' \, \frac{1}{N}\sum_i \sigma_i(t') = \xi, \qquad \xi = \pm 1. \tag{1.35}$$

If the system's temperature is hot, $\beta \gg 1$, the spins do not feel as great an urge to align to their local field. The state of a single spin therefore changes often, resulting in a vanishing magnetization. We see that the order in the system, the alignment of the spins to one of the groundstates, is well reflected in the value for the magnetization. In statistical mechanics, the magnetization is therefore called an *order parameter*.

1.4.2 Mattis magnet

The ferromagnet has the remarkable feature that its two ground states are the only stable states of the system and that they are *unfrustrated*, that is $\sigma_i J_{ij}\sigma_j > 0$ for all $i < j$. Each pair of spins

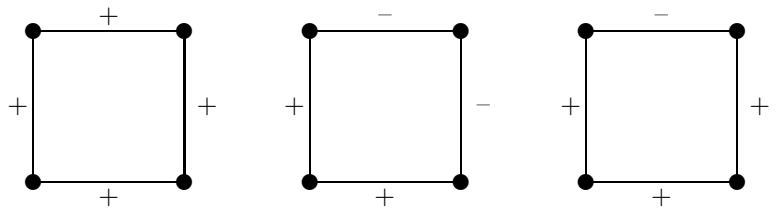


Figure 1.5: The ferromagnet shown in a) and the Mattis magnet shown in b) are both examples of unfrustrated systems. There are two configurations of the spins such that all the couplings are simultaneously satisfied. Figure c) shows a configuration of couplings that does not allow an unfrustrated placement of the spins.

is aligned according to the sign of their bond. As the couplings in the ferromagnet are all the same and positive, this is a rather difficult way of saying a simple thing. However, if we change the signs of some couplings the concept of frustration becomes more useful. Assume we have a set $\xi_i \in \{-1, +1\}$ for all spin sites $i \in \{1, \ldots, N\}$. The Hamiltonian $H(\sigma) = -\sum_{i<j} \sigma_i J_{ij} \sigma_j$ (see (1.32) is invariant under the local gauge transformation:

$$\forall\, i, j \quad \sigma_i \to \xi_i \sigma_i, \quad J_{ij} \to \xi_i J_{ij} \xi_j. \tag{1.36}$$

The resulting system has the couplings

$$J_{ij} = \frac{1}{N} \xi_i \xi_j \tag{1.37}$$

and is called a Mattis magnet. As its two ground- and stable states the spin system has the patterns $\sigma_i = \xi_i$ and $\sigma_i = -\xi_i$ for all i. Both ground states are free of frustration, see figure 1.5. If we do measurements or calculations on a Mattis magnet and we already know the ground state pattern $\pm\xi_i$ we could, of course, use a gauge transformed magnetization or *pattern overlap*,

$$m(\vec{\sigma}) = \frac{1}{N} \sum_i \widehat{\xi_i \sigma_i} \tag{1.38}$$

and get for this order parameter exactly the same behavior as for the magnetization of the ferromagnet. What if we are confronted with a Mattis magnet without detailed information of its ground state or couplings, but that we know that the pattern bits are chosen with an equal probability to be either plus one or minus one. In thermodynamic limit we can not use the standard magnetization because the Central Limit Theorem shows that it will be zero.

The following function would do a better job:

$$q = \frac{1}{N} \sum_i \widehat{\sigma_i}^2. \tag{1.39}$$

For $\beta > 1$ the individual spin time averages $\widehat{\sigma_i}$ are zero and so will be q, while at very low temperature all spins are almost fixed in one of two stable positions and the spin averages are of norm one. At this temperature q will be nearly one, which rightly reflects the ordering of the spins. In the next chapters, an order parameter very similar to q will play a dominant role in the analysis.

Notice that if not the ground state pattern but the weights are randomly distributed, we have a very different system. Large systems will certainly not have an unfrustrated ground state and there might be many metastable states instead of only two. Such a system is a spin glass model. In the next chapter more is said about a (different, but related) spin glass model.

1.4.3 Hopfield model

Now consider a set of p patterns (words, pictures, sonar data, etc.) that can each be binary coded with N bits, or as Hopfield realized, with N neurons. If one has symmetric bonds between neurons, one can easily prove (for instance by using the Hamiltonian as a Lyapunov-function) that a noiseless network reaches a static state, i.e. after a certain finite time no neuron will ever flip again. Choosing the weights of the neurons as a linear superposition of the p connected Mattis weight configurations, one can create a system with the p patterns quenched into the system as static states. Let ξ_i^μ denote digit $i \in \{1,\ldots,N\}$ of pattern $\mu \in \{1,\ldots,p\}$. The noiseless network with its weights set as[3]:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{1.40}$$

has among its static states, the p patterns, if the correlation between the patterns is negligible. When no external field is applied, the network also has the inverses of the patterns as static states. Both properties can be shown easily if the digits of the patterns are independently chosen and minus and plus one are equally probable, i.e. $\langle \xi_i^\mu \xi_j^\nu \rangle = \delta_{ij}\delta_{\mu\nu}$. Consider the local field on neuron i when the rest of the neurons is already representing pattern number α, $\sigma_j = \xi_j^\alpha$ for $j \neq i$:

$$\begin{aligned} h_i(\vec{\sigma}) &= \frac{1}{N}\sum_{j\neq i}\sum_{\mu} \xi_i^\mu \xi_j^\mu \sigma_j \\ &= \frac{1}{N}(N-1)\xi_i^\alpha + \frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\alpha} \xi_i'^\mu \xi_j'^\mu, \end{aligned} \tag{1.41}$$

where the original patterns $\mu \neq \alpha$ are changed according to

$$\xi_j'^\mu = \xi_j^\mu \xi_j^\alpha \quad \text{for } j \neq i, \qquad \xi_i'^\mu = \xi_i^\mu. \tag{1.42}$$

When N grows to infinity, the Central Limit Theorem tells us that the sum over j of the independently chosen random variables $\xi_j'^\mu$ becomes a Gaussian distributed random variable. The mean and the standard deviation of the pattern averages for the overlap due to the other patterns (the last term in (1.41)) will in the thermodynamic limit contain all information. Calculation of the pattern averages for finite N yields:

$$\begin{aligned} \langle \frac{1}{\sqrt{N}}\sum_{\mu\neq\alpha} \xi_i'^\mu \frac{1}{\sqrt{N}}\sum_{j\neq i} \xi_j'^\mu \rangle &= 0, \\ \langle (\frac{1}{\sqrt{N}}\sum_{\mu\neq\alpha} \xi_i'^\mu \frac{1}{\sqrt{N}}\sum_{j\neq i} \xi_j'^\mu)^2 \rangle &= \frac{(N-1)(p-1)}{N^2}. \end{aligned} \tag{1.43}$$

If $p = o(N)$, the standard deviation vanishes in the thermodynamic limit. In this case, application of the Central Limit Theorem show that the local field on neuron i will be ξ_i^α. If the temperature is zero, the neuron will be set according to pattern α and will remain so. This argument applies to all neurons and to all patterns and thus we can conclude that the patterns are static states if no noise is present. We have not shown here, that all patterns are stable attractors, but others have shown this to be the case.

If $p = \alpha N$, the standard deviation of the overlap due to other patterns does not vanish, and there will be a finite probability that the overlap causes disalignment of neuron i with the condensed pattern, even in the thermodynamic limit. For N large, the standard deviation from (1.43) is equal to $\sqrt{\alpha}$. The probability of the overlap with other patterns being larger than the contribution to the local field h_i of pattern α is:

$$P_\alpha \equiv \text{Prob}\left(\frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\alpha} \xi_i'^\mu \xi_j'^\mu > |\frac{1}{N}(N-1)\xi_i^\alpha| \right) = \frac{1}{\sqrt{2\pi\alpha}} \int_1^\infty e^{-\frac{1}{2\alpha}x^2} dx \tag{1.44}$$

[3]Hopfield used binary neurons with the states $\{0,1\}$, but by the transformations (1.2) and (1.3) his choice of weights can easily be cast into a form for $\{-1,+1\}$-type neurons.

For $\alpha = 0.1$, $P_\alpha \approx 10^{-3}$ and some neurons will not align according to the right pattern, but it will certainly be recognizable. For $\alpha = 0.2$, $P_\alpha \approx 10^{-2}$. The pattern will be severely damaged. The pattern will even be more distorted than this figure indicates, because if some neurons are not properly aligned, the contribution to the local fields due to the right pattern will be much smaller. Somewhere between $\alpha N = .1N$ and $\alpha N = .2N$ lies the maximum number of patterns to be stored in an N-neuron Hopfield network.

Hopfield analyzed (mostly numerically) his network only without any noise, or equivalently in the zero temperature regime and found a maximum storage capacity of approximately $0.15N$ patterns. The neural network with Hopfield couplings and with a finite temperature was analyzed by Amit, Gutfreund and Sompolinsky. They analyzed the $p = o(N)$ case in [2] with traditional tools from statistical mechanics. The much more difficult case $p = \mathcal{O}(N)$ was seen to mathematically resemble spin glasses models. For this analysis [3] they used the same machinery that I will use in this thesis. They found that in general for large N no more that $0.14N$ uncorrelated N-bit patterns can be stored in the Hopfield network. Other work showed that when the patterns are all orthogonal a maximum number of N patterns can be stored.

Another notable element of the Hopfield network that surfaced in the analysis of Amit *et al.* is that noise is a feature, not a bug. When patterns are stored as static states in the network configuration, automatically some static states are created, that are a combination of an (odd) number of patterns. This effect is of course unwanted. When a little noise is introduced in the neuron dynamics it is possible to destabilize these *spurious* states, while the desired patterns remain stable.

Storing $0.14N^2$ bits of information in N^2 weights is not the most impressive storing algorithm ever encountered, but at least the cost scales like the information. The importance of the Hopfield network is that it includes a robust retrieving algorithm. Once the patterns are stored in the weights and thresholds, a large part of the neurons can be set to resemble one of the patterns. This results initially in a very distorted pattern, but the distorted pattern is not a stable state and the neuron dynamics will steer the state of the network to a nearby stable attractor. If the distortion was not too big, the nearest attractor will be the undistorted pattern. The Hopfield network is thus able to 'recognize' noisy patterns.

The Hopfield model made many physicists interested in the field of neural networks and it continues to do so.

I have not included the short description of the ferro- and Mattismagnet only for a natural construction of the Hopfield model. We will meet both models again in the course of this thesis. The ferromagnet will be used to illustrate some of the general properties of infinite-dimensional Ising-systems. And the Mattismagnet will come up when we try to interpret the analysis of the neural network model of the next section.

1.5 Dynamical Synapses

So far we have considered the weights to be fixed. Yet to explain the most intriguing property of biological neural networks, the fact that they can learn new behavior, plasticity of the synapses is crucial. Although the states of the about 10^{11} neurons in the human brain could in principal represent 100 gigabytes of information[4], it is improbable that our memories are stored in this way. People hit by an electrical discharge, like lightning or, more common, epilepsy, that resets a large fraction of the neurons, in general do not lose their memories. Memories stored in our brain are most likely located in the efficacies and location of the synapses (as is the case in the Hopfield model). For the way in which this is done, the neurobiological research provides remarkable little clues at present. Illustrative of this point is that in a 750^+ page book that offers *the essentials of neural science and behavior*[19] the evidence of learning at the cellular level comprises only five pages. Having almost no biological lead gives us some freedom in choosing a rule for the dynamics of the weights. Models where the weights and the neuron are coupled are notoriously

[4]The robustness of our brain under cell death hints at a large redundancy in information storage. The actual information content that could be stored in this way will be at least an order of magnitude smaller.

hard to study, but in 1993 Penney, Coolen and Sherrington [35] used this freedom to propose a model where both the states and the connections of the neurons are dynamic. Their model, and a variation on it independently found by Dotsenko, Franz and Mezard [9], is devised so that analysis with statistical mechanics is possible.

The PCS-model is at the heart of this thesis. It consists of N neurons, labeled with indices $i, j, k, \ldots$, each connected to all other neurons with couplings $J_{ij} = J_{ji}$. The neuron dynamics are described by (1.4) and are called *Glauber dynamics*. The weights will also evolve in time, but in response to the states of the neurons and on a much larger time scale. To make the analysis of this interacting system of fast and slow variables tenable, first the number of neurons is send to infinity. Subsequently the time-scale of the weight dynamics is taken in such a way that the weights only respond to the thermal average of the neuron states.

The rule proposed by PCS is

$$\tau \frac{d}{dt} J_{ij} = \frac{1}{N}\langle \sigma_i \sigma_j \rangle - \frac{1}{N} A_{ij} - \frac{1}{\mu_{ij}} J_{ij} + \frac{1}{\sqrt{N}} \eta_{ij}(t) \quad \text{for} \quad i < j. \tag{1.45}$$

In the remainder of this section every element of this rule will be defined, explained and motivated.

Hebbian learning: $\frac{1}{N}\langle \sigma_i \sigma_j \rangle$

In 1949 the psychologist Hebb hypothesized that the associative learning seen macroscopically in the behavior of man and animal could be explained when microscopically a same associative learning rule would apply: "When an axon of cell A ... excites cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells so that A's efficiency as one of the cells firing B is increased," [13]. There is some biological evidence that this process takes place in the hippocampus of mammals, but it would be too much to say that Hebb's rule is a cell biologically motivated rule for weight adaption. Next to the successfull application of Hebb's rule in artificial associative memories, like Hopfield's model[5] Hebbian learning has shed some light on the early development of the visual system in mammals, which we will discuss in section (1.7). In the study of artificial feed-forward networks many other rules for weight change have been proposed, but these rules not only lack any biological evidence, but also all biological plausibility (of course, this argument did not keep us from considering only symmetric interactions).

Considering only symmetric neural networks we use a symmetric version of Hebb's learning term, $N^{-1}\langle \sigma_i \sigma_j \rangle$. The factor N^{-1} is present to ensure non-trivial behavior of the neuron dynamics in the thermodynamic limit. We want $\langle \sigma_i \sigma_j \rangle = \mathcal{O}(1) \quad (N \to \infty)$, therefore the local field on each of the neurons must be of $\mathcal{O}(1)$. This necessitates the average value of a weight to be of $\mathcal{O}(N^{-1})$ (the range of the weights will be of $\mathcal{O}(N^{-1/2})$ as we will see later).

If the Hebbian term would not be present, then at any given moment this system would be a Sherrington-Kirkpatrick spin glass. The inclusion of the Hebbian term in the coupling dynamics, can also be interpreted as modeling an aging-effect of real-life spin glasses. The pollution atoms, that influence the value of the couplings between two grid points, may wander slowly through the spin glass. The time spent in a certain position is proportional to the Boltzmann factor of the entire energy of the system in such a configuration. In this interpretation, the heat bath of the spin glasses is the cause of the spin noise and the weight noise. The system only has one temperature, i.e. there is a constant n such that $\tilde{\beta} = n\beta$.

Bias: $-\frac{1}{N} A_{ij}$

The bias term can be used to steer the weights in a certain direction. When a non-negative correlated external field is applied to the network, then having $A_{ij} > 0$ creates a competition between the correlation induced growth and the decrease due to the biases. Although we are only looking at the evolution of the weights J_{ij} for $i < j$, it is convenient to define the entire symmetric matrix, $A_{ij} = A_{ji}$.

[5] Consider the Hebb rule: $\dot{J}_{ij} = \sigma_i \sigma_j$. Starting with $J_{ij}(0) = 0$ and fixing the neurons in each pattern ξ_i^μ for a time J, the Hopfield weights $J_{ij} = J \sum_\mu \xi_i^\mu \xi_j^\mu$ are generated by these dynamics.

Decay: $-\frac{1}{\mu_{ij}}J_{ij}$
If there are no constraints on the weights, Hebbian learning will cause the weights to grow beyond all bounds. The way nature solves this problem is not clear. In a model one has a couple of options if one wants to consider only local (i.e. within the range of one neuron) rules. One could set *strong* or *weak* bounds for the individual couplings, $|J_{ij}| \leq C$, which a weight can not outgrow. Weights get fixed reaching a strong boundary, while from weak boundaries they can escape. Another option is setting bounds for the sum of weights from or to a certain neuron: $\sum_j |J_{ij}| \leq C$ or by constraining this sum: $\sum_j |J_{ij}| = C$. The enforcement of all these rigid bounds very much complicates the analysis and we will not attempt to solve the problem is this way.

Another way to prevent the weights to grow indefinitely is to introduce a weight decay proportional to the individual weight or (when the weights are not required to be symmetric) to the sum of the weights onto or from a neuron. Adding the decay term $-J_{ij}/\mu_{ij}$ to the time evolution, is effectively introducing a cost proportional to J_{ij}^2/μ_{ij} for the height of the weight. If only the Hebbian learning term and the weight decay terms are present, then we see an other aspect of the inverse decay parameter μ_{ij}. When such a system is in a stationary state and a weight J_{ij} is slightly perturbed from its stationary value, then the relaxation time is $(\tau\mu_{ij})^{-1}$.

Noise: $\frac{1}{\sqrt{N}}\eta_{ij}(t)$
This last term is a Gaussian white noise, completely defined by the first two moments:

$$\begin{array}{rcl} \overline{\eta_{ij}(t)} & = & 0, \\ \overline{\eta_{ij}(t)\eta_{kl}(t')} & = & 2\tau\tilde{\beta}^{-1}\delta_{ij,kl}\delta(t-t'). \end{array} \tag{1.46}$$

The noise parameter $\tilde{\beta}$ will play a role similar to an inverse temperature used in statistical mechanics. The scaling factor $N^{-1/2}$ is used to make the total influence of the noise on the time-evolution of the weights of the same orders as the other terms. A scaling factor of lower order would cause the influence of the noise to vanish, a higher order would lead to complete domination of the noise, regardless of the temperature.

1.5.1 Langevin equation

The time-evolution given by (1.45) is a non linear Langevin equation (see appendix A). In general systems described by a Langevin equation do not have to go to thermal equilibrium. The Langevin equation (1.45) has a nice feature that was first seen by Shinomoto, who studied this equation in the absence of noise [41]. The average $\langle\sigma_i\sigma_j\rangle$ is integrable with respect to J_{ij}:

$$\langle\sigma_i\sigma_j\rangle_{\{J_{ij}\}} = \frac{\partial}{\partial J_{ij}}\frac{1}{\beta}\log Z_\beta(\{J_{ij}\}), \tag{1.47}$$

where Z_β is the normalization constant defined in (1.31). As the other terms of (1.45) can be integrated as well, the force in the Langevin equation is conservative and (1.45) can be written as

$$\tau\frac{d}{dt}N^{\frac{1}{2}}J_{ij} = -\frac{\partial\mathcal{H}(J)}{\partial N^{\frac{1}{2}}J_{ij}}, \tag{1.48}$$

with the energy function

$$\mathcal{H}(\{J_{ij}\}) = -\frac{1}{\beta}\log Z_\beta(\{J_{ij}\}) + \frac{1}{2}N\sum_{i<j}\frac{J_{ij}^2}{\mu_{ij}} + \sum_{i<j}A_{ij}J_{ij}.$$

The J_{ij}'s are of $\mathcal{O}(N^{-1/2})$ and are therefore not the variables to consider in the thermodynamic limit. We will use $N^{1/2}J_{ij}$ instead. This is made explicit in the transfer of a factor $N^{1/2}$ to the left hand side.

In appendix A it is proven that Langevin models with conservative forces converge in time to a unique equilibrium distribution, given by the Gibbs-Boltzmann measure of an inverse temperature $\tilde{\beta}$ and energy $\mathcal{H}$. If we define the ratio of the two noise parameters as

$$n \equiv \tilde{\beta}/\beta, \tag{1.49}$$

we can write the equilibrium weight distribution as

$$p(\{N^{\frac{1}{2}} J_{ij}\}) = \tilde{Z}^{-1} \exp -\tilde{\beta}\mathcal{H}(\{J_{ij}\}), \tag{1.50}$$

$$\tilde{Z} = \int \left[\prod_{i<j} dJ_{ij}\, N^{\frac{1}{2}}\right] [Z_\beta]^n \exp\left[-\frac{1}{2}\tilde{\beta}N \sum_{i<j} \frac{J_{ij}^2}{\mu_{ij}} - \tilde{\beta}\sum_{i<j} A_{ij}J_{ij}\right]. \tag{1.51}$$

The normalizing constant $\tilde{Z}$ will be the starting point for the analysis of the model in the next chapter. A model in which θ_i, A_{ij} and μ_{ij} can be specified for all i, j separately, is extremely hard to analyze. We will restrict the networks under consideration. The way we will do this, we will specify soon, but in the next section we first review some work closely related to the model as setup in this section. Subsequently we discuss the neural network model that is the main motivation for my choice of the restriction.

1.6 Related Work

Since 1993 several variations on the above theme have been studied.

Random biases

Penney, Coolen and Sherrington [35] considered the decay and bias to be site-independent fixed variables, i.e. $\mu_{ij} = \mu$ and $A_{ij} = A$ for all i,j. They mainly considered an unbiased system, but did some calculations for a system with a ferromagnetic bias. Penney and Sherrington [36] have taken the biases to be time-independent random variables. The presence of biases can be seen as forced learning. Note that the structure of the uncorrelated biases is certainly different from the weights chosen in the Hopfield model. In the Hopfield model the weights are correlated. If the bits ξ_i^μ of the p patterns are independently chosen such that the pattern average is m, i.e. $\langle \xi_i^\mu \rangle = m$, we find for the correlation between the couplings of three neurons:

$$\begin{aligned} \langle J_{ij}J_{jk}J_{ki}\rangle &= \langle \frac{1}{N^3}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu \sum_{\nu=1}^{p}\xi_j^\nu\xi_k^\nu \sum_{\rho=1}^{p}\xi_k^\rho\xi_i^\rho\rangle \\ &= \frac{1}{N^3}\left[p + 3p(p-1)m^4 + p(p-1)(p-2)m^6\right] > 0. \end{aligned} \tag{1.52}$$

Whereas for independently chosen biases the average of $A_{ij}A_{jk}A_{ki}$ vanishes by definition. If the weights mirror the biases precisely, the system will have the SK-spin glass property that it has number of stable states that grows exponentially with the number of neurons.

In this thesis we will see that for calculation of the partition function of a site-independent bias we need a mathematical method called the *replica method*. At this point we do not want to discuss this method except for saying that in a replica method calculation a set of replicas of the system with fixed weights are introduced. For a 'simple' unbiased system, a detailed calculation takes quite a number of pages. To calculate a system with a random bias Penney and Sherrington need a second instance of this replica method, making the calculation even much more intricate.

Oscillator neurons

The PCS-model and analysis has been ported to a different neuron model by Anemüller [4] under supervision of Coolen. Instead of Ising-type neurons, they considered oscillator neurons. In neural

tissue, experiments have shown that sometimes neurons tend to fire action potentials consistently synchronously or asynchronously. Oscillator neurons are used as a neural model that can exhibit such synchronization. Rather than being either firing (+1) or at rest (-1), the oscillator neuron is at each moment in the process of firing an action potential. The process is not instantanuous, but takes some time. For each neuron, a phase is used to describe at which stage in the firing procedure the neuron currently is. The interaction of the neurons is through their difference in phase only.

The analysis for these neurons is more complicated than it is for the Ising-neurons, but it yields a phase diagram that is identical, save a factor two in the critical neuron temperature, to the Ising-neuron phase diagram. Jongen, Bollé and Coolen are repeating the Penney and Sherrington analysis for this type of neurons. In this calculation an error in the original work surfaced, but except for this error, the analysis seems to give results similar to the Ising-type neurons [16].

Anti-Hebbian learning

A different model with fast neural and slow weight dynamics has been studied by Dotsenko, Franz and Mezard [9]. Instead of including the frustration removing Hebbian term into the evolution of the weights they took the opposite 'anti-Hebbian' learning term: $-\langle\sigma_i\sigma_j\rangle$. Instead of allowing all possible values for the weights, they constrained the weights, by only allowing Hopfield type interactions (see 1.40):

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \tag{1.53}$$

where $\xi_i^\mu = \pm 1$ are the stored patterns. They considered the number of patterns p to be proportional to N, i.e. $p = \alpha N$. Their restriction is not made explicit in the weight dynamics, but only in the sum over states in the partition function. They sum over all possible patterns instead of over the entire weight space. Their partition function is of the form

$$\tilde{Z} = \sum_{\xi} [Z_J]^{-n}\,. \tag{1.54}$$

Compare this to the partition function (1.51) we use. Dotsenko *et al.* found that for this system in the zero neuron temperature limit each of the patterns can be retrieved perfectly up to $\alpha = 1$. This contrasts with the static Hopfield model, where perfect retrieval is only possible up to $\alpha = 0.14$. They think that the anti-Hebbian learning (if one can speak of learning when one does not specify an explicit learning rule) tends to orthogonalize the patterns. Or rather, a system storing p patterns that are orthogonal has a smaller energy, than a system with an equal number of patterns where the patterns have a finite overlap. Their interpretation is motivated by the fact that a Hopfield network can store up to N orthogonalized patterns.

1.7 Linsker's model

In 1986, Linsker [23] did computer simulations of a neural network inspired by the layered architecture of the primary visual cortex of mammals. Although certainly not meant as a detailed model of this part of the brain, Linsker's model yielded results that give insight into the pre-natal development of orientation-selective neurons in certain primates. At first sight his model will look very different of the neural net model we have elaborated on so far, but I will explain how I adjust our model in such a way that a comparison might be possible.

1.7.1 System specification

In contrast to the two-state neurons (1.4) we have considered so far, Linsker uses continuous neurons. The activation of a neuron is the sum of its weighted input minus a constant threshold.

$$s_i = \beta' \sum_{j\in V_i} J_{ij} s_j - \vartheta, \tag{1.55}$$

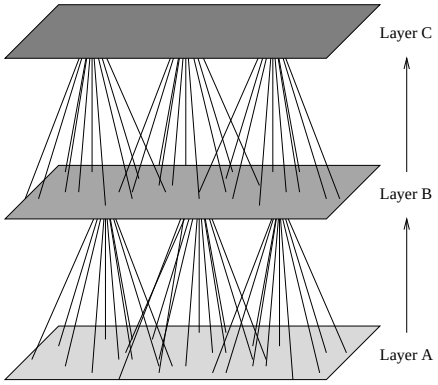


Figure 1.6: Structure of Linsker's neural network. The interaction between layers is strictly feed-foward.

where V_i is the set of all neurons that supply input to neuron i. The constant $\beta' > 0$ in front of the input sum is certainly not identical in meaning to the noise parameter β in the earlier model, but there is some similarity in the effect both parameters have on the functioning of the network[6]. Two other differences with our model also surface here. First, the Linsker model is not fully connected, secondly the weights are asymmetric.

The architecture of the net consists of a number of layers with feed-forward interaction. Layer A provides input to layer B, B is input for C, etc. Within each layer the interactions can be symmetric. For the moment we do not consider the interactions within layers. The structure of the network is then pictured in figure 1.6. The neurons within each layer are supposed to be lying in a two-dimensional plane. The planes are stacked on top of each other. A neuron in plane M is the post-synaptic neuron of cells in layer L directly below M. For neuron $i \in M$ the set of these cells is called the *receptive field* of neuron i, and is denoted by V_i. The density of the cells in V_i depends on the distance of the cells from the vertical projection of neuron i from layer M to layer L. Linsker used an average density proportional to $\exp(-a^M r^2)$, where r is the distance from the projection of neuron i. The exact results will be dependent on a^M, but the essential behavior of the network does not dependent on the choice of a Gaussian synapse density. A flat 'pill-box' density function yields roughly the same behavior [24].

Another difference is that the connections change on the same time-scale as the neurons. Each update of the neural states evokes a change in the weights, but it is assumed that this change is very small. The weight between neuron l of layer L and neuron m of the layer M above changes according to a generalized Hebbian associative rule given by

$$\Delta J_{ml} = k_a + k_b(s_m - t_M)(s_l - t_L), \tag{1.56}$$

where k_a,k_b,t_M and t_L are constants and $k_b > 0$ to make it Hebbian.

As mentioned before, Hebbian learning without any constraints, typically causes the modulus of the weights to grow beyond all limits. To solve this problem, strong boundaries are applied to the weights. When weights reach one of the boundary values -1 or +1, they get stuck at that value. Linsker divided the synapses in two groups, inhibitory, with weight values ranging from -1 to 0, and excitatory, ranging from 0 to 1. This was done to make the model more biologically plausible. However for the behavior of the network this is not an essential feature.

Now assume that the cells in layer A, which do not get any input from within the network, provide a random, spontaneous, uncorrelated activity. At each time step a new random pattern is produced in layer A. The new pattern is fed to the neurons in layer B, and from there on to higher levels. After the neurons in the highest level have been updated, the weights are changed by (1.56). As the weights are assumed to change only by very small amounts at each time step,

[6]In equilibrium, $\langle \sigma_i \rangle = \langle \tanh \beta h_i(\vec{\sigma}) \rangle$ holds for the stochastic network. For small $\beta h_i(\vec{\sigma})$, this equation is approximately: $\langle \sigma_i \rangle = \beta h_i(\langle \vec{\sigma} \rangle)$. Using linear neurons can be seen as modeling $\langle \sigma_i \rangle$ instead of σ_i.

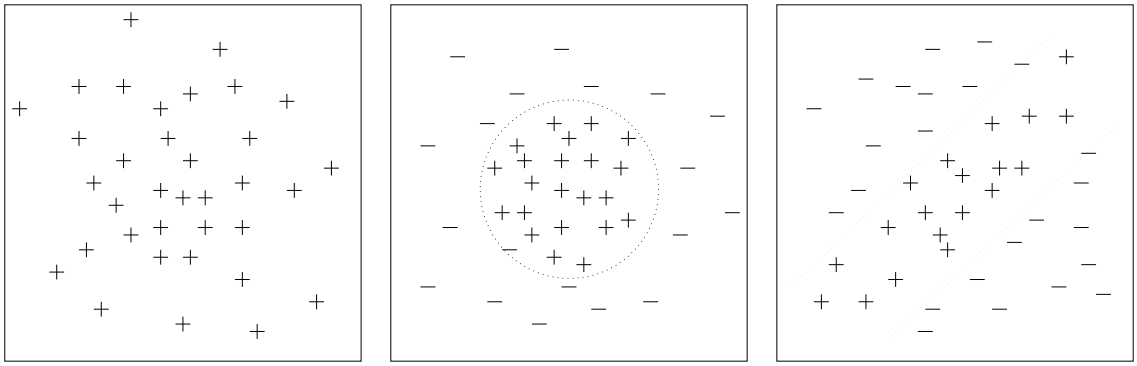


Figure 1.7: The input connections of neurons in three different layers of Linsker's neural network model: layer B, no spatial differentiation occurs; layer C, an on-center cell has evolved; layer G, a bilobed cell, sensitive to lines oriented in the /-direction.

we can accumulate the change of a weight over a large number of time steps and then divide by the number of steps. If the time between the presentations of new random patterns in layer A is very small with respect to the time scale on which the weights make notable changes, this average yields a time derivative of the weight.

When we insert the neuron update function (1.55) for the neurons in layer B into (1.56) and introduce a new time scaling we find for the mean rate of change of weight J_{ml}:

$$\frac{d}{dt}J_{ml}(t) = \underbrace{k_a - k_b(\vartheta + t_M)(\langle s_l\rangle - t_L)}_{k_1} + k_b\beta \sum_{l'\in V_m} J_{ml'}(t)\Big(\underbrace{\langle(s_{l'} - \langle s_{l'}\rangle)(s_l - \langle s_l\rangle)\rangle}_{Q^L_{ll'}} + \underbrace{\langle s_{l'}\rangle(\langle s_l\rangle - t_L)}_{k_2}\Big) \tag{1.57}$$

The bracket $\langle\cdot\rangle$ denote average over a sample of noise patterns generated in layer A. Linsker assumed the averaged activation of a neuron to be site independent, i.e. $\langle s_l\rangle = \langle s'_l\rangle$ for all neurons l and l' in layer L. This assumption makes the constants k_1 and k_2 independent of the sites l and l' but not layer independent.

The matrix Q^L is the correlation matrix of layer L. The linearity of the neuron activation and the linear Hebb rule have caused the correlation within layer L to have a very direct effect on the development of the connections between neurons in layer L to the higher layer M.

1.7.2 Results

Linsker chose for each layer the constants k_1, k_2 (not the k_a and k_b!) and the size for the dendritic trees, i.e. the width of the Gaussian distribution of the connections and starting simulating equation (1.57). For large ranges in the parameter space one gets interesting results for layer C and higher layers. In this subsection we very briefly describe the results and give a qualitative explanation.

In layer A random uncorrelated patterns are produced and therefore $Q^A = 0$. The evolution of the weights connecting layer A to B depends only on k_1 and k_2 and no spatial structure develops in the connections to neurons in layer B. Either all connections with neurons in the receptive fields are excitatory (positive) or all are inhibitory (negative), see figure 1.7.

Neurons in layer B that are close neighbors share a large portion of their receptive fields. As a result they are positively correlated. The Q^B matrix will be positive on distances of the order of the dendritic tree width and will be zero at greater distances. For the evolution of the weights from layer B to C this can have a remarkable effect. Consider a neuron c in layer C. If $k_1 > 0$ then initially all weights J_{cb} grow as fast. The largest portion of the input of cell c is coming from neurons located close to the center of its dendritic tree. These input neurons are positively

correlated and neuron c starts to mirror the behavior of the center of its receptive field. The neurons on the border of the receptive field on neuron c do not have a high correlation with most other neurons in the receptive field. If $k_2 < 0$, the sum of the weights stabilizes and a competition arises for excitatory weights. Due to the Hebbian learning, the neurons with high correlation with neuron c grow faster and therefore take the available excitatory weights. The neurons outside the center lose the competition and become inhibitory towards neuron c. When all weights have reached the boundary values ± 1, neuron c has become a *on-center* cell, with excitatory connections near the center of its receptive field and inhibitory connections near the border (see figure 1.7). A more rigorous analysis is done in the continuum limit by MacKay and Miller [24].

In higher layers the correlation matrix is no longer non-negative. The correlation between close neighbor neurons will be positive, but on a somewhat larger distance the correlation becomes negative. Neurons far apart still exhibit no correlation. The mexican-hat like form of the correlation as function of the distance can lead to the evolution of cells with a non-circular symmetric distribution of the weights over its receptive fields (see figure 1.7). These neurons are orientation sensitive. The appearance of these neurons is also discussed in [24]

1.7.3 Hebbian based self-organization

The appearance of the orientation selective neurons caused most of the interest in the work of Linsker, because precisely this kind of neurons had been found in young kittens [43]. The kittens seem to come into this world possessing neurons that respond to bars of a certain orientation, even though they have never seen a straight line in the outside world. In cats and macaque monkeys on-center neurons, also present in Linsker's model, have been identified as well. Linsker was not the first to offer a possible explanation for the development of both types of neurons, but earlier attempts to explain the orientation selective neurons needed the explicit presentation of bars to the network.

In the third article in the series [23] Linsker also assumed the presence of intra-layer interactions. These interactions are of a mexican hat type: positive on short distances, negative on somewhat longer distances and vanishing for large distances. This resulted in a clustering of neurons that are sensitive to bars of almost the same orientation. A feature that is also found in cats. Miller, Keller and Stryker [28] used a model very similar to Linsker's to explain the onset of occular dominance patches in many adult mammals. These patches are groups of cells that, although initially all are receiving input from both eyes, have specialized in processing information from just on eye.

These and other results show that a low level self-organization can develop in neural network models that share three basic properties:

i. Hebbian learning
ii. Spatial structure in the connections
iii. Linear activation of the neurons

1.8 The Final Formulation of the Model

I do not see why the linear activation of the neurons, the third item in the list at the end of the previous section, might be essential for the development of self-organization. The analysis and simulation of linear neurons consist of simple matrix multiplication and are easy in comparison to the model of the neuron we have discussed in the lion share of this introduction. This is the primary reason why Linsker and others use this type of neurons. In this thesis I want to make a start in analyzing neural networks that share the first two properties of the list above but have Ising-type neurons.

In section 1.5 we have already precisely explained how we will deal with Hebbian learning. Spatial structure we introduce by dividing the N neurons into Λ clusters. The clusters, labeled by Greek indices λ and κ and occasionally ρ, each consist of $V \equiv N/\Lambda$ neurons. The important assumption that we make in order to be able do calculations is that the decay and the bias of the connection between two neurons is depends only on the clusters in which the neurons are

contained, i.e. for all neurons i in cluster λ and all neurons j in cluster κ (possibly the same cluster) we assume

$$\mu_{ij} = \mu_{\lambda\kappa}, \qquad A_{ij} = A_{\lambda\kappa}. \tag{1.58}$$

By this assumption the information of the setup of the network can be substantially reduced from two $N \times N$-matrices to two $\Lambda \times \Lambda$-matrices.

The inverse decay matrix $\mu_{\lambda\kappa}$ is a powerful handle to create a topology for the fully-connected network. If $\mu_{\kappa\lambda} = 0$ then the connections between the clusters κ and λ will all vanish. Choosing the strength of the decay is certainly different from setting the synapse or connection density as was done in Linsker's model. The latter means setting the number of connections for each two clusters. This is possible by making the decay strength a random variable with the distribution specified for each combination of clusters. The analysis of this approach will have a similar structure (i.e. two level replica theory) as the analysis of Penney and Sherrington [36] of a network where the biases are random variables. This line is not pursued in this thesis.

I have chosen the term $-A_{ij}$ in the Langevin equation to denote the bias. This is to suggest that we are mainly interested in negative biases, $A_{ij} > 0$. If an external field is applied to align all spins, the bias creates a competition between the ferromagnetic and anti-ferromagnetic behavior. At this stage we can just express our hope that this will cause interesting behavior.

Making a feed-forward network with symmetric interactions is not possible. However, the structure of a feed-forward network with intra-layer symmetrical connections can be approximated by choosing convenient external fields and decays for intercluster connections: consider the clusters 1,2,3. By applying external fields to cluster 1 one can set the neurons. Choose a small (with respect to the external fields) inverse decay between cluster 1 and 2. The small connection strength between cluster 1 and 2 has the effect that the neurons of cluster 1 are not influenced very much by cluster 2. Cluster 2, however, does feel the input of the cluster 1 neurons. Choosing the connection strength (i.e. the inverse decay) between cluster 2 and 3 much smaller than the strength between 1 and 2 creates the same influence relation between clusters 2 and 3 as there is between 1 and 2.

We will use this explicit structure only somewhere in chapter three. For now just assume the neural net to be clustered in Λ clusters of equal size and its configuration described by certain yet unspecified $\Lambda \times \Lambda$-matrices μ and A.

Chapter 2

Order Parameters

Having set up the model and reached an expression for the partition function we are at the proper starting point for an exploration of the neural network in the framework of statistical mechanics. In this chapter we completely leave the biology and become entangled in the calculation of the free energy of the system. In this calculation order parameters emerge, which hold the key to the behavior of the system on a macroscopic scale.

Before we begin with this task, we start with a short review of the *replica method*. This replica method we will use excessively in the calculation of the partition function.

2.1 Replica Method

In the utopian world where mathematics is trivial, the knowledge of the energy as a function of the microscopic states enables the physicist to predict the average value of any observable expressable as function of the microscopic states of an N particle ergodic system immersed in a heat bath of temperature $1/k\beta$. Repeated measurements of the observable A will average to

$$\langle A \rangle = \frac{\sum_\sigma A(\sigma) \exp -\beta E(\sigma)}{\sum_\sigma \exp -\beta E(\sigma)}. \tag{2.1}$$

The denominator of the right hand side fraction is the *partition function*, $Z_N = \sum_\sigma \exp -\beta E(\sigma)$. The partition function is part of the definition of a more important thermodynamic variable, the *free energy*

$$F_N = -\frac{1}{\beta} \log Z_N. \tag{2.2}$$

Any observable which, as a function of the microscopic state variables, is expressible in parts and derivatives of the microscopic energy function, can be expressed as a function of this free energy and its derivatives. The calculations in statistical mechanics often come down to calculation of the free energy, or actually the thermodynamic limit of the free energy per particle

$$f = \lim_{N \to \infty} F_N / N. \tag{2.3}$$

The free energy per particle does not scale with the system size and is therefore called an *intensive* observable, whereas the free energy itself is an *extensive* observable that scales as N for large systems. Looking only at the thermodynamic limit is not something we do because we fancy large numbers, but is a necessity in this non utopian world in order to be able to actually do the calculations.

In the nineteen seventies theorists proposed several models to account for the measured behavior of the earlier mentioned spin-glasses. Success came for models with frozen random disorder. The bonds between spins were taken to be independent random variables. The natural and necessary assumption for anyone trying to calculate the properties of such a system is the *self-averaging* of

the free energy in thermodynamic limit. Self-averaging means that as the number of spins tends to infinity the specific realization of their interactions will be of decreasing importance to the value of the free energy, i.e. for large N almost all realization of the weights will yield the same free energy. This assumption, which takes the form

$$\overline{F_N^2} - \overline{F_N}^2 = \mathcal{O}(N^{-1}), \quad N \to \infty, \tag{2.4}$$

where $\bar{\cdot}$ denotes the average over all the possible bond realizations, can be proven by 'standard thermodynamical arguments'. In the limit of an infinite number of spins the chance will be one that only the moments of the bond distribution will matter to the outcome of the free energy of a system. Note that this does not imply that if we take two particular possible realizations of the interactions, the free energy will be the same. A bond distribution with zero mean and non-zero variance allows the possibility of realizations with only positive bonds and with only negative bonds. The system with only positive couplings will behave like a ferromagnet in the zero temperature limit, whereas this is not expected for the second case. The free energy as a function of the system parameters will therefore differ significantly for these two instances. It is therefore not a good idea to actually choose a realization, calculate the free energy and hope that this will be a valid value in general. One needs to take an average of the free energy over all possible realizations of the disorder. Although sometimes hard to imagine, physicists are trying to explain real system and that means that the thermodynamic limit should be taken as late as possible,

$$f = \lim_{N\to\infty} \frac{1}{N}\overline{F_N} = -\lim_{N\to\infty} \frac{1}{\beta N}\overline{\log Z_N}. \tag{2.5}$$

Averaging before taking the logarithm would not have posed as big a problem as this would have been equivalent to annealed disorder instead of frozen or quenched, which can in many cases be tackled with the same techniques used to evaluate the partition function without disorder. The presence of the logarithm however prevents the straightforward use of the usual thermodynamical tricks and methods.

The first to take this hurdle were Edwards and Anderson [10]. They used a primitive version of a trick which is now known as the *replica method*. A more explicit version of this replica trick was used to attempt solving the Sherrington Kirkpatrick (SK) spin glass model [39, 21] The replica trick essentially is the naive application of the following mathematical identity:

$$\overline{\log Z} = \lim_{n\to 0} \frac{1}{n} \log \overline{Z^n}. \tag{2.6}$$

Evaluating $\log Z$ is changed into evaluating Z^n. The n-fold product can, for n integer and positive, be written as

$$Z^n = \prod_{\alpha=1}^{n} Z_\alpha. \tag{2.7}$$

The right hand side can be seen as the partition function of n identical independent replicas of the system. The disorder average is not longer frustrated by the non-linearity of the logarithm and can be performed to lead to an effective partition function were the replicas are coupled to each other. Considering the replica number n to be integer and positive, this partition function can be transformed into a set of $\frac{1}{2}n(n+1)$ coupled non-linear equations in the thermodynamic limit. One can try to numerically find solutions to these equations. The results should then somehow be continued to small non-integer n to finally perform the limit n to zero. Sherrington and Kirkpatrick used the trivial continuation by just substituting zero for n. In this way they could explain some aspects of spin glasses, but there was still a discrepancy between theory and (numerical) experiments. A negative zero temperature entropy also puzzled the spin glass society.

It was not immediately clear what precisely caused the problems and how to solve them. Along the way Sherrington and Kirkpatrick also interchanged the limits N to infinity and n to zero. The correctness of the interchange was proved three years later by Van Hemmen and Palmer [14]. This pinpointed the problem to the continuation from integer n to small n. Around 1980, Parisi devised

a different continuation to small n. The uniqueness of the Parisi continuation still has not been proved, but its results are in line with most of the Monte Carlo calculations of the SK model. The results found by using the replica trick have also been confirmed by theoretical work following a mean field technique by Thouless, Anderson and Palmer [42]. These confirmations elevated the 'replica trick' into the 'replica method', which is now used in problems ranging from traveling salesman to molecule folding.

2.2 Calculation of the Free Energy

The above intermezzo would have been pointless if we had not encountered a power of the partition function Z_β for the fixed weights system in the expression 1.51 for the partition function $\tilde{Z}$ of the entire system of changing neuron states and weights. Now we can use this replica method, described above, to simplify the overall partition function. During the calculation we will meet the same problems that earlier hampered the spin glass community.

We start by assuming $n \equiv \tilde{\beta}/\beta$ is integer and introduce the n replicas, labeled by the Greek letter α. The partition function (1.51) takes the form:

$$\tilde{Z} = \int \left[\prod_{i<j} dJ_{ij} N^{\frac{1}{2}}\right] \sum_{\vec{\sigma}} \exp\left[\beta \sum_{\alpha=1}^{n} \left(\sum_{i<j} \sigma_i^\alpha J_{ij} \sigma_j^\alpha + \sum_i \theta_i \sigma_i^\alpha\right) - \frac{1}{2}\tilde{\beta} N \sum_{i<j} \frac{J_{ij}^2}{\mu_{ij}} - \tilde{\beta} \sum_{i<j} A_{ij} J_{ij}\right],$$

where the $\sum_{\vec{\sigma}}$ means a sum over all 2^N possible states of all the n replicas of the system. In what follows $\vec{x}$ will be shorthand for $(x^1, \ldots, x^n)$, the vector of the n replica instances of a certain state variable x.

Next we make the substitutions $z_{ij} = (\tilde{\beta} N/\mu_{ij})^{1/2} J_{ij}$,

$$\begin{aligned} \tilde{Z} &= \int \left[\prod_{i<j} dz_{ij} \left(\frac{\mu_{ij}}{\tilde{\beta}}\right)^{\frac{1}{2}}\right] \sum_{\vec{\sigma}} \exp\left[\beta \sum_i \sum_\alpha \theta_i \sigma_i^\alpha\right] \\ &\quad \times \exp \sum_{i<j} \left[\frac{\beta}{\sqrt{N\tilde{\beta}}} \sum_\alpha \sqrt{\mu_{ij}} \sigma_i^\alpha \sigma_j^\alpha z_{ij} - \frac{1}{2} z_{ij}^2 - \sqrt{\frac{\tilde{\beta}}{N}} \sqrt{\mu_{ij}} A_{ij} z_{ij}\right], \end{aligned} \tag{2.8}$$

and subsequently perform the Gaussian integrations over the transformed weights:

$$\begin{aligned} \tilde{Z} &= \left[\prod_{i<j} \left(\frac{2\pi\mu_{ij}}{\tilde{\beta}}\right)^{\frac{1}{2}}\right] \sum_{\vec{\sigma}} \exp\left[\beta \sum_i \sum_\alpha \theta_i \sigma_i^\alpha\right] \\ &\quad \times \exp \sum_{i<j} \left[\frac{\beta^2}{2N\tilde{\beta}} \mu_{ij} \sum_{\alpha,\beta} \sigma_i^\alpha \sigma_i^\beta \sigma_j^\alpha \sigma_j^\beta - \frac{\beta}{N} \mu_{ij} A_{ij} \sum_\alpha \sigma_i^\alpha \sigma_j^\alpha + \frac{\tilde{\beta}}{2N} \mu_{ij} A_{ij}^2\right]. \end{aligned} \tag{2.9}$$

Completing the sum over i and j, we have:

$$\begin{aligned} \tilde{Z} &= \left[\prod_{i<j} \left(\frac{2\pi\mu_{ij}}{\tilde{\beta}}\right)^{\frac{1}{2}}\right] \exp\left[-\frac{\beta n}{4N} \sum_i \mu_{ii} - \frac{\tilde{\beta}}{4N} \sum_i A_{ii}^2 \mu_{ii} + \frac{\tilde{\beta}}{4N} \sum_{i,j} A_{ij}^2 \mu_{ij}\right] \\ &\quad \times \sum_{\vec{\sigma}} \exp \sum_{i,j} \left[\frac{\beta^2}{4N\tilde{\beta}} \mu_{ij} \sum_{\alpha,\beta} \sigma_i^\alpha \sigma_i^\beta \sigma_j^\alpha \sigma_j^\beta - \frac{\beta}{2N} \mu_{ij} A_{ij} \sum_\alpha \sigma_i^\alpha \sigma_j^\alpha\right] \exp\left[\beta \sum_i \sum_\alpha \theta_i \sigma_i^\alpha\right]. \end{aligned} \tag{2.10}$$

Recall that to account for the spatial structure of the system due to μ_{ij}, we have divided the neural network into a large number Λ of non-overlapping clusters of adjacent neurons. The set I_λ consists of the indices of the neurons contained in cluster λ. All clusters consist of the same number

of neurons $V \equiv N/\Lambda$. We have assumed that the decay rate and the bias of the weight between two neurons are functions of their containing cluster only: $\mu_{ij} = \mu(\lambda(i), \lambda(j))$ and $A_{ij} = A(\lambda(i), \lambda(j))$.

Now it is time to start isolating the sum over states. For this reason, we plug in a set of variables m_λ^α and $q_\lambda^{\alpha\beta}$ for the magnetization of replica α and the overlap between the replicas α and β in cluster λ respectively.

These variables will be our handle to analyzing the equilibrium state of the system. Later it will become clear that they are closely linked to the parameters characterizing order in the system. We anticipate this and already call them 'order parameters'. A justification for this nomenclature follows in the next paragraph. Ignoring for the moment some prefactors,

$$\begin{aligned}
\tilde{Z} \;\propto\; & \int \left[\prod_{\substack{\alpha<\beta \\ \lambda}} dq_\lambda^{\alpha\beta}\right] \left[\prod_{\substack{\alpha \\ \lambda}} dm_\lambda^\alpha\right] \exp\left[\frac{\tilde{\beta}V^2}{4N}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}A_{\kappa\lambda}^2 - \frac{\tilde{\beta}V}{4N}\sum_\kappa \mu_{\kappa\kappa}A_{\kappa\kappa}^2\right] \\
& \times \exp\left[\frac{N\beta^2}{4\Lambda^2\tilde{\beta}}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}\sum_{\alpha,\beta} q_\kappa^{\alpha\beta}q_\lambda^{\alpha\beta} - \frac{N\beta}{2\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}A_{\kappa\lambda}\sum_\alpha m_\kappa^\alpha m_\lambda^\alpha + \frac{N\beta}{\Lambda}\sum_\kappa \theta_\kappa \sum_\alpha m_\kappa^\alpha\right] \\
& \times \sum_{\vec{\sigma}} \left[\prod_{\substack{\alpha<\beta \\ \lambda}} \delta\left(q_\lambda^{\alpha\beta} - \frac{1}{V}\sum_{i\in I_\lambda}\sigma_i^\alpha\sigma_i^\beta\right)\right]\left[\prod_{\substack{\alpha \\ \lambda}} \delta\left(m_\lambda^\alpha - \frac{1}{V}\sum_{i\in I_\lambda}\sigma_i^\alpha\right)\right]. \qquad (2.11)
\end{aligned}$$

In an only partly successful attempt to keep the equations clear we have introduced an surplus of n local constants:

$$q_\lambda^{\alpha\alpha} = 1$$

The Dirac delta function owes its entrance to the introduction of the order parameters. Now we substitute for the delta function its integral representation,

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} d\hat{x}\exp[i\hat{x}x].$$

As we are about to take the limits of N and Λ going to infinity, we have to make sure we introduce the new integration variables with the right scaling properties. This results in

$$\begin{aligned}
\tilde{Z} \;\propto\; & \int \left[\prod_{\substack{\alpha<\beta \\ \lambda}} dq_\lambda^{\alpha\beta}\right]\left[\prod_{\substack{\alpha<\beta \\ \lambda}} d\hat{q}_\lambda^{\alpha\beta}\frac{N}{2\pi\Lambda}\right]\left[\prod_{\substack{\alpha \\ \lambda}} dm_\lambda^\alpha\right]\left[\prod_{\substack{\alpha \\ \lambda}} d\hat{m}_\lambda^\alpha\frac{N}{2\pi\Lambda}\right] \\
& \times \exp\left[\frac{\tilde{\beta}N}{4\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}A_{\kappa\lambda}^2 - \frac{\tilde{\beta}}{4\Lambda}\sum_\kappa \mu_{\kappa\kappa}A_{\kappa\kappa}^2\right]\exp N\left[\frac{i}{\Lambda}\sum_{\substack{\alpha<\beta \\ \lambda}}\hat{q}_\lambda^{\alpha\beta}q_\lambda^{\alpha\beta} + \frac{i}{\Lambda}\sum_{\substack{\alpha \\ \lambda}}\hat{m}_\lambda^\alpha m_\lambda^\alpha\right] \\
& \times \exp N\left[\frac{\beta^2}{4\Lambda^2\tilde{\beta}}\sum_{\substack{\alpha,\beta \\ \kappa,\lambda}}\mu_{\kappa\lambda}q_\kappa^{\alpha\beta}q_\lambda^{\alpha\beta} - \frac{\beta}{2\Lambda^2}\sum_{\substack{\alpha \\ \kappa,\lambda}}\mu_{\kappa\lambda}A_{\kappa\lambda}m_\kappa^\alpha m_\lambda^\alpha + \frac{\beta}{\Lambda}\sum_{\substack{\alpha \\ \kappa}}\theta_\kappa m_\kappa^\alpha\right] \\
& \times \sum_{\vec{\sigma}}\exp\left[-i\sum_{\substack{\alpha<\beta \\ \lambda}}\hat{q}_\lambda^{\alpha\beta}\sum_{i\in I_\lambda}\sigma_i^\alpha\sigma_i^\beta - i\sum_{\substack{\alpha \\ \lambda}}\hat{m}_\lambda^\alpha\sum_{i\in I_\lambda}\sigma_i^\alpha\right].
\end{aligned}$$

The advantage of all these manipulations is that we are now able to reduce the sum over the states of all neurons to a sum over the states of just a single neuron and its replicas,

$$\sum_{\vec{\sigma}}\exp\left[-i\sum_{\substack{\alpha<\beta \\ \lambda}}\hat{q}_\lambda^{\alpha\beta}\sum_{i\in I_\lambda}\sigma_i^\alpha\sigma_i^\beta - i\sum_{\substack{\alpha \\ \lambda}}\hat{m}_\lambda^\alpha\sum_{i\in I_\lambda}\sigma_i^\alpha\right] = \prod_\lambda\left(\sum_{\vec{\sigma}_1}\exp\left[-i\sum_{\alpha<\beta}\hat{q}_\lambda^{\alpha\beta}\sigma_1^\alpha\sigma_1^\beta - i\sum_\alpha\hat{m}_\lambda^\alpha\sigma_1^\alpha\right]\right)^{\frac{N}{\Lambda}}$$

Here the mean field character of the analysis shows itself clearly. The interactions of all the original neurons are effectively replaced by the N-fold product of an interaction term due to n copies of one neuron interacting with a mean field.

Now we are ready to take the limit of N to infinity. Recall that the proper quantity to look at in this limit is not the partition function, nor the free energy, but the free energy per neuron:

$$\tilde{f} = \lim_{N\to\infty} \frac{\tilde{F}}{N} = -\lim_{N\to\infty} \frac{1}{\tilde{\beta}N} \log \tilde{Z}. \tag{2.12}$$

We now write the free energy per neuron in the form

$$\tilde{f} = -\lim_{N\to\infty} \frac{1}{\tilde{\beta}N} \log \int \left[\prod_{\substack{\alpha<\beta \\ \lambda}} dq_\lambda^{\alpha\beta}\right] \left[\prod_{\substack{\alpha<\beta \\ \lambda}} d\hat{q}_\lambda^{\alpha\beta}\right] \left[\prod_{\substack{\alpha \\ \lambda}} dm_\lambda^\alpha\right] \left[\prod_{\substack{\alpha \\ \lambda}} d\hat{m}_\lambda^\alpha\right] \exp[-N\tilde{\beta}\phi(q,\hat{q},m,\hat{m})],$$

where the difficult parts have been hidden in the function ϕ:

$$\begin{aligned}
\phi(q,\hat{q},m,\hat{m}) \;=\; & -\frac{1}{4\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa,\lambda}A_{\kappa,\lambda}^2 - \frac{\beta^2}{4\tilde{\beta}^2\Lambda^2}\sum_{\kappa\lambda}\mu_{\kappa\lambda}n \\
& -\frac{\beta^2}{2\tilde{\beta}^2\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}\sum_{\alpha<\beta} q_\kappa^{\alpha\beta}q_\lambda^{\alpha\beta} - \frac{i}{\tilde{\beta}\Lambda}\sum_{\substack{\alpha<\beta \\ \lambda}} \hat{q}_\lambda^{\alpha\beta}q_\lambda^{\alpha\beta} \\
& +\frac{\beta}{2\tilde{\beta}\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}A_{\kappa\lambda}\sum_\alpha m_\kappa^\alpha m_\lambda^\alpha - \frac{\beta}{\tilde{\beta}\Lambda}\sum_{\substack{\alpha \\ \kappa}}\theta_\kappa m_\kappa^\alpha - \frac{i}{\tilde{\beta}\Lambda}\sum_{\substack{\alpha \\ \lambda}}\hat{m}_\lambda^\alpha m_\lambda^\alpha \\
& -\sum_\lambda \frac{1}{\tilde{\beta}\Lambda}\log\sum_{\vec{\sigma}_1}\exp\left[-i\sum_{\alpha<\beta}\hat{q}_\lambda^{\alpha\beta}\sigma_1^\alpha\sigma_1^\beta - i\sum_\alpha \hat{m}_\lambda^\alpha\sigma_1^\alpha\right].
\end{aligned} \tag{2.13}$$

The partition function is now in the perfect shape for using the *saddle-point method*. In appendix B the saddle-point method is stated in theorem 4 and a sketch of a proof is given. The usual way, see for instance [7, 37], in which the calculation continues, is to calculate the saddle-points of $\phi(q,\hat{q},m,\hat{m})$ and not to worry too much about the legitimacy of subsequently saying that the partition function is proportional to the value of ϕ in the saddle-points. I have tried to prove the legitimacy, but have failed to do so. The only justification I can give is that for the cases where an alternative way of calculating the partition function and the order parameters is available, the naïve iterated application of the saddle-point method gives the right answer. The alternative way is using a Gaussian integral as we have used before to avoid using the delta function. This route is considered in appendix C, where in addition its validity is proven under certain condition. The system we actually want to look at does not satisfy these conditions, as one of the conditions is that the $-\mu_{\kappa\lambda}A_{\kappa\lambda}$, interpreted as a $\Lambda\times\Lambda$-matrix, is positive definite. Other conditions are that $\mu_{\kappa\lambda}$ is a positive definite matrix and that for all clusters λ, $\mu_{\lambda\lambda}>0$ and $\mu_{\lambda\lambda}A_{\lambda\lambda}<0$.

We just do as if our noses are bleeding (a strange Dutch saying) and apply the saddle-point method. The saddle point integration yields $\tilde{f} = \text{extremum}\,\phi(q,\hat{q},m,\hat{m})$. In the saddle-point method the function ϕ is closely linked to the free energy. To make this connection explicit, we will often call ϕ the *precursor of the free energy*. When it is clear from the context whether we mean f or ϕ, we will succumb to the common, sloppy, practice of using just the word *free energy* for ϕ. As the derivatives of $\phi(q,\hat{q},m,\hat{m})$ with respect to q and m are zero in a saddle-point, we find

$$\begin{aligned}
\hat{q}_\lambda^{\alpha\beta} &= i\frac{\beta^2}{\tilde{\beta}\Lambda}\sum_\kappa \mu_{\kappa\lambda}q_\kappa^{\alpha\beta}, \\
\hat{m}_\lambda^\alpha &= -i\frac{\beta}{\Lambda}\sum_\kappa \mu_{\kappa\lambda}A_{\kappa\lambda}m_\kappa^\alpha + i\beta\theta_\lambda.
\end{aligned} \tag{2.14}$$

Here the symmetry of $\mu_{\kappa\lambda}$ and $A_{\kappa\lambda}$ in their indices is used.

As also the derivatives with respect to $\hat{q}$ and $\hat{m}$ are taken to be zero, we find in addition to the above conditions:

$$\begin{aligned} q_\lambda^{\gamma\delta} &= \left[\sum_{\vec{\sigma}_1} \sigma_1^\gamma \sigma_1^\delta \exp -\tilde{H}_\lambda(\hat{q},\hat{m})\right]\left[\sum_{\vec{\sigma}_1} \exp -\tilde{H}_\lambda(\hat{q},\hat{m})\right]^{-1}, \\ m_\lambda^\gamma &= \left[\sum_{\vec{\sigma}_1} \sigma_1^\gamma \exp -\tilde{H}_\lambda(\hat{q},\hat{m})\right]\left[\sum_{\vec{\sigma}_1} \exp -\tilde{H}_\lambda(\hat{q},\hat{m})\right]^{-1}, \\ \tilde{H}_\lambda(\hat{q},\hat{m}) &= i\sum_{\alpha<\beta} \hat{q}_\lambda^{\alpha\beta}\sigma_1^\alpha\sigma_1^\beta + i\sum_\alpha \hat{m}_\lambda^\alpha \sigma_1^\alpha. \end{aligned} \tag{2.15}$$

Combining these relations between the capped and uncapped variables with (2.14) we establish that $\phi(q,\hat{q},m,\hat{m})$ has an extremum at q,m if and only if q,m obey

$$\begin{aligned} \forall\lambda \quad \forall\gamma,\delta \quad q_\lambda^{\gamma\delta} &= \left[\sum_{\vec{\sigma}_1} \sigma_1^\gamma \sigma_1^\delta \exp -H_\lambda(q,m)\right]\left[\sum_{\vec{\sigma}_1} \exp -H_\lambda(q,m)\right]^{-1}, \\ \forall\lambda \quad \forall\gamma \qquad m_\lambda^\gamma &= \left[\sum_{\vec{\sigma}_1} \sigma_1^\gamma \exp -H_\lambda(q,m)\right]\left[\sum_{\vec{\sigma}_1} \exp -H_\lambda(q,m)\right]^{-1}, \\ H_\lambda(q,m) &= -\frac{\beta^2}{\tilde{\beta}\Lambda}\sum_{\substack{\alpha<\beta \\ \rho}} \mu_{\lambda\rho} q_\rho^{\alpha\beta}\sigma_1^\alpha\sigma_1^\beta + \frac{\beta}{\Lambda}\sum_{\substack{\alpha \\ \rho}} \mu_{\lambda\rho}A_{\lambda\rho}m_\rho^\alpha\sigma_1^\alpha - \beta\sum_\alpha \theta_\lambda\sigma_1^\alpha. \end{aligned} \tag{2.16}$$

These equations have the form of fixed-point equations and might have one or many solutions, depending on the system settings. In the steepest descent *approximation* all solutions corresponding to a local minimum in the free energy function will contribute to the free energy density, but in the limit of N to infinity the only contributing solutions will be those in which ϕ attains a global minimum.

2.2.1 A different view

We can also look at the derived equations from a different point a view as a neural network consisting of n sub-networks could yield the same results. This is most clear when we consider the expression (2.10) for the partition function of a one cluster network. Neglecting irrelevant prefactors, the partition function can be written as:

$$\tilde{Z} = \sum_{\vec{\sigma}} \exp\left[\frac{\beta}{4Nn}\mu\sum_{\alpha,\beta}\left(\sum_i \sigma_i^\alpha\sigma_i^\beta\right)^2 - \frac{\beta}{2N}\mu A\sum_\alpha\left(\sum_i \sigma_i^\alpha\right)^2 + \beta\sum_\alpha\theta\sum_i\sigma_i\right]. \tag{2.17}$$

Consider an entirely different system of $n\times N$ neurons subdivided in n clusters of N neurons. The neurons within a subsystem are labeled with the Roman indices i,j, the subsystems themselves are labeled by the Greek α and β. Within a subsystem the neurons are uniformly connected with the weight $-\mu A/N$. With the usual Ising Hamiltonian, the subsystems each contribute an internal energy of:

$$H_i(\sigma^\alpha) = -\frac{1}{N}\sum_{i<j}\sigma_i^\alpha\mu A\sigma_j^\alpha, \tag{2.18}$$

to the total energy. Furthermore, any two subsystems tend to align their neurons with respect to each other. This tendency is expressed by the energy term:

$$H_a(\sigma^\alpha,\sigma^\beta) = -\frac{\mu}{4Nn}\left(\sum_i \sigma_i^\alpha\sigma_i^\beta\right)^2 \tag{2.19}$$

This part of the Hamiltonian also allows two subsystems to be entirely anti-parallel in the ground state. This degeneracy can be removed in the thermodynamic limit by applying a small inhomogenuous external field. This system directly yields the above partition function.

2.3 Interpretation of the Order Parameters

An important notion in statistical mechanics is that of *order parameters*. A state or phase of a system can be characterized by the value of these special variables. The set of variables m^α and $q^{\alpha\beta}$ as they were introduced in the calculation of the partition function are referring explicitly to the replicas. These n replicas are nothing but a mathematical construction and have no *a priori* physical relevance. The label 'order parameter' which was given to them earlier is therefore a bit tendentious. We want order parameters to be measurable, at least in theory, variables of just the one original system we are studying. In this section we will establish a link between the replica order parameters and some real order parameters.

For a cluster of Ising-spins it is natural to look at the average magnetization of the cluster as a starting point for classifying the order:

$$m_\lambda \equiv \frac{1}{V}\sum_{i\in I_\lambda}\langle\sigma_i\rangle = \frac{1}{V}\sum_{i\in I_\lambda}\frac{\sum_\sigma \sigma_i \exp -\beta H(\sigma)}{\sum_\sigma \exp -\beta H(\sigma)} \tag{2.20}$$

At this point the ergodicity of the system becomes relevant. Another short excursion into the domain of the usual ferromagnet illustrates this.

The name of Marie Curie might have become obsolete as a measure of radiation, but the name of her husband Pierre is still very much in use in the description of the behavior of a ferromagnet. When held at a temperature much below the Curie temperature, a ferromagnet can have a definite magnetic moment. On a cloudy day or night this property of the ferromagnet is very much appreciated by sailor and boy scout. Looking at the Gibbs-Boltzmann distribution for the Ising-models for ferromagnetism this property is not at all so obvious. If there is no external field applied the energy of the system is invariant under the transformation $\forall i \quad \sigma_i \to -\sigma_i$. When calculating the magnetization of the system according to the definition above, for every configuration of σ there is an opposite configuration with identical Boltzmann weight. Straightforward calculation of the magnetization will yield zero.

Spontaneous magnetization has to be put into the equations manually by introducing a uniform external magnetic field. If this external field is positive and much larger than $1/N$, the energy of the state in which the spins are aligned to the field is much lower than the opposite configuration and we find a positive magnetization. After performing the limit of N to infinity, we can set the field to zero and regain the original system, but with a non zero magnetization. If we used a negative external field instead we would have found an negative magnetization.

The point of this excursion is that an infinite volume spin array is no longer necessarily *ergodic* and that the techniques should be adjusted to this situation. The stationary state is not the usual Gibbs-Boltzmann average $p(\sigma) \propto \exp -\beta H(\sigma)$, but a *pure* state in which the symmetry of the system is broken. A pure state is a state where only configurations belonging to a single valley in the free energy landscape have a non-zero probability[1]. Pure states arise because when the thermodynamic limit is taken, the mountains between the free energy valleys grow infinitely high and the dynamical system gets stuck in a single valley. In calculating a thermal average for such a system, the sum over states should no longer be performed over the entire configuration space but only over the fraction of configurations which can be reached by the system in a finite time. This is the ergodic component of the state space corresponding with the pure state. The Gibbs-Boltzmann state is a sum over all states and is a weighted average of these pure states.

In the infinite-range Ising-model for the ferromagnet we can stick to definition (2.20) for m with only a minor adjustment, because of our knowledge of the system symmetry responsible for

[1] A pure state can be told apart from the Gibbs state by looking at the correlation function [27].

the splitting of the Gibbs-Boltzmann distribution. Consider, for instance, the following solution:

$$m = \lim_{ext.field \downarrow 0} \left\{ \lim_{N\to\infty} \frac{1}{V} \sum_{i\in I_\lambda} \langle \sigma_i \rangle \right\} \tag{2.21}$$

The order of the limits is crucial. In finite systems there is no ergodicity breaking. That the compass needle nevertheless has a definite magnetic moment is due to the fact that it has a very large number of magnetic components. The chance of a flip in the magnetic field may not be zero, as in the infinite volume case, but is extremely small.

In the early eighties, it became clear in the study of infinite range spin-glasses (see for review [27, 11]) that not only the spin-flip symmetry causes more than one pure state to exist, but that frustration caused by the disorder in the spin-glasses might create a very complex free energy landscape with many valleys corresponding to pure states of a less trivial nature. Because we are actually studying a more general model, in which the Sherrington-Kirkpatrick spin-glass is just a special-case ($\tilde{\beta}/\beta$=0), we have to be prepared for the existence of multiple pure states. We therefore stop looking at *the* value of the magnetization and start looking at the distribution of possible values for different pure states.

$$P(m) = \overline{\langle \delta(m - \frac{1}{V}\sum_i \sigma_i) \rangle} \tag{2.22}$$

The variable m refers here to the magnetization of a certain cluster as does the sum over i, but for notational convenience we omit the cluster label. Writing out this formula, treating n as integer, we find the replicas appearing again:

$$\begin{aligned} P(m) &= \tilde{Z}^{-1} \int \left[\prod_{i<j} dJ_{ij} \right] \exp\left[-\tilde{\beta}\mathcal{H}(J) \right] Z_J^{-1} \sum_\sigma \delta(m - \frac{1}{V}\sum_i \sigma_i) \exp -\beta H(\sigma) \\ &= \tilde{Z}^{-1} \int \left[\prod_{i<j} dJ_{ij} \right] \exp\left[-\tilde{\beta}H_J(J) \right] \sum_{\vec{\sigma}} \exp\left[-\beta \sum_{\alpha=1}^{n} H(\sigma^\alpha) \right] \delta(m - \frac{1}{V}\sum_i \sigma_i^1) \end{aligned} \tag{2.23}$$

We can calculate this expression in exactly the same way that we calculated the free energy. In the beginning of this calculation we introduced the delta function for the order parameter m^1. As this delta function peaks at the same value as the delta function for m we find

$$P(m) = \frac{1}{|\mathcal{S}_0|} \sum_{(m_0^\alpha, q_0^{\alpha\beta}) \in \mathcal{S}_0} \delta(m - m_0^1) \tag{2.24}$$

where $\mathcal{S}_0$ is the set of all solutions of the saddle-point equations 2.16 with the minimal corresponding $\phi(q, \hat{q}, m, \hat{m})$. From the saddle-point theorem in appendix B, it is not clear whether we really have to consider all, some or just one of the solutions. To answer the question we need to analyze the landscape of the real part ϕ. For this particular ϕ, this is a task that I cannot perform. It seems logical not to favor one of the saddle-points above others and thus to include them all. For the value of the free energy the number of saddle-points on the path is not important in the thermodynamic limit (as long as it is a finite number).

Equation (2.24) is the key in interpreting the meaning of the variables m^α. The possible magnetizations of replica number 1, that is the possibles values of m^1 at which ϕ is in a lowest value saddle-point, are also the possible magnetizations of the real system. If there is just one saddle-point where ϕ takes its lowest value, then m^1 has just one definite value and the distribution of the magnetization is a delta function:

$$P(m) = \delta(m - m_0^1) \tag{2.25}$$

The existence of multiple minimal saddle-points corresponds to a less trivial distribution of the magnetization and thus to broken ergodicity.

A system without any external field has the spin-flip symmetry. From this we know, that if it is possible to encounter a non-zero magnetization m, then it is also possible to encounter a system with opposite magnetization $-m$. This should be reflected in the structure of the saddle-points and indeed it is (see e.g. [14]). If $\mu^\alpha = \pm 1$ for $\alpha = 1, 2, \ldots, n$, the *sign transformation*

$$T_\mu : \begin{cases} q^{\alpha\beta} \to \mu_\alpha \mu_\beta q^{\alpha\beta} & \hat{q}^{\alpha\beta} \to \mu_\alpha \mu_\beta \hat{q}^{\alpha\beta} \\ m^\alpha \to \mu_\alpha m^\alpha & \hat{m}^\alpha \to \mu_\alpha \hat{m}^\alpha \end{cases} \tag{2.26}$$

leaves $\phi(q, \hat{q}, m, \hat{m})$ as defined in (2.13) invariant. Including the identity there are 2^n sign transformations. A saddle-point $(m_0^\alpha, q_0^{\alpha\beta})$, which has at least one non-zero replica magnetization and all positive $q^{\alpha\beta}$, has $2^n - 1$ sign transformed sisters with identical value of ϕ. If all replica magnetizations are zero, then each saddle-point only has 2^{n-1} siblings, because the transformation with $\mu_\alpha = -1$ for all replicas α is the identity transformation. The function ϕ also is invariant to another, even more obvious, transformation, namely the replica permutation. Any relabeling of the replicas will not alter ϕ. There are $n!$ such permutations. It is common practice to identify saddle-points that are mapped to each other by these two types of transformation. Instead of the sum over all minimal saddle-points in (2.24), a sum is performed over saddle-points that are not a transformation of another one present in the sum. Now the symmetry in the replicas can be broken and (2.24) should manually be symmetrized:

$$P(m) \quad = \quad \frac{1}{n|\mathcal{S}_0'|} \sum_\gamma \sum_{(m_0^\alpha, q_0^{\alpha\beta}) \in \mathcal{S}_0'} \delta(m - m_0^\gamma) \tag{2.27}$$

where $\mathcal{S}_0'$ is the set of all solutions of the saddle-point equations (2.16) with the minimal corresponding $\phi(q, \hat{q}, m, \hat{m})$ that can not be transformed into each other by a replica permutation or sign transformation.

In the limit of an infinite number of neurons the saddle-point method used is exact and the only contributing saddle-points are the ones that give the lowest value for $\phi(q, \hat{q}, m, \hat{m})$. For a large but finite number of neurons, the saddle-point method is not exact but a good approximation. For a finite number N, all saddle-points contribute to the free energy. The non-minimal saddle-points can correspond to local minima in the free energy landscape. In the dynamical picture these solutions are the metastable states of the system. For N very large we find by using the saddle point approximation:

$$P(m) \quad \approx \quad \frac{1}{|\mathcal{S}|\tilde{Z}} \sum_{(m_0^\alpha, q_0^{\alpha\beta}) \in \mathcal{S}} \exp\left[-N\tilde{\beta}\phi(m_0^\alpha, q_0^{\alpha\beta})\right] \delta(m - m_0^1) \tag{2.28}$$

where $\mathcal{S}$ now is the set of all solutions of the saddle-point equations 2.16. The Gibbs-Boltzmann average magnetization can be split in a weighted sum of pure states corresponding to the different saddle-points:

$$m_{GB} \equiv \frac{1}{V} \sum_i \langle \sigma_i \rangle = \frac{1}{V} \sum_i \sum_a w_a \langle \sigma_i \rangle_a, \tag{2.29}$$

where a labels the saddle-points and

$$w_a = \tilde{Z}^{-1} \exp\left[-N\tilde{\beta}\phi(m_a^\alpha, q_a^{\alpha\beta})\right] \tag{2.30}$$

2.3.1 Edwards-Anderson order parameter

In the realm of spin glasses it was found that (for non-biased weights) there is a phase in which the magnetization of the entire system is zero, like in a ferromagnet above the Curie-temperature, but the magnetization of the individual spins is non-zero, quite unlike the hot ferromagnet. To characterize this behavior, Edwards and Anderson [10] introduced the following order parameter:

$$q_{EA} \equiv \lim_{t \to \infty} \lim_{N \to \infty} \overline{\langle \sigma_i(t_0) \sigma_i(t_0 + t) \rangle} \tag{2.31}$$

which now bears their name. If no external field is applied and the system stays ergodic, even in the infinite volume limit, this order parameter will be zero due to the spin-flip symmetry. If however the ergodicity is broken, there is no obvious reason why the correlation between the states of one spin will vanish. The infinite system is trapped in a, perhaps local, minimum of the free energy. The Edwards Anderson order parameter q_{EA} measures the average over all these minima of the squared local magnetization in such a valley. We use the roman a for describing a certain free-energy valley and the chance of starting in a valley by w_a:

$$q_{EA} = \overline{\sum_a w_a \langle \sigma_i \rangle_a^2} \tag{2.32}$$

The average $\langle \sigma_i \rangle_a$ is the average over an ensemble of systems with identical weights and all restricted to the same free energy valley or ergodic component in the infinite volume case. The weighted sum over the valleys is over all valleys of a given weight configuration. The $\overline{\cdot}$-average is an average over all configurations of the weights. Note that because of the average over the weights and the fact that the system setup parameters $A_{\kappa\lambda}$ and $\mu_{\kappa\lambda}$ solely depend on the clusters involved, we can also write

$$q_{EA} = \frac{1}{V} \sum_i \overline{\sum_a w_a \langle \sigma_i \rangle_a^2}, \tag{2.33}$$

in which the cluster label again is left out to avoid heavily indexed notation.

Although the notation might suggest otherwise, the Edwards-Anderson order parameter can not directly be identified with our order parameters $q^{\alpha\beta}$. These are more closely linked to the mean squared equilibrium magnetization. We will therefore study the distribution of the following variable

$$q \equiv \frac{1}{V} \sum_i \overline{\langle \sigma_i \rangle^2} = \frac{1}{V} \sum_i \overline{\left(\sum_a w_a \langle \sigma_i \rangle_a \right)^2} \tag{2.34}$$

This variable is different from q_{EA}, because here overlaps between two different (meta)stable states are contributing. Taking a very long time average squared magnetization of a system with a very large, but finite volume q is the outcome of the measurement as transitions between pure states are possible. On a shorter timescale one will measure q_{EA}.

In writing the distribution we use of the fact that the square of an average of a quantity in one system can be written as the averaged product of the same quantity in two identical uncoupled systems.

$$P(q) = \overline{\langle \delta(q - \frac{1}{V} \sum_i \sigma_i \sigma_i') \rangle}_{\sigma,\sigma'} \tag{2.35}$$

The identification of σ and σ' as the replicas σ^1 and σ^2 leads to the earlier introduced $n(n-1)/2$ order parameters $q^{\alpha\beta}$. Using $H_J(\{J_{ij}\})$ to denote all terms in the Hamiltonian $\mathcal{H}$ that do not depend on the neuron dynamics (terms coming from the bias and the decay), we can see the connection between the replicas and the physical copies in:

$$\begin{aligned}
P(q) &= \tilde{Z}^{-1} \int \left[\prod_{i<j} dJ_{ij} \right] \exp\left[-\tilde{\beta} H_J(\{J_{ij}\}) \right] Z_\beta^{-2} \sum_\sigma \delta(q - \frac{1}{V} \sum_i \sigma_i \sigma_i') \exp\left[-\beta H(\sigma) - \beta H(\sigma') \right] \\
&= \tilde{Z}^{-1} \int \left[\prod_{i<j} dJ_{ij} \right] \exp\left[-\tilde{\beta} H_J(\{J_{ij}\}) \right] \sum_{\vec{\sigma}} \exp\left[-\beta \sum_{\alpha=1}^n H(\sigma^\alpha) \right] \delta(q - \frac{1}{V} \sum_i \sigma_i^1 \sigma_i^2) \\
&= \frac{1}{|\mathcal{S}_0|} \sum_{(m_0^\alpha, q_0^{\alpha\beta}) \in \mathcal{S}_0} \delta(q - q^{12})
\end{aligned}$$

In the interpretation of m, we concluded that the possible magnetizations of the systems were reflected in the magnetizations of the replicas. Now we see that the possible overlaps between two

pure states of the system also are the overlaps between the replicas. And thus we have linked the replica order parameters $q^{\alpha\beta}$ and m^α, which we can try to calculate from the saddle-point equations, to the real order parameters m and q.

In the literature the expressions for $P(m)$ and $P(q)$ always are without the sum over the saddle-points. This usual formulation obscures the possible presence of several really different minimal saddle-points. It also gives a sense of reality to the *number* of replicas n as it looks like an average over n different possible solutions. This is a bit misleading, because the number of replicas does not have such a direct relation to the number of solutions. But what we do see is that the possible magnetizations of the replicas are also the possible magnetizations of the real system. We will assume here that all the minimal saddle-points are identical up to a permutation of the replicas. The degeneration due to the spin flip is removed by hand by considering only non negative overlaps. Now we can comply to the customary notation of $P(m)$ and $P(q)$:

$$\begin{aligned} P(m) &= \frac{1}{n}\sum_\alpha \delta(m - m^\alpha), \\ P(q) &= \frac{1}{n(n-1)}\sum_{\alpha\neq\beta} \delta(q - q^{\alpha\beta}), \end{aligned} \tag{2.36}$$

where m^α and $q^{\alpha\beta}$ are taken to be a minimal saddle-point of the free energy precursor ϕ.

2.4 Replica Symmetric Solution

In the article that initiated replica theory, Sherrington and Kirkpatrick [39] simply wrote: "Since the replicas are indistinguishable, we consider only the extremum of the exponential for which all the $[q^{\alpha\beta}, \alpha \neq \beta]$ are equal, as are all the $[m^\alpha]$." This is now known as the *Replica Symmetry Ansatz*.

$$\begin{aligned} m_\lambda^\alpha &= m_\lambda \\ q_\lambda^{\alpha\beta} &= (1 - \delta^{\alpha\beta})q_\lambda + \delta^{\alpha\beta} \end{aligned}$$

Replica Symmetric Solution (2.37)

At the time they did not deem it necessary to use more words to justify this assumption. The impossible negative entropy they found at zero temperature caused many people to look much more closely at the truth of the assumption. For positive integer value of n it was proved by Lieb that the minimal saddle-point was indeed a replica symmetric one. His proof was published by Van Hemmen and Palmer [14] and a generalised version can be found in appendix D. I have not been able to prove replica symmetry for a system with more than one cluster. For small real n the replica symmetric assumption is certainly not correct. We will come back to this point in section 2.6. We now first start exploring the implications of replica symmetry.

For the mean magnetization and the mean squared magnetization, replica symmetry results in:

$$\begin{aligned} \frac{1}{V}\sum_{i\in I_\lambda} \overline{\langle\sigma_i\rangle} &= m_\lambda \\ \frac{1}{V}\sum_{i\in I_\lambda} \overline{\langle\sigma_i\rangle^2} &= q_\lambda \end{aligned} \tag{2.38}$$

Both are now single delta functions. A Ising system having a weight configuration that only allows for one stable state (if the global spin flip symmetry is removed) will have delta function for its order parameters. Such a system also is ergodic. This leads one to suspect that replica symmetry holds only for ergodic Ising system, as is stated in [7]. I do not understand why this is true. Consider, for instance, the thermodynamic limit of an Hopfield network with two imprinted

patterns. If we choose the patterns orthonormal, the overlap between the patterns is zero. The energy belonging to both patterns is the same:

$$H(\vec{\xi}^1) = -\sum_{i<j}\left[\xi_i^1\xi_i^1\xi_j^1\xi_j^1 + \xi_i^1\xi_i^2\xi_j^2\xi_j^1\right] = -\frac{1}{2}N(N-1) = -\sum_{i<j}\left[\xi_i^2\xi_i^1\xi_j^1\xi_j^2 + \xi_i^2\xi_i^2\xi_j^2\xi_j^2\right] = H(\vec{\xi}^2). \tag{2.39}$$

This means that the power of the partition function Z_β^n, present in the partition function belonging to neural and weight dynamics, will have the same value no matter how we choose the replicas to represent the patterns. If we choose all replica's to represent one of the two patterns, we will have a replica symmetric solution as stable as any other distribution of the two patterns over the replicas. This choice leads to a single delta function for the order parameter q, but the two pattern Hopfield network is not ergodic in the thermodynamic limit. We see that replica symmetry does not automatically implies ergodicity. We will come back to physical interpretation of the Replica Symmetry Ansatz in subsection 2.4.1. For the moment we continue with the algebraic consequences.

The most important effect from a computational point of view of the Replica Symmetry Ansatz is the simplification of the saddle-point equations (2.16). The number of variables is reduced from $n + n(n-1)/2$ to a mere two. The equations are reduced to

$$\begin{aligned} q_\lambda &= \frac{\sum_{\vec{\sigma}_1}\sigma_1^1\sigma_1^2\exp -H_\lambda[q,m]}{\sum_{\vec{\sigma}_1}\exp -H_\lambda[q,m]}, \\ m_\lambda &= \frac{\sum_{\vec{\sigma}_1}\sigma_1^1\exp -H_\lambda[q,m]}{\sum_{\vec{\sigma}_1}\exp -H_\lambda[q,m]}, \end{aligned} \tag{2.40}$$

$$H_\lambda[q,m] = -\frac{\beta}{2n}\left(\sum_\alpha \sigma_1^\alpha\right)^2\sum_\kappa\frac{1}{\Lambda}\mu_{\lambda\kappa}q_\kappa + \beta\left(\sum_\alpha\sigma_1^\alpha\right)\left[\sum_\kappa\frac{1}{\Lambda}\mu_{\lambda\kappa}A_{\lambda\kappa}m_\kappa + \theta_\lambda\right].$$

Note that this H is slightly different from the earlier one in (2.16).

Having carried out a Gaussian integral a page or two before, we are now inserting another one back into the fixed-point equation for purpose of linearization, using

$$\int \mathcal{D}z\,\exp[xz] = \exp[\frac{1}{2}x^2], \quad \text{where} \quad \mathcal{D}z \equiv \frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}z^2]\,dz.$$

This results in a change of $\exp -H_\lambda[q,m]$:

$$\exp -H_\lambda[q,m] = \int \mathcal{D}\xi\,\exp\left(\sum_\alpha\sigma_1^\alpha\right)\left[\sqrt{\beta/n}\sqrt{Q_\lambda}\xi + \beta M_\lambda + \beta\theta_\lambda\right].$$

Here we have introduced two convenient abbreviations:

$$\begin{aligned} Q_\lambda &\equiv \frac{1}{\Lambda}\sum_\kappa\mu_{\lambda\kappa}q_\kappa, \\ M_\lambda &\equiv -\frac{1}{\Lambda}\sum_\kappa\mu_{\lambda\kappa}A_{\lambda\kappa}m_\kappa. \end{aligned} \tag{2.41}$$

At last we are able to carry out the remaining sum over $\vec{\sigma}_1$ and find:

$$\boxed{\begin{aligned} q_\lambda &= \left[\int\mathcal{D}\xi\,\cosh^n\Xi\tanh^2\Xi\right]\left[\int\mathcal{D}\xi\,\cosh^n\Xi\right]^{-1}, \\ m_\lambda &= \left[\int\mathcal{D}\xi\,\cosh^n\Xi\tanh\Xi\right]\left[\int\mathcal{D}\xi\,\cosh^n\Xi\right]^{-1}, \\ \Xi &= \xi\sqrt{\beta/n}\sqrt{Q_\lambda} + \beta M_\lambda + \beta\theta_\lambda. \end{aligned}} \tag{2.42}$$

This set of equations is the main result of this chapter. In the next chapter we will analyze their solutions in two simple configurations.

2.4.1 Mattis glass

In the SK spin glass model and our neural network model, the couplings are not confined to ± 1 and have values in $\mathbb{R}$. This has no impact on the concept of frustration, which remains very clear. If $J_{ij}\sigma_i\sigma_j < 0$ the bond is frustrated. The effect that Hebbian learning has on a system is the systematic removal of frustration. In a low weight temperature $\tilde{T}$ regime we expect an unbiased system to move to a highly unfrustrated weight configuration, much like the Mattis model. An unfrustrated, but random distribution of the weights, makes the term *Mattis glass* seem appropriate. However, in section 2.5 we will see that the fluctuation on the individual weights due to the noise are of order $N^{-1/2}$ whereas the the value of the weights in a Mattis magnet is of order N^{-1}. The fluctuations are such that there will always be frustrated bonds. The similarity between the Hebbian learning neuron model we are discussing and the Mattis magnet discussed in chapter one, is not the complete absence of frustration in the ground states, but the existence of a unique (up to a global spin flip) stable state contrasting with the possibly many ground or metastable states of a general frustrated system.

The assumption that the equilibrium state of our system resembles a Mattis-like state in certain parameter and temperature regions, does not only imply the ergodicity (up to the $\mathbb{Z}_2$-spin flip symmetry) of the spin dynamics, but also hints at the non-ergodicity of the weight dynamics. There are 2^{N-1} possible Mattis magnets, thus when we assume that our system in equilibrium resembles in a glassy sort of way one of these configurations, there are an equal number of possible equilibrium states or ergodic components.

In section 3.1.4 of the next chapter we will prove that the weight dynamics of a one cluster system are ergodic for any non-zero weight temperature. The proof given there can be easily extended to systems with any finite number of clusters. Because the system is ergodic, we will see frustration removal neither in the theoretical equilibrium expressions for the weight values nor in their long term measurements. However on short time scales (short with respect to the weight dynamics, still very long with respect to the spin dynamics), we might experience some effects of the Mattis-like behavior of the system.

If at some short time scale the weights indeed allow only one pair of stable states, it is not appropriate to perform the weight average over the entire configuration space when looking at an expression like $\langle\sigma_i\sigma_j\rangle$ for explanation of the measurements. One needs to perform the average only around one of the Mattis-magnet states. This is in sharp contrast to the SK-model, where the weights are not dynamical variables and have no preference for unfrustrated configuration. A division of the configuration space is in the SK-model shear nonsense.

In an average over the entirety of the weight space as is performed in the SK-model and in [35], nothing distinguishes one site from another. Therefore the magnetization and overlap of a given spin are identical to respectively the magnetization and overlap of the cluster containing the site, (cf. 2.38). Still assuming replica symmetry, this means:

$$\overline{\langle\sigma_i\rangle} = m_{\lambda_i}, \qquad \overline{\langle\sigma_i\rangle^2} = q_{\lambda_i}. \tag{2.43}$$

Let us define a *Mattis component* of the weight space belonging the patterns $\vec{\xi}$ and $-\vec{\xi}$ as the set of all weight configurations that have this pair of patterns as its only two ground states. In this way the weight space is divided in 2^{N-1} Mattis components and a spin glass component of configurations that have more than two ground states. When we average over just one particular Mattis component of the weightspace, hereafter this average is denoted by $\overline{\overline{\cdot}}$, we still expect this to be true for the overlap:

$$\overline{\overline{\langle\sigma_i\rangle^2}} = q_{\lambda_i} \tag{2.44}$$

This will not be the case for the magnetization, as some of the spins are consistently down, while others are consistently up (having removed the spin-flip symmetry). We assume that the modulus

of the average single site magnetization is near the square root of the averaged single site overlap:

$$\overline{\overline{\sigma_i}} = \pm\sqrt{q_{\lambda_i}}. \tag{2.45}$$

The number of plusses and minusses is of course dependent on around which Mattis configuration the average is taken and is subject to the constraint

$$\frac{1}{V}\sum_i \overline{\overline{\sigma_i}} = m_{\lambda_i}. \tag{2.46}$$

This condition implies the fractions of positive and negative spins to be

$$p(\overline{\overline{\sigma_i}} = \pm\sqrt{q_{\lambda_i}}) = \frac{1}{2}(1 \pm m_{\lambda_i}). \tag{2.47}$$

As a consequence the averages of the covariance of two spins differ:

$$\begin{aligned} \overline{\langle\sigma_i\sigma_j\rangle} &= \overline{\langle\sigma_i\rangle\langle\sigma_j\rangle} = m_{\lambda_i}m_{\lambda_j}, \\ \overline{\overline{\langle\sigma_i\sigma_j\rangle}} &= \overline{\overline{\langle\sigma_i\rangle\langle\sigma_j\rangle}} = \pm\sqrt{q_{\lambda_i}q_{\lambda_j}}. \end{aligned} \tag{2.48}$$

The first identity is only valid for the limit of N to infinity and expresses the vanishing of the correlation between two sites in a system with infinite range couplings. We will see that the difference of the two averages has consequences for the equilibrium values calculated for the weights.

A reminder is needed at the end of this section. The interpretation of the system as a Mattis-glass is only valid because we assume that the frustration removal effect is not completely overshadowed by the fluctuation of the weights. Whether the interpretation and its consequences given above are valid, remains to be checked by further analysis or numerical experiment.

I have encountered the term Mattis glass only in Anemüller [4]. Anemüller uses the name Mattis glass to describe the region of phase space where the Replica Symmetry Ansatz is assumed to be correct and a non-trivial solution of the self-consistency equations (2.42) is possible. In the rest of this thesis, I will use *Mattis glass* only in this sense, using the name *spin glass* phase for the region where replica symmetry does not hold.

2.5 From Order Parameters to Weights

It was stated earlier that the free energy (density) can be used to calculate properties of the system. Now it is time to put this wisdom into practice.

The presence of external fields in the free energy expression enables us to check the above interpretation of the replica order parameters. Using the expression (1.51) for the partition function, differentiating the free energy with respect to the external field shows:

$$-\Lambda\frac{\partial\tilde{f}}{\partial\theta_i} = \frac{1}{\tilde{\beta}N}\frac{\partial}{\partial\theta_i}\log\tilde{Z} = \overline{\langle\sigma_i\rangle}. \tag{2.49}$$

But performing the same differentiation when the partition function is of the form (2.10) and subsequently using the saddle-point method gives:

$$-\Lambda\frac{\partial\tilde{f}}{\partial\theta_\lambda} = -\Lambda\frac{\partial\phi}{\partial\theta_\lambda} = \frac{1}{n}\sum_\gamma m_\lambda^\gamma. \tag{2.50}$$

This confirms the result about the magnetization of the previous section. Along the same line we find with $i \in I_\lambda$:

$$n - n^2\overline{\langle\sigma_i\rangle}^2 + n(n-1)\overline{\langle\sigma_i\rangle^2} = \frac{1}{\beta^2}\frac{\partial^2}{\partial{\theta_i}^2}log\tilde{Z} = n - n^2(\frac{1}{n}\sum_\alpha m_\lambda^\alpha)^2 + n(n-1)\frac{1}{n(n-1)}\sum_{\alpha\neq\beta} q_\lambda^{\alpha\beta}$$

$$\Rightarrow \overline{\langle\sigma_i\rangle^2} = \frac{1}{n(n-1)} \sum_{\alpha\neq\beta} q_{\lambda_i}^{\alpha\beta} \tag{2.51}$$

Again this is in agreement with the earlier results.

The presence of the bias A_{ij} in the original partition function provides a useful handle on the distribution of the final weight values. In the SK spin glass model the distribution of the weights was put into equations *ab initio* and therefore it was no use calculating them. In our model the weights are dynamical variables and their equilibrium value distribution is not known from the start. In the modelling of spin glasses people do not really care how the weights are distributed. However, from the viewpoint of neural information processing the distribution of the weights might give one some information.

The first and second order derivatives of the free energy ($\tilde{F} = -\tilde{\beta}^{-1}\log\tilde{Z}$, where $\tilde{Z}$ is the original partition function from (1.51)) to the bias yield:

$$\begin{aligned} \frac{\partial\tilde{F}}{\partial A_{ij}} &= \overline{J_{ij}}, \\ -\frac{1}{\tilde{\beta}}\frac{\partial^2\tilde{F}}{\partial A_{ij}\partial A_{kl}} &= \left[\overline{J_{ij}J_{kl}} - \overline{J_{ij}}\,\overline{J_{kl}}\right]. \end{aligned} \tag{2.52}$$

For expression (2.10) for the same partition function, this gives the following list:

$$\begin{aligned} \frac{\partial\tilde{F}}{\partial A_{ij}} &= -\frac{1}{N}A_{ij}\mu_{ij} + \frac{\beta}{\tilde{\beta}N}\mu_{ij}\sum_{\alpha}\overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle}, \\ -\frac{1}{\tilde{\beta}}\frac{\partial^2\tilde{F}}{\partial A_{ij}\partial A_{ij}} &= \frac{1}{\tilde{\beta}N}\mu_{ij} + \left(\frac{\beta}{\tilde{\beta}N}\right)^2\mu_{ij}^2\sum_{\alpha,\beta}\left[\overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle\langle\sigma_i^\beta\sigma_j^\beta\rangle} - \overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle}\,\overline{\langle\sigma_i^\beta\sigma_j^\beta\rangle}\right], \\ -\frac{1}{\tilde{\beta}}\frac{\partial^2\tilde{F}}{\partial A_{ij}\partial A_{ik}} &= \left(\frac{\beta}{\tilde{\beta}N}\right)^2\mu_{ij}\mu_{ik}\sum_{\alpha,\beta}\left[\overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle\langle\sigma_i^\beta\sigma_k^\beta\rangle} - \overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle}\,\overline{\langle\sigma_i^\beta\sigma_k^\beta\rangle}\right], \\ -\frac{1}{\tilde{\beta}}\frac{\partial^2\tilde{F}}{\partial A_{ij}\partial A_{kl}} &= \left(\frac{\beta}{\tilde{\beta}N}\right)^2\mu_{ij}\mu_{kl}\sum_{\alpha,\beta}\left[\overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle\langle\sigma_k^\beta\sigma_l^\beta\rangle} - \overline{\langle\sigma_i^\alpha\sigma_j^\alpha\rangle}\,\overline{\langle\sigma_k^\beta\sigma_l^\beta\rangle}\right]. \end{aligned} \tag{2.53}$$

The spin averages, like $\langle\sigma_i^\alpha\rangle$, are identical to the spin averages of the original system in the sense that they use the same Hamiltonian $H = -\sum_{i<j}\sigma_i^\alpha J_{ij}\sigma_j^\alpha - \sum_i\theta_i\sigma_i^\alpha$. The replica indices refer to the indices used in the weight average.

If the entire Gibbs-Boltzmann average over the weight configuration is taken, this leads to expressions for the first and second moments of J_{ij}.

$$\begin{aligned} \overline{J_{ij}} &= -\frac{1}{N}A_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_j} + \frac{1}{Nn}\mu_{\lambda_i\lambda_j}\sum_{\alpha} m_{\lambda_i}^\alpha m_{\lambda_j}^\alpha, \\ \overline{J_{ij}^2} - \overline{J_{ij}}^2 &= \frac{1}{\tilde{\beta}N}\mu_{\lambda_i\lambda_j} + \frac{1}{N^2n^2}\mu_{\lambda_i\lambda_j}^2\sum_{\alpha,\beta}\left[q_{\lambda_i}^{\alpha\beta}q_{\lambda_j}^{\alpha\beta} - m_{\lambda_i}^\alpha m_{\lambda_j}^\alpha m_{\lambda_i}^\beta m_{\lambda_j}^\beta\right]. \end{aligned} \tag{2.54}$$

where λ_i is the cluster containing neuron i. The second terms in both these expression are not present in the SK-model. The Hebbian learning term in the Langevin equation does therefore certainly effect the weights. Another difference with the SK-model is that there is a finite correlation between the weights on one neuron:

$$\begin{aligned} \overline{J_{ij}J_{ik}} - \overline{J_{ij}}\,\overline{J_{ik}} &= \frac{1}{N^2n^2}\mu_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_k}\sum_{\alpha,\beta}\left[q_{\lambda_i}^{\alpha\beta}m_{\lambda_j}^\alpha m_{\lambda_k}^\beta - m_{\lambda_i}^\alpha m_{\lambda_j}^\alpha m_{\lambda_i}^\beta m_{\lambda_k}^\beta\right], \\ \overline{J_{ij}J_{kl}} - \overline{J_{ij}}\,\overline{J_{kl}} &= o(N^{-2}). \end{aligned} \tag{2.55}$$

Let us consider the average weight more closely for the case that replica symmetry holds. The full Gibbs-Boltzmann average will be

$$\begin{aligned} \overline{J_{ij}} &= -\frac{1}{N}A_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_j} + \frac{1}{N}\mu_{\lambda_i\lambda_j}\overline{\langle\sigma_i\rangle\langle\sigma_j\rangle} \\ &= -\frac{1}{N}A_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_j} + \frac{1}{N}\mu_{\lambda_i\lambda_j}m_{\lambda_i}m_{\lambda_j}. \end{aligned} \tag{2.56}$$

This Gibbs-Boltzmann average will be the outcome of measuring the value of a particular weight for a very long time. On a shorter timescale, I have given a reason to believe that the weight configuration is trapped near a certain Mattis-magnet configuration. A short measurement therefore will not yield the Gibbs-Boltzmann result, but an average over a restricted region of phase space.

$$\begin{aligned}\overline{\overline{J_{ij}}} &= -\frac{1}{N}A_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_j} + \frac{1}{N}\mu_{\lambda_i\lambda_j}\overline{\overline{\langle\sigma_i\rangle\langle\sigma_j\rangle}} \\ &= -\frac{1}{N}A_{\lambda_i\lambda_j}\mu_{\lambda_i\lambda_j} \pm \frac{1}{N}\mu_{\lambda_i\lambda_j}\sqrt{q_{\lambda_i}q_{\lambda_j}}.\end{aligned} \tag{2.57}$$

The fraction of plusses and minusses is dictated by the cluster magnetizations:

$$p(\pm) = \frac{1}{2}\left[1 \pm m_{\lambda_i}m_{\lambda_j}\right]. \tag{2.58}$$

The funny thing about this section is that we could have guessed the outcome of these calculations very easily from the original Langevin equation (1.45). Instead of taking all these fancy derivatives, we can perform a time average over the Langevin equation. Introduce the time average by:

$$\widehat{A(t)} = \lim_{T\to\infty}\frac{1}{T}\int_S^T A(t)dt \tag{2.59}$$

Performing this average we find

$$\widehat{J_{ij}} = \frac{1}{N}\left[\mu_{ij}A_{ij} + \widehat{\langle\sigma_i\sigma_j\rangle}\right] \tag{2.60}$$

We have thus confirmed that the long time average yields the same result as the ensemble average.

2.6 Replica Symmetry Breaking

The way in which we derived the self-consistency equations (2.16) for m_λ^α and $q^{\alpha\beta}$ does not give an easy way to find out which solutions are stable. This is caused by the introduction of the integral expressions for the delta functions as this led us to evaluation of the free energy precursor in the complex plane. Under the conditions that the $\Lambda\times\Lambda$-matrices $\mu_{\kappa\lambda}$ and $-\mu_{\kappa\lambda}A_{\kappa\lambda}$ are positive definite, we can avoid the introduction of the delta function and subsequent use of complex analysis. This route is followed in appendix C. The way the partition function is expressed in this appendix in (C.4) as $\tilde{Z} \propto \int dqdm \exp -N\phi(m,q)$ provides a way to answer for general n the question whether or not the replica symmetric solutions are *local* minima of the free energy, or equivalently, whether the replica symmetric solutions are unstable or not. If the Hessian of $\phi(m,q)$ is positive definite at a certain solution, then it is indeed a local minimum and for finite N the solution will certainly contribute to the saddle-point approximation (see appendix B for details). Using the symmetry of the replica symmetric Hessian, De Almeida and Thouless [1] were able to calculate all its eigenvalues and determine in the limit of n to zero the conditions for the replica symmetric solution to be a local minimum of ϕ. They numerically calculated a line (now called the AT-line) in the applied field against spin temperature graph above which the replica symmetric solution is a (local) minimum of the free energy. Below that line the system will certainly not feature replica symmetry and another ansatz for the $q^{\alpha\beta}$ and m^α is needed.

Here we are interested in the stability of the replica symmetric solution for general n. In the calculations of [1] only at the very end the limit of n to zero is taken and they are therefore easily adapted to our model. As the precise location of the AT-line is very much dependent on the precise constitution of the network, the calculation is postponed until the next chapter where we choose configurations.

2.6.1 One step replica symmetry breaking

When we cannot make the Replica Symmetry Ansatz, we are forced to continue with the replica trick. We have to carry on pretending n is a positive integer, but not only that, we now also pretend

$$\begin{pmatrix} 0 & q^1 & q^0 & q^0 & q^0 & q^0 \\ q^1 & 0 & q^0 & q^0 & q^0 & q^0 \\ q^0 & q^0 & 0 & q^1 & q^0 & q^0 \\ q^0 & q^0 & q^1 & 0 & q^0 & q^0 \\ q^0 & q^0 & q^0 & q^0 & 0 & q^1 \\ q^0 & q^0 & q^0 & q^0 & q^1 & 0 \end{pmatrix}$$

Figure 2.1: Example of one step broken replica symmetry. n=6, l=3.

that n is a very large number. We have seen in section 2.3 that the assumptions we make about the configuration space of the original system are projected onto the n replicas and vice versa. For example, if the configuration space can be divided into l pure states, then n/l of the n replicas are considered to be in a particular state. That l divides n is something we just take for granted (there is not so much difference in n being a multiple of 1 or l). Without any further assumption we have only replaced the $n(n+1)/2$ replica overlaps and magnetizations by $l(l+1)/2$ pure state overlaps and magnetizations. The step forward is made when we adopt pure state symmetry:

$$\begin{array}{ll} m_a = m & \text{for all pure states } a \\ q_{aa} = q^1 & \text{for all pure states } a \\ q_{ab} = q^0 & \text{for all pure states } a \neq b \end{array} \tag{2.61}$$

The nature of the pure state overlaps requires q^1 to be equal to or larger than q^0. The assumptions can easily be ported to our many cluster network, if we take the division of the replicas into valleys the same for all clusters.

An example of the $q^{\alpha\beta}$-matrix one gets, is shown in figure 2.1. We have succumbed to the customary fashion to have zeros on the diagonal instead of the more logical ones. The first to try such a matrix for $q^{\alpha\beta}$ with $n_1 \equiv n/l = 2$ was Blandin. Parisi [31] generalized this attempt to general l. We pick up the calculation of the free energy at (2.13). As the conjugated replica parameters $\hat{q}$ and $\hat{m}$ are still in there, we adopt the same division for them with the roles of q^0 and q^1 played by $\hat{q}^0$ and $\hat{q}^1$. For a system with only one cluster we can see by relation (2.14), that connects the capped with the uncapped variables, that we have no other option. For a multicluster system one could choose a different division, but using the same division is the natural (and easiest) choice.

Writing, like before, the free energy density in the form:

$$\tilde{f} = -\lim_{N\to\infty} \frac{1}{\tilde{\beta}N} \log \int \dots \exp\left[-N\tilde{\beta}\phi(q^0, \hat{q}^0, q^1, \hat{q}^1, m, \hat{m})\right], \tag{2.62}$$

we can express the exponent ϕ as:

$$\begin{aligned} \phi(q^0, \hat{q}^0, q^1, \hat{q}^1, m, \hat{m}) \;=\; & -\frac{1}{4\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa,\lambda}A^2_{\kappa,\lambda} - \frac{\beta^2 n}{4\tilde{\beta}^2\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}\left[1 + (n_1-1)q^1_\kappa q^1_\lambda + (n-n_1)q^0_\kappa q^0_\lambda\right] \\ & -\frac{in}{2\tilde{\beta}\Lambda}\sum_{\lambda}\left[(n_1-1)\hat{q}^1_\lambda q^1_\lambda + (n-n_1)\hat{q}^0_\lambda q^0_\lambda\right] + \frac{\beta n}{2\tilde{\beta}\Lambda^2}\sum_{\kappa,\lambda}\mu_{\kappa\lambda}A_{\kappa\lambda}m_\kappa m_\lambda \\ & -\frac{\beta n}{\tilde{\beta}\Lambda}\sum_{\kappa}\theta_\kappa m_\kappa - \frac{in}{\tilde{\beta}\Lambda}\sum_{\lambda}\hat{m}_\lambda m_\lambda - \sum_{\lambda}\frac{1}{\tilde{\beta}\Lambda}\log Z_\lambda, \end{aligned} \tag{2.63}$$

$$Z_\lambda = \sum_{\vec{\sigma}_1} \exp -\frac{i}{2}\left[\hat{q}^0_\lambda\left(\sum_\alpha \sigma^\alpha_1\right)^2 + (\hat{q}^1_\lambda - \hat{q}^0_\lambda)\sum_{k=0}^{l-1}\left(\sum_{\alpha=kn_1}^{(k+1)n_1}\sigma^\alpha_1\right)^2 - n\hat{q}^1_\lambda + 2\hat{m}_\lambda\left(\sum_\alpha \sigma^\alpha_1\right)\right].$$

Using, yet again, a Gaussian integral transform, we linearize the spin dependencies and work out the trace:

$$\begin{aligned} Z_\lambda &= \sum_{\vec{\sigma}_1} \int \mathcal{D}z^0 \int \left[\prod_{k=1}^{n/n_1} \mathcal{D}z_k^1\right] \exp\left[\frac{i}{2} n \hat{q}_\lambda^1\right] \\ &\quad \times \exp\left[\left(-i\hat{q}_\lambda^0\right)^{\frac{1}{2}} \left(\sum_\alpha \sigma_1^\alpha\right) z^0 + \sum_{k=1}^{n/n_1} \left(-i\left(\hat{q}_\lambda^1 - \hat{q}_\lambda^0\right)\right)^{\frac{1}{2}} \left(\sum_{\alpha=kn_1}^{(k+1)n_1} \sigma_1^\alpha\right) z_k^1 - i\hat{m}_\lambda \left(\sum_\alpha \sigma_1^\alpha\right)\right] \\ &= 2^n \exp\left[\frac{i}{2} n \hat{q}_\lambda^1\right] \int \mathcal{D}z^0 \left\{\int \mathcal{D}z^1 \cosh^{n_1}\left[\left(-i\hat{q}_\lambda^0\right)^{\frac{1}{2}} z^0 + \left(-i\left(\hat{q}_\lambda^1 - \hat{q}_\lambda^0\right)\right)^{\frac{1}{2}} z^1 - i\hat{m}_\lambda\right]\right\}^{n/n_1}. \end{aligned}$$

Application of the saddle-point method does not yield any surprises for the capped variables. Considering ϕ to be a critical point with respect to the uncapped q's and m shows:

$$\begin{aligned} \hat{q}_\lambda^0 &= \frac{i\beta^2}{\Lambda\tilde{\beta}} \sum_\kappa \mu_{\lambda\kappa} q_\kappa^0, \\ \hat{q}_\lambda^1 &= \frac{i\beta^2}{\Lambda\tilde{\beta}} \sum_\kappa \mu_{\lambda\kappa} q_\kappa^1, \\ \hat{m}_\lambda &= -\frac{i\beta}{\Lambda} \sum_\kappa A_{\lambda\kappa}\mu_{\lambda\kappa} m_\kappa + i\beta\theta_\lambda. \end{aligned}$$

We substitute the above into the equations coming from the other derivatives. The result of the one step replica symmetry breaking scheme are three sets of coupled non linear equations.

$$\begin{aligned} q_\lambda^0 &= C \int \mathcal{D}z^0 \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi\right\}^{n/n_1-2} \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi \tanh \Xi\right\}^2, \\ q_\lambda^1 &= C \int \mathcal{D}z^0 \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi\right\}^{n/n_1-1} \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi \tanh^2 \Xi\right\}, \\ m_\lambda &= C \int \mathcal{D}z^0 \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi\right\}^{n/n_1-1} \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi \tanh \Xi\right\}. \end{aligned} \tag{2.64}$$

The following shorthands were introduced to improve the readability

$$\begin{aligned} \Xi &= \left(\frac{\beta^2}{\tilde{\beta}} Q_\lambda^0\right)^{\frac{1}{2}} z^0 + \left(\frac{\beta^2}{\tilde{\beta}}\left(Q_\lambda^1 - Q_\lambda^0\right)\right)^{\frac{1}{2}} z^1 - \beta M_\lambda - \beta\theta_\lambda, \\ C^{-1} &= \int \mathcal{D}z^0 \left\{\int \mathcal{D}z^1 \cosh^{n_1} \Xi\right\}^{n/n_1}. \end{aligned}$$

The capital Q^i are the obvious generalizations of the Q earlier defined in (2.41). Solutions of these equations should be found by numerical methods. One consistency check is that indeed when either $n_1 = n$, $n_1 = 1$ or $q^1 = q^0$, the above equations reduce to the replica symmetric saddle point equations (2.42).

For the SK-spin glass in the region of AT-instability, the results obtained with this one step replica symmetry breaking are in much better agreement with the Monte Carlo calculations then the replica symmetric approach. Among the improvements is the reduction of the zero temperature entropy at zero magnetic field from $S_{RS}(0) \approx -0.16$ to $S_{1RSB}(0) \approx -.01$ [31].

2.6.2 The full Parisi scheme

The fact that the one step replica symmetry breaking still leaves us with a negative value for the zero temperature entropy indicates that we still do not have a completely accurate mean field

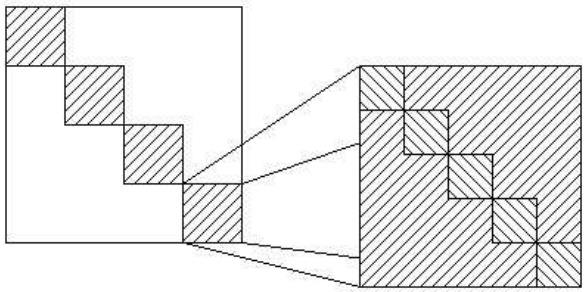


Figure 2.2: An example of two replica symmetry breaking steps. The replicas of the left matrix are divided into four groups. Any two replicas from separate groups share the same mutual overlap. The replicas of one group in the left matrix are again divided into five new groups. Any two replicas from separate groups again share the same mutual overlap. The division and overlaps are equal for all groups.

theory for the SK spin glass as the entropy should be proportional to the logarithm of the number of possible states. Parisi [33, 32] proposed to iterate the procedure above. At the time, the replica overlap matrix was not yet explicitly thought of as representing the structure of the phase space. Parisi originally made assumptions about the $n \times n$ overlap matrix $q^{\alpha\beta}$ and not about the overlaps of valleys. He divided the l groups of n_1 replicas again into smaller groups of n_2 replicas and these again into groups of n_3 replicas and again and again. If the procedure is to be iterated K times, the numbers $n_1 \geq n_2 \geq \ldots \geq n_K$ have to be chosen, $n_{K+1} = 1$. All replicas have the same magnetization, independent of the groups they belong to:

$$m_\alpha = m \qquad \text{for all replicas } \alpha \tag{2.65}$$

This is not true for the overlap. If replicas α and β both belong to the same group of n_i replicas, but are not in the same subdivision of n_{i+1} replicas, they have a mutual overlap of q_i. This is formalized in:

$$q_{\alpha\beta} = q_i \qquad \text{if } \lceil \frac{\alpha}{n_i} \rceil = \lceil \frac{\beta}{n_i} \rceil \text{ and } \lceil \frac{\alpha}{n_{i+1}} \rceil \neq \lceil \frac{\beta}{n_{i+1}} \rceil, \tag{2.66}$$

where $\lceil x \rceil$ denotes the *ceiling* of x, that is the smallest integer larger than or equal to x. This iterative scheme is well illustrated by figures 2.2 and 2.3 taken from [27]. The overlaps of the different replicas are assumed to reflect the hierarchical structure of the division of the replicas in groups,

$$q_0 \leq q_1 \leq \ldots \leq q_K. \tag{2.67}$$

This choice for the order parameter $q^{\alpha\beta}$ is not an immediately obvious one, but is a result of a trial and error period of looking for calculable extensions of the previous scheme. The structure assumed for $q^{\alpha\beta}$ will be mirrored in the structure of the pure state overlaps. This particular choice implies a phase space structure named *ultrametricity*. More is said about the physical implications in the next section.

Ultrametricity

The particular choice made by Parisi for the overlap matrix has an interesting consequence for the structure of the phase space. Consider three replicas α,β and γ. There will always be a step number k, for which $\lceil \frac{\alpha}{n_k} \rceil = \lceil \frac{\beta}{n_k} \rceil$ and $\lceil \frac{\alpha}{n_{k+1}} \rceil \neq \lceil \frac{\beta}{n_{k+1}} \rceil$. The overlap in the Parisi ansatz will be $q^{\alpha\beta} = q_k$ by definition. Define the level l for the replicas α and γ similar to the way we defined the level k for the replicas α and β. It is possible that the replica γ is in the same group as α up to exactly the same subdivision as β, i.e. $l = k$. In that case all three overlaps are equal to q_k. If l is smaller than k, then $q_l < q_k$ and the overlaps given by the Parisi ansatz are: $q^{\alpha\gamma} = q_l$

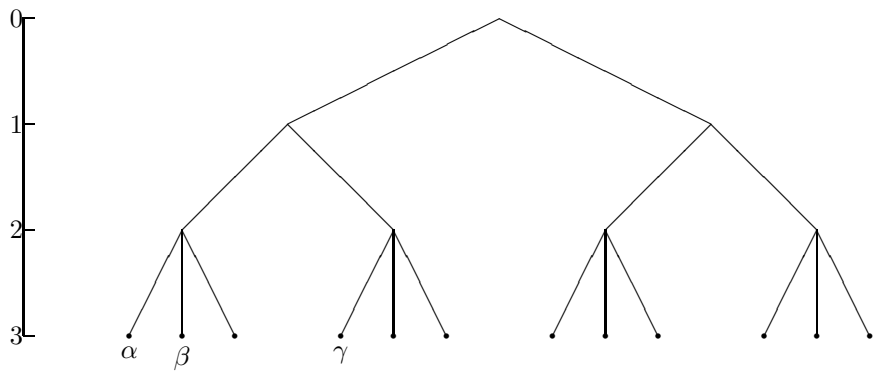


Figure 2.3: The Parisi matrix assumes an hierarchical structure for the replicas. In this picture the replica symmetry is broken twice. Replicas α and β are different, but they are closer relatives to each other than they are to replica γ. This is reflected in the overlaps, $q^{\alpha\beta} > q^{\alpha\gamma} = q^{\beta\gamma}$.

and $q^{\beta\gamma} = q_l$. The last possibility, $k < l$, gives the overlaps: $q^{\alpha\gamma} = q_l$ and $q^{\beta\gamma} = q_k$, where now $q_l > q_k$. So given three replicas, either all overlaps are identical, or two of them are identical and smaller than the third. The way the Parisi replica matrix yields a structure similar to a family tree is visualized in figure 2.3. A natural distance between the two replicas α and β is given by $1 - q^{\alpha\beta}$. If we consider this distance for the space of Parisi replica states, the metric space is called *ultrametric*.

Now consider three real copies of a system with a fixed set of weights J. The overlap of the systems, labeled 1,2,3, have a probability distribution given by

$$P_J(q_{12}, q_{13}, q_{23}) = \sum_{a,b,c} w_a w_b w_c \delta(q_{12} - q_{ab}) \delta(q_{13} - q_{ac}) \delta(q_{23} - q_{bc}), \tag{2.68}$$

where $\sum_a$ is a sum over the free energy valleys and q_{ab} is defined by

$$q_{ab} = \frac{1}{V} \sum_i \langle \sigma_i \rangle_a \langle \sigma_i \rangle_b. \tag{2.69}$$

By a calculation similar to the one performed in section 2.3.1, we can link the weight averaged version of this probability to the Parisi overlap matrix.

$$\begin{aligned} P(q_{12}, q_{13}, q_{23}) &= \overline{P_J(q_{12}, q_{13}, q_{23})} \\ &= \frac{1}{n(n-1)(n-2)} \sum_{\substack{\alpha,\beta,\gamma \\ \text{all different}}} \delta\left(q_{12} - q^{\alpha\beta}\right) \delta\left(q_{13} - q^{\alpha\gamma}\right) \delta\left(q_{23} - q^{\beta\gamma}\right). \end{aligned}$$

The ultrametric structure of the Parisi ansatz is directly reflected in the structure of the phase space. If the Parisi ansatz is correct, we should be able to find the ultrametricity in measurements of the system. Unfortunately, it is not easy to extract the ultrametricity property from (numerical) experiments. Up to now we can only say that the results from numerical experiments still allow ultrametricity.

The only reason why the Parisi ansatz is being used, is that it gives numerical results of for example the entropy and the ground state energy which are in very good agreement with the experiments. Of course, this is one of the best reasons one can have for using a formula in physics, but we would like to know if the peculiar structure of ultrametricity is indeed a property of the system or if there are other replica overlap matrices possible which give equally good results. People, in particular Parisi himself, are still trying to analyticly prove the existence of ultrametricity in the system, see for example [34], but so far nobody has succeeded.

We proceed now with the computational analysis and continue the replica symmetry breaking iteration beyond all bounds.

2.6.3 Near the critical temperature

The Parisi scheme can be used step by step to improve the numerically found results in any region of the phase diagram. The full scheme, $K = \infty$, can be applied to the expansion of the free energy precursor ϕ in the order parameters. We can expand the precursor of the free energy $\phi(q, m)$ given in C.4 in appendix C. The drawback is that we have to comply to condition 1, thus limiting the networks we can study. For an unbiased system we can expand the replicated single neuron trace that appears in the free energy expression (C.4) in orders of q. Using

$$\begin{aligned} \exp x &= \sum_{n=0}^{N} \frac{x^n}{n!} + \mathcal{O}(x^{N+1}) \\ \log 1 + x &= \sum_{n=1}^{N} (-1)^{n-1} \frac{x^n}{n} + \mathcal{O}(x^{N+1}) \end{aligned} \qquad x \downarrow 0,$$

we find

$$\begin{aligned} &\log \sum_{\vec{\sigma}_1} \exp \frac{\eta}{2} \sum_{\alpha,\beta} Q_\lambda^{\alpha\beta} \sigma_1^\alpha \sigma_1^\beta \\ &= \log \left[2^n \left(1 + \frac{\eta^2}{4} \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^2 + \frac{\eta^3}{6} \sum_{\alpha,\beta,\gamma} Q_\lambda^{\alpha\beta} Q_\lambda^{\beta\gamma} Q_\lambda^{\gamma\alpha} + y' \eta^4 \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^4 + \mathcal{O}(q^4) \right) \right] \\ &= n \log 2 + \frac{\eta^2}{4} \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^2 + \frac{\eta^3}{6} \sum_{\alpha,\beta,\gamma} Q_\lambda^{\alpha\beta} Q_\lambda^{\beta\gamma} Q_\lambda^{\gamma\alpha} + y \eta^4 \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^4 + \mathcal{O}(q^4), \end{aligned}$$

where we temporarily use the variable $\eta = \beta^2/\tilde{\beta}$. The sum over states in the original expression only keeps terms where the replica indices appear in pairs. Of the terms of order q^4 only the one term is kept which is responsible for replica symmetry breaking following [32]. This makes the number $y = 1/12$ rather arbitrary, but more detailed investigations including all fourth order terms have shown that this only fraction is relevant. Forgetting the irrelevant (if we do not want to compute the energy) $n \log 2$, this makes $\phi(q)$:

$$\phi(q) = \frac{1}{\Lambda} \sum_\lambda \left[\frac{\eta}{4} \sum_{\alpha,\beta} q_\lambda^{\alpha\beta} Q_\lambda^{\alpha\beta} - \frac{\eta^2}{4} \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^2 - \frac{\eta^3}{6} \sum_{\alpha,\beta,\gamma} Q_\lambda^{\alpha\beta} Q_\lambda^{\beta\gamma} Q_\lambda^{\gamma\alpha} - y \eta^4 \sum_{\alpha,\beta} \left(Q_\lambda^{\alpha\beta} \right)^4 \right] + \mathcal{O}(q^4)$$

If the spin temperature is very high, the only minimum is at $q_\lambda^{\alpha\beta} = 0$ for all clusters and all replicas. When the temperature is lowered, a continuous transition to a state corresponding to less trivial critical point is possible at a certain critical temperature. To compute the inverse critical temperature β_c, we bring the $\Lambda \times \Lambda$-matrix μ in diagonal form. Because $\mu_{\kappa\lambda}$ is symmetric it is always possible to find an orthogonal $\Lambda \times \Lambda$-matrix Ω, i.e. $\Omega^T \Omega = \Omega \Omega^T = \mathbb{I}$, such that

$$D \equiv \frac{1}{\Lambda} \Omega^T \mu \Omega$$

is a diagonal matrix with on its diagonal the Λ eigenvalues d_i of μ ordered as $d_1 \geq d_2 \geq \ldots \geq d_\Lambda$. Further we define the transforms of the Λ-vectors $q^{\alpha\beta}$ and $Q^{\alpha\beta}$:

$$\begin{aligned} p^{\alpha\beta} &= \Omega^T q^{\alpha\beta}, \\ P^{\alpha\beta} &= \Omega^T Q^{\alpha\beta} = \frac{1}{\Lambda} \Omega^T \mu \Omega \Omega^T q^{\alpha\beta} = Dp. \end{aligned}$$

We define a new function in terms of this new vector, $\phi(\tilde{\Omega^T} q) = \phi(q)$. The lowest order terms of this new 'free energy' are:

$$\tilde{\phi}(p) = \frac{1}{\Lambda} \sum_\lambda \frac{\eta d_\lambda}{4} (1 - \eta d_\lambda) \sum_{\alpha,\beta} \left(p_\lambda^{\alpha\beta} \right)^2 + \mathcal{O}(p^3). \tag{2.70}$$

The extrema of this free energy function satisfy

$$p_\lambda^{\alpha\beta} = \eta d_\lambda p_\lambda^{\alpha\beta} + \mathcal{O}(p^2). \tag{2.71}$$

Near $p = 0$, this equation only allows non-trivial solutions when $\eta d_\lambda = 1$ for one or more clusters λ. When decreasing the temperature β^{-1}, the first bifurcation (a second order phase transition) one encounters, is located at

$$\beta_c^2 = \tilde{\beta} d_1^{-1}, \tag{2.72}$$

as d_1 is by definition the largest eigenvalue of μ. Around this temperature the expansion in q makes sense (if no first order transition has occurred before that point). To stress that we expand near the critical temperature, we introduce a new variable

$$\tau = \frac{T_c - T}{T_c} = 1 - \frac{\beta_c}{\beta}.$$

Although [9] did not give any justification for this step, using this new variable is only allowed because this critical spin temperature is independent of n. If this had not been the case, we would have risked leaving the original model at this point, because the model we have set up in chapter one is parametrized by β and $\tilde{\beta}$, not by β and n. In general, the first Λ' eigenvalues of μ might be identical. The bifurcation from $q = 0$ can initially develop only along the corresponding Λ' eigenvectors. We make the extra assumption that d_1 is strictly larger than the other eigenvalues, i.e. $\Lambda' = 1$. The consequence is that the bifurcation is only along the first eigenvector p_1. Near $\tau = 0$, $q = \mathcal{O}(\tau)$ which justifies (except for the already discussed last term) the expansion made. We substitute the approximation $p = p_1$ back into the full form of (2.70)

$$\phi(p) = \frac{1}{2\Lambda}\left[-\tau \sum_{\alpha,\beta}\left(p_1^{\alpha\beta}\right)^2 - \frac{1}{3}c_3 \sum_{\alpha,\beta,\gamma} p_1^{\alpha\beta} p_1^{\beta\gamma} p_1^{\gamma\alpha} - 2y c_4 \sum_{\alpha,\beta}\left(p_1^{\alpha\beta}\right)^4\right] + \mathcal{O}(p^4), \tag{2.73}$$

where the constants c_3 and c_4 are remnants of the transformation of q to p. Other than the second order terms, the third and fourth orders terms in 2.70 are not invariant under the orthogonal transformation, resulting in

$$c_3 = \sum_\lambda (\Omega_{\lambda 1})^3, \qquad c_4 = \sum_\lambda (\Omega_{\lambda 1})^4.$$

This last expression for $\tilde{\phi}$ is fit for analytical investigation in the case of infinite replica symmetry breaking.

2.6.4 Infinite replica symmetry breaking

Consider a K-step replica symmetry breaking scheme. If $n \gg 1$, it consists of a sequence $n \equiv n_0 \geq n_1 \geq \cdots \geq n_k \geq n_{k+1} \equiv 1$ and a sequence $0 \leq q_0 \leq q_1 \leq \cdots \leq q_k \leq 1$. A sum over all co-efficients of the matrix $q^{\alpha\beta}$ is

$$\sum_{\alpha,\beta} q^{\alpha\beta} = n \sum_{i=0}^{k} (n_i - n_{i+1})\, q_i. \tag{2.74}$$

In the limit of $K \to \infty$, infinite breaking, we can replace the discrete q_i's by the continuous function $q(x)$,

$$\forall x \in [1, n], \quad n_i < x \leq n_{i+1}: \quad q(x) = q_i. \tag{2.75}$$

The sum over the breaking steps gets replaced by an integral:

$$\sum_{\alpha,\beta} q^{\alpha\beta} = n \int_1^n dx\, q(x). \tag{2.76}$$

In the usual fashion, we forget the initial constraints of n being a large positive integer. Parisi considered directly the limit of $n \to 0$. Kondor [22] was the first to look at the results of the RSB-scheme for finite $n < 1$. Here we also only look at n smaller than one as for larger n, we assume the replica symmetric saddle-point to be stable. How the order parameter function $q(x)$ relates to the experiment can be seen by looking at the distribution of q (2.36):

$$\begin{aligned} P(q) &= \frac{1}{1-n}\int_n^1 dx\,\delta\left(q-q(x)\right) \\ &= \frac{1}{1-n}\int_{q(n)}^{q(1)} d\tilde{q}\,x'(\tilde{q})\delta\left(q-\tilde{q}\right) = \frac{1}{1-n}x'(q), \end{aligned} \tag{2.77}$$

where $x(q)$ is the inverse of $q(x)$.

By following a procedure outlined in [9], all the appearances of the matrix $q_\lambda^{\alpha\beta}$ (or $p_\lambda^{\alpha\beta}$) in (2.73) can be expressed in terms of $q_\lambda(x)$ (or its transform $p_\lambda(x)$). A matrix Q in the linear space V of Parisi matrices completed with the identity matrix $I^{\alpha\beta} = \delta^{\alpha\beta}$ can be completely characterized by its diagonal element $g(Q) = Q^{\alpha\alpha}$ and its corresponding Parisi function $f(Q) = q$. The linear space V is closed with respect to the matrix product. In [9] it is noted that, if A, B are matrices in V, f and g behave under the matrix product as:

$$\begin{aligned} g(AB) &= g(A)g(B) - \langle f(A)f(B)\rangle, \\ f(AB) &= -nf(A)f(B) + \Big(g(A) - \langle f(A)\rangle\Big) f(B) + \Big(g(B) - \langle f(B)\rangle\Big) f(A), \\ &\quad + \int_n^x dy\ \left(f(A)(x) - f(A)(y)\right)\left(f(B)(x) - f(B)(y)\right) \end{aligned}$$

where

$$\langle q\rangle = \int_n^1 dx\,q(x). \tag{2.78}$$

All terms in (2.73) can be expressed in terms of $p_\lambda(x)$:

$$\begin{aligned} \sum_{\alpha,\beta}\left(p_\lambda^{\alpha\beta}\right)^t &= -\int_n^1 dx\,p_\lambda(x)^t, \\ \sum_{\alpha,\beta,\gamma} p_\lambda^{\alpha\beta}p_\lambda^{\beta\gamma}p_\lambda^{\gamma\alpha} &= g(p_\lambda^3) = \int_n^1 dx\,xp_\lambda(x)^3 + 3\int_n^1 dx\,p_\lambda(x)\int_n^x dy\,p(y)_\lambda^2. \end{aligned}$$

And the 'free energy' (2.70) can be written as

$$\tilde{\phi}(p) = \frac{1}{2\Lambda}\sum_{\lambda=1}^{\Lambda'}\int_n^1 dx\left[\tau p_\lambda(x)^2 - \frac{1}{3}c_3xp_\lambda(x)^3 - c_3p_\lambda(x)\int_n^x dy\,p_\lambda(y)^2 + 2yc_4p_\lambda(x)^4\right]. \tag{2.79}$$

Using a infinite-dimensional saddle-point argument (see appendix B), we vary $\phi(p)$ with respect to p_λ and find with $I(1,n) = [\min\{1,n\}, \max\{1,n\}]$ for $x \in I(1,n)$:

$$2\tau p_\lambda(x) - c_3xp_\lambda(x)^2 - 2c_3p_\lambda(x)\int_x^1 dy\,p_\lambda(y) - c_3\int_n^x dy\,p_\lambda(y)^2 + 8yc_4p_\lambda(x)^3 = 0. \tag{2.80}$$

To solve this equation we differentiate once with respect to x and find:

$$\tau - c_3xp_\lambda(x) - c_3\int_x^1 dy\,p_\lambda(y) + 12yc_4p_\lambda(x)^2 = 0 \quad \text{or} \quad p_\lambda'(x) = 0 \tag{2.81}$$

Differentiating the left hand side once more shows that

$$x \in I(1,n): \qquad p_\lambda(x) = \frac{c_3}{24yc_4}x \quad \text{or} \quad p_\lambda'(x) = 0. \tag{2.82}$$

To calculate the order parameter q or rather its distribution $P(q)$, we need more information about the transformation matrix Ω. Further conclusions are therefore postponed to the next chapter, when a specific configuration is chosen.

Chapter 3

Translation Invariant Networks

In the previous chapter we have come a long way in reducing the partition function into a set of non-linear coupled equations. In this chapter we will numerically find some solutions for the order parameters and determine the temperature regions where different type of solutions are possible. We will focus mainly on the one cluster model, reproducing the 1993 results. For comparison with the diagram of Penney *et al.* we present the results with the parameters they chose and with the parameters of the original model. For the one-cluster system it is shown that the weight dynamics remains ergodic in the thermodynamic limit.

At the end of the chapter we will use the cluster structure. For an example system we will show that a spatial structure in the magnetization might arise. Finally, we discuss what happens if we implement Linsker's layered structure in this network model.

3.1 The One Cluster Model

When neither a spatial structure nor an external field are forced on the system, the system is like the one studied by Penney, Coolen and Sherrington (PCS) in [35, 5, 40]. They made the assumption of a zero or positive bias, whereas here implicitly a negative bias is assumed. However, up to now the sign of the bias has been irrelevant as soon as we assume replica symmetry and we expect to find their results. Indeed, if $\mu_{\kappa\lambda} = \mu$ and $A_{\kappa\lambda} = A$, then Q_λ and M_λ turn out to be cluster independent by definition (2.41).

The fixed-point equations (2.42) are simplified to

$$\begin{aligned} q_\lambda &= \left[\int \mathcal{D}\xi\, \cosh^n \Xi \tanh^2 \Xi\right]\left[\int \mathcal{D}\xi\, \cosh^n \Xi\right]^{-1}, \\ m_\lambda &= \left[\int \mathcal{D}\xi\, \cosh^n \Xi \tanh \Xi\right]\left[\int \mathcal{D}\xi\, \cosh^n \Xi\right]^{-1}, \qquad \Xi = \sqrt{\beta/n}\sqrt{Q}\xi + \beta M. \end{aligned} \tag{3.1}$$

This is consistent with the formulae encountered in [35], but gives the additional information that the cluster order parameters are all equal to the order parameters of the entire system.

Another extreme structure for which we can anticipate the outcome is the one which has only interaction within the clusters. This is the case when the weight decays are of the form $\mu_{\kappa\lambda} = \mu\delta_{\kappa\lambda}$. The resulting fixed-point equations are the expected ones: a series of decoupled systems like the one described right above. q_λ and m_λ are the order parameters for the infinite volume system of cluster λ.

The PCS-model also is the one-cluster case of the model under investigation in this thesis. As shown by PCS even this 'simple' case exhibits a phase diagram with interesting features. Before taking on systems with an infinite number of clusters we will analyze and calculate the phase diagram of this one-cluster system.

3.1.1 AT-line

The first task in drawing a phase diagram is to determine in which region we can use the replica symmetry ansatz. Van Hemmen and Palmer [14] concluded that the analytic continuation of the formula found for positive integral n is unique for real n larger than one. For these n we are sure to have a global minimum in the replica symmetric solution. For smaller n we do not have certainty of having a global minimum, but we can find where the replica symmetric solution is at least a local minimum. For this we have to calculate the AT-line as outlined in the previous chapter.

The starting point is the function $\phi(q,m)$ appearing in appendix D as a precursor of the free energy:

$$\begin{aligned}\phi(q,m) &= \frac{1}{2}\eta^2 \sum_{\alpha<\beta} \left(q^{\alpha\beta}\right)^2 + \frac{1}{2}\kappa \sum_{\alpha} \left(m^{\alpha}\right)^2 - \log \sum_{\vec{\sigma_1}} \exp -H(\sigma_1),\\ H(\sigma_1) &= -\eta^2 \sum_{\alpha<\beta} q^{\alpha\beta}\sigma_1^{\alpha}\sigma_1^{\beta} - \kappa \sum_{\alpha} m^{\alpha}\sigma_1^{\alpha}.\end{aligned}$$

The sufficient condition to start with this function is that this system has a positive bias, i.e. $K \equiv -A > 0$. In the model introduced in chapter one and in the PCS-model the constants are

$$n = \tilde{\beta}/\beta, \qquad \eta = \mu\beta^2/\tilde{\beta}, \qquad \kappa = -A\mu\beta. \tag{3.2}$$

We will call this the *Hebbian Learning Neurons-* or HLN-parametrization. In their publications Penney *et al.* took a different set of parameters in order to make the expressions correspond more clearly with the expressions for the SK model. They introduced the new parameters:

$$J_0 = -A\mu, \qquad \tilde{J} = \frac{\mu}{\tilde{\beta}} = \frac{\mu}{\beta n}. \tag{3.3}$$

In the SK model J_0/N represents the mean-value coupling between two spins and $\tilde{J}/n$ the variance of the coupling. Note that this interpretation is linked to the $n \to 0$ limit taken in this model. These parameters also bring the expressions in harmony with the ones Sherrington used in [38]. In this article Sherrington analyzes the behavior of the SK equations with integral $n \geq 2$ without taking the $n \to 0$ limit. The η and κ consistent with their choice of constants are:

$$\tilde{\beta} = n\beta, \qquad \eta = \beta^2\tilde{J}, \qquad \kappa = \beta J_0. \tag{3.4}$$

By treating $\tilde{J}$ as a constant the model is actually altered, as now the decay term becomes dependent on the temperature of the weights. In this section we will calculate the phase diagram once for the SK constants and once for the model originally under investigation.

We now closely follow [1]. The general Hessian of ϕ, denoted by the matrix G, has seven different types of matrix elements. In a replica symmetric point of the (q,m) space these coefficients can be written concisely when we introduce next to the two familiar order parameters q and m, two new variables t and r:

$$\begin{aligned} m &= t_1(\eta q, \kappa m),\\ q &= t_2(\eta q, \kappa m),\\ t &= t_3(\eta q, \kappa m),\\ r &= t_4(\eta q, \kappa m),\end{aligned} \tag{3.5}$$

where t_2 and t_1 are abbreviations for the functions encountered earlier:

$$t_k(x,y) = \left[\int \mathcal{D}\xi\, \cosh^n \Xi \tanh^k \Xi\right]\left[\int \mathcal{D}\xi\, \cosh^n \Xi\right]^{-1}, \qquad \Xi = \xi\sqrt{x} + y.$$

Define the shorthands

$$G_{\alpha,\beta} = \frac{\partial^2 G}{\partial m^{\alpha}\partial m^{\beta}}, \quad G_{\alpha,\beta\gamma} = \frac{\partial^2 G}{\partial m^{\alpha}\partial q^{\beta\gamma}}, \quad G_{\alpha\beta,\gamma\delta} = \frac{\partial^2 G}{\partial q^{\alpha\beta}\partial q^{\gamma\delta}}. \tag{3.6}$$

The two types of matrix elements for the magnetization derivatives are:

$$G_{\alpha,\alpha} = \kappa\left[1-\kappa(1-m^2)\right] \equiv A, \tag{3.7}$$
$$G_{\alpha,\beta} = -\kappa^2(q-m^2) \equiv B. \tag{3.8}$$

Three types for the Edwards-Anderson parameter derivatives

$$G_{\alpha\beta,\alpha\beta} = \eta^2\left[1-\eta^2(1-q^2)\right] \equiv P, \tag{3.9}$$
$$G_{\alpha\beta,\alpha\gamma} = -\eta^4(q-q^2) \equiv Q, \tag{3.10}$$
$$G_{\alpha\beta,\beta\gamma} = -\eta^4(r-q^2) \equiv R. \tag{3.11}$$

And finally two cross term types

$$G_{\alpha,\alpha\beta} = -\eta^2\kappa(m-mq) \equiv C, \tag{3.12}$$
$$G_{\alpha,\beta\gamma} = -\eta^2\kappa(t-mq) \equiv D. \tag{3.13}$$

Define a $n(n+1)/2$ column vector $\vec{\mu}$ from the replica order parameters as

$$\vec{\mu} = \begin{pmatrix} \{m^\alpha\} \\ \{q^\alpha\beta\} \end{pmatrix}. \tag{3.14}$$

The matrix G is a real symmetric matrix of order $n(n+1)/2$ and thus has the same number of eigenvectors. In order to solve the eigenvalue equation $G\mu = \lambda\mu$, De Almeida and Thouless made an educated guess about the form of the eigenvectors, based of course on the replica symmetry of the matrix. Their analysis is repeated here below. The calculations are not difficult, but due to the large number of different coefficient types the expressions one encounters during the algebra are quite long. Therefore all intermediate steps are left out.

First consider the replica symmetric vector $\vec{\mu_0}$ with the elements:

$$\forall\alpha \quad m^\alpha = a, \qquad \forall\alpha\neq\beta \quad q^{\alpha\beta} = b. \tag{3.15}$$

Substitution of this vector in the eigenvalue equation yields two eigenvalues:

$$\begin{aligned}\lambda_{0\pm} = {} & \frac{1}{2}\left[A+(n-1)B+P+2(n-2)Q+\frac{1}{2}(n-2)(n-3)R\right] \\ & \pm\frac{1}{2}\Bigg[\left\{A+(n-1)B-P-2(n-2)Q-\frac{1}{2}(n-2)(n-3)R\right\}^2 \\ & \qquad +2(n-2)\left\{2C+(n-2)D\right\}^2\Bigg]^{1/2}.\end{aligned}$$

Next we consider a vector in which a replica θ is excepted from the replica symmetry

$$\begin{array}{llll} \alpha=\theta & m^\alpha = a, & \alpha=\theta\vee\beta=\theta & q^{\alpha\beta}=c, \\ \alpha\neq\theta & m^\alpha = b, & \alpha,\beta\neq\theta & q^{\alpha\beta}=d. \end{array} \tag{3.16}$$

For $n>2$, eigenvectors of this form correspond with two eigenvalues:

$$\begin{aligned}\lambda_{1\pm} = {} & \frac{1}{2}\left[A-B+P+(n-4)Q-(n-3)R\right] \\ & \pm\frac{1}{2}\left[\left\{A-B-P-(n-4)Q+(n-3)R\right\}^2+4(n-2)\left\{C-D\right\}^2\right]^{1/2}.\end{aligned}$$

And now substitute vectors where two replicas θ and ν are together singled out from the rest:

$$\begin{array}{llll} \alpha\in\{\theta,\nu\} & m^\alpha = a, & & q^{\theta\nu}=c, \\ \alpha\notin\{\theta,\nu\} & m^\alpha = b, & \alpha\notin\{\theta,\nu\} & q^{\alpha\nu}=q^{\alpha\theta}=d, \\ & & \alpha,\beta\notin\{\theta,\nu\} & q^{\alpha\beta}=e. \end{array} \tag{3.17}$$

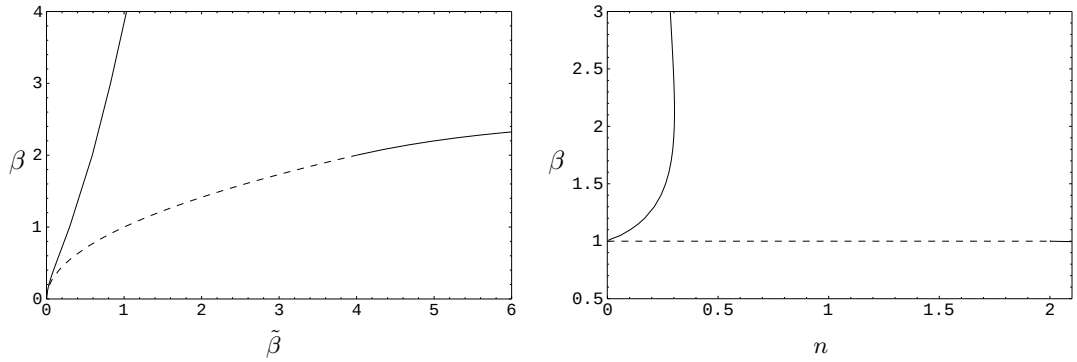


Figure 3.1: The AT-instability is indicated by the full line shown on the left in the HLN-parametrization (3.2) with $\mu = 1$ and on the right in the SK-parametrization (3.4) with $J_0 = 1$. No bias is present. The dashed line is the second order paramagnet to Mattis/spin glass transition.

hich, for $n > 2$, yields just one eigenvalue:

$$\lambda_2 \quad = \quad P - 2Q + R. \tag{3.18}$$

Counting the multiplicities of these five eigenvalues (respectively 1,1,(n-1),(n-1),n(n-3)/2) it can be concluded that all the eigenvalues have been now found.

De Almeida and Thouless studied all three sets of eigenvalues and they concluded that only the third eigenvalue can be negative. The equation $\lambda_2 = 0$ thus defines in the $\eta - n$-phase diagram a line which separates the region where the replica symmetric solution is at least a local minimum of the free energy functional from the region where it is not even a local minimum. The AT-instability is plotted in figure 3.1. For parameter values left of the AT-line, the replica symmetric solution is an unstable solution of the saddle-point equations and replica symmetry thus will be broken.

3.1.2 Stable replica symmetry

To characterize the system in a replica symmetric state, we have the two order parameters q and m, with the restriction $q > m^2$. This leads to the identification of the following three phases: *paramagnet* in case $q = 0, m = 0$, *Mattis glass* in case $q \neq 0, m = 0$, and *ferromagnet* in case $q \neq 0, m \neq 0$.

Mattis glass to paramagnetic phase transition

Without a bias A (or equivalently J_0), κ is zero. The function $t_1(x, y)$ is odd in y and in the absence of a bias, m will be zero by (3.5) (see also section (3.1.4)). In this case we only have to deal with the first of equations (3.5) and transition from the Mattis glass to paramagnetic state. When $\beta = 0$ while n is finite, η is zero in both parameter regimes and thus (3.5) only has $q = 0$ as a solution. If we expand $t2(\eta q)$ around $\eta q = 0$,

$$t_2(\eta q) = \eta q + (n-2)(\eta q)^2 + \mathcal{O}((\eta q)^3), \tag{3.19}$$

it is apparent from the second term that for $n \leq 2$ there will be a second order transition (a bifurcation) of the trivial solution $q = 0$ to a non-trivial but initially very small $q \neq 0$. The transition takes place when $\eta = 1$. If $n > 2$ there will be a first order transition. The precise location of this transition has to be calculated numerically for all $n > 2$. The results of these calculations are given in figure (3.2) for the SK constants and for our original model. The phase diagram with $\tilde{J} = 1$ looks at first sight exactly the same as the one published by PCS, as it should. A closer examination shows that the precise location of the first order transition is slightly different. I have no explanation for this discrepancy.

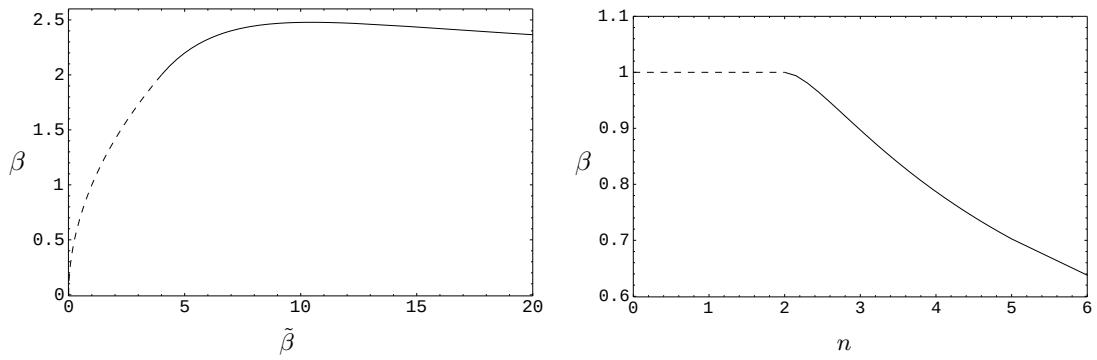


Figure 3.2: Phase transition between paramagnetic and Mattis glass state. On the left, the HLN parametrization with $\mu = 1$, on the right the SK parametrization with $\tilde{J} = 1$. No bias is present. Full lines are first order transitions, dashed lines are of second order.

Ferromagnetic phase transitions

In the presence of a positive bias, that is $A < 0$ or $J_0 > 0$, it is less trivial to find the phase diagram as now the fixed points are sought in a two dimensional space, because the magnetization comes into the play. Again substituting $\beta = 0$ in (3.1) while considering n to be finite, shows that in the very high temperature regime only the trivial paramagnetic solution $q = m = 0$ is possible. In a not extremely sophisticated, but tedious expansion of t_2 and t_1 Sherrington [38] showed that the paramagnetic to ferromagnetic phase transition is of second order if $\eta^{-1} > 3n - 2$ and of first order if the contrary is true. The second order transition will take place when $\kappa = 1$. Diagrams with a ferromagnetic phase are shown in figures 3.3 and 3.4. Possibly because of numerical difficulties PCS did not include a phase diagram with a ferromagnetic region in their articles.

3.1.3 Broken replica symmetry

Near paramagnet-spin glass transition

Left of the AT-line ($n < 1$) and near the boundary between the paramagnetic state and the spin glass state, we can apply the full Parisi scheme as derived in section 2.6.4. For just one cluster $p(x) = q(x)$, and the only non-trivial possible continuous solution of (2.82) is:

$$q(x) = \begin{cases} q_n & n \leq x < x_n \\ \frac{1}{24y} & x_n \leq x < x_1 \\ q_1 & x_1 \leq x \leq 1 \end{cases} \tag{3.20}$$

with

$$x_n = 24yq_n, \qquad x_1 = 24yq_1. \tag{3.21}$$

Furthermore, from equations (2.80) and (2.81) follows (for small τ)

$$q_n = \frac{n}{16y}, \qquad q_1 = \tau. \tag{3.22}$$

The existence of this solution is subject to the condition $\tau > n/16y$. For $\tau < n/16y$, the only non-negative solution is $q(x) = 0$. An example of a solution $q(x)$ and the corresponding $P(q)$ are shown in figure 3.5.

3.1.4 Non-existence of ferromagnetic state and ergodicity

In [35] Penney *et al.* it is mentioned briefly that a ferromagnetic state is only expected when a finite, in their case meaning positive, uniform bias is present. The reason for this is of course the

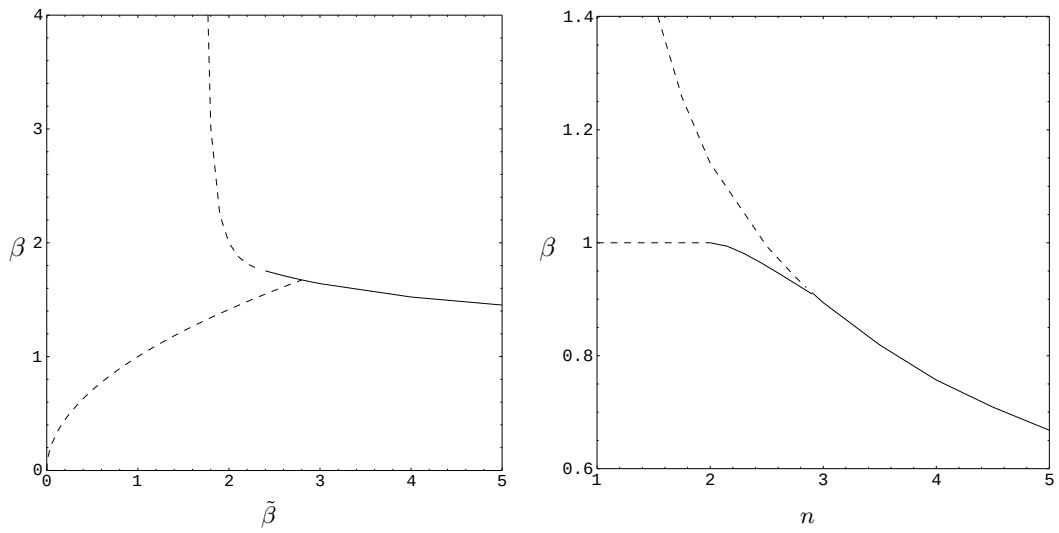


Figure 3.3: Phase diagrams of systems with a small ferromagnetic bias. On the left the HLN parametrization with $\mu = 1$, $A = -0.5$; on the right the SK parametrization with $\tilde{J} = 1$, $J_0 = 0.5$. Full lines stand for first order transitions, dashed lines for second order transitions. The phases are in the upper right the ferromagnetic, left the Mattis glass and down below the paramagnetic phase.

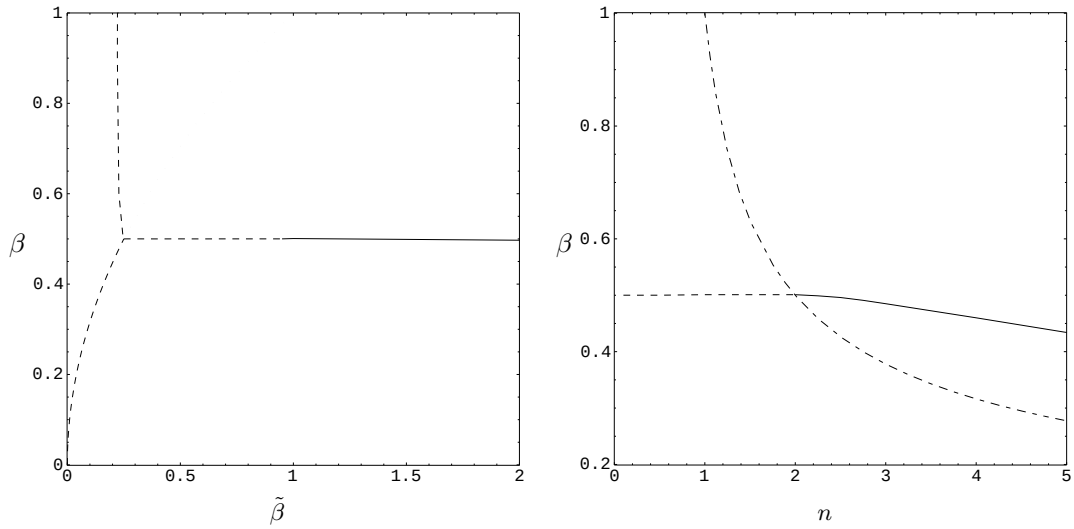


Figure 3.4: Phase diagrams of systems with a large ferromagnetic bias. On the left the HLN parametrization with $\mu = 1$, $A = -2$; on the right the SK parametrization with $\tilde{J} = 1$, $J_0 = 2$. Full lines stand for first order transitions, dashed lines for second order transitions. The HLN-parametrization show clockwise from the upper right: ferromagnetic, paramagnetic and Mattis glass order. The SK parametrization does not feature a Mattis glass phase for this bias. The positions of the transition from first to second order for different values of the biases are given by the dashed-dotted line.

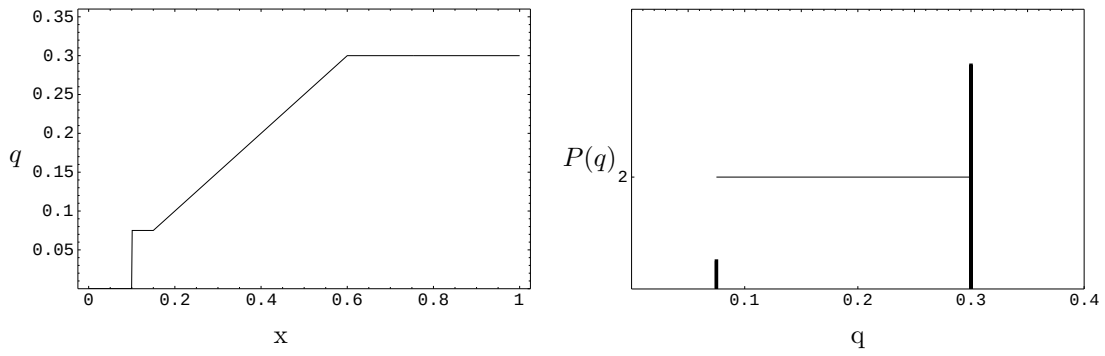


Figure 3.5: A solution for $n = 0.1$ and $\tau = 0.3$ of the order parameter function $q(x)$ is shown on the left and $P(q)$, the corresponding distribution for the order parameter q, is shown on the right. The thick lines symbolize delta functions with integrals $2q_n - n$ and $1 - 2q_1$.

spin flip symmetry. If the system is unbiased there is nothing to favor a ferromagnet more than any other Mattis magnet configuration. The question I have asked myself, is whether it is possible that, when starting in a ferromagnetic system, the system can remain ferromagnetic even in the presence of a negative bias. Imagine a system in which the initial weights are all positive and which has a uniform negative bias between zero and one ($0 < A < 1$). If the neuron temperature is sufficiently low, the neurons will align. The neuron correlation which drives the weight change will be close to one. Disregarding the noise in the weights, which is equivalent to a zero weight-temperature limit, the weight dynamics are

$$\tau \frac{d}{dt} J_{ij} = \frac{1}{N}(1 - A) - \frac{1}{\mu} J_{ij} + \frac{1}{\sqrt{N}} \eta_{ij}(t) \quad \text{for} \quad i < j. \tag{3.23}$$

This noiseless equation is easily integrated to

$$J_{ij}(t) = \mu/N(1 - A)(1 - \exp -t/\mu\tau) + J_{ij}(0) \exp -t/\mu\tau. \tag{3.24}$$

Starting with all weights positive, they will remain so because the bias A is smaller than one. In this state the magnetization will be very near plus or minus one. Adding a little noise to the biased Hebbian dynamics shouldn't change much. So I thought initially.

However, for finite non-zero weight and neuron temperatures this ferromagnetic state is not a solution of the saddle-point equations. We can see this by looking at the structure of the right hand side of the magnetization equation. When m is zero, $\cosh^n \Xi(\xi) \tanh \Xi(\xi)$ is an odd function, positive for ξ greater than zero. If m is positive, the function is shifted to the right. Integrated with the Gaussian measure $\exp -\xi^2/2$, this function will yield a negative value. For negative m, a left shift occurs, and the integration results in a positive value. The only solution therefore for a zero or negative bias is a zero magnetization.

We can check if the same behavior persists in the zero temperature limits. If we want to take low temperature limits in the saddle-point equations, we can choose to perform one of them first or performing them both at the same time. The second case is done by fixing the temperature quotient $n \equiv \tilde{\beta}/\beta$, while sending β and $\tilde{\beta}$ to infinity.

$$\begin{aligned}
& \lim_{\substack{\beta\to\infty \\ n \text{ fixed}}} \tanh \Xi = \lim_{\substack{\beta\to\infty \\ n \text{ fixed}}} \tanh \beta \left(\sqrt{\frac{Q}{\beta n}} \xi + M \right) = \operatorname{sgn}(M) \\
\Rightarrow \quad & \lim_{\substack{\beta\to\infty \\ n \text{ fixed}}} m = \lim_{\substack{\beta\to\infty \\ n \text{ fixed}}} \operatorname{sgn}(M) \left[\int \mathcal{D}\xi \, \cosh^n \Xi \right] \left[\int \mathcal{D}\xi \, \cosh^n \Xi \right]^{-1} = -\operatorname{sgn}(Am) \\
\Rightarrow \quad & m = 0.
\end{aligned}$$

This means that there is no saddle-point with a finite magnetization, thus after taking the thermodynamic limit we do not find a stable ferromagnetic state in this way of taking the zero temperature limits.

Another way of taking the zero temperature limit, is taking them one at the time. Following the argument which led me to a stable state when looking at the dynamics, I should set the neuron temperature to zero before liquidating the weight noise.

$$\begin{aligned} & \lim_{\substack{\beta\to\infty \\ \tilde{\beta}\,\text{fixed}}} \tanh\Xi = \lim_{\substack{\beta\to\infty \\ \tilde{\beta}\,\text{fixed}}} \tanh\beta\left(\sqrt{\frac{Q}{\tilde{\beta}}}\xi + M\right) = \operatorname{sgn}\left(\sqrt{\frac{Q}{\tilde{\beta}}}\xi + M\right) \\ \Rightarrow \quad & \lim_{\substack{\beta\to\infty \\ \tilde{\beta}\,\text{fixed}}} m = \lim_{\substack{\beta\to\infty \\ \tilde{\beta}\,\text{fixed}}} \int \mathcal{D}\xi \operatorname{sgn}\left(\sqrt{\frac{Q}{\tilde{\beta}}}\xi + M\right) \\ \Rightarrow \quad & m = 0 \end{aligned}$$

The last line follows from the fact the Gaussian function hidden in $\mathcal{D}\xi$ is an even function in ξ, whereas $\operatorname{sgn}(\xi\sqrt{Q/\tilde{\beta}} + M)$ is an odd function shifted by the presence of a non-zero M. If M is positive, the shift of the sgn-function is to the right, so that $\int \mathcal{D}\xi \operatorname{sqn}(\xi\sqrt{Q/\tilde{\beta}} + M)$ is smaller than zero. For negative M the same argument applies. We have to conclude that the saddle-point equations indeed do not allow a ferro-magnetic equilibrium state when a negative bias is present.

The last doubt about the presence of such a state can be blown away by examining the dynamics of the Langevin equation more precisely. Assume that a) the neuron temperature is zero and b) all weights are initially positive. Let $J := \mu(1-A)/N$ and $X_{ij} := J_{ij} - J$, then *as long as for each neuron the sum of its weights is positive* the original Langevin equation can be written:

$$\frac{d}{dt}X_{ij}(t) = -\frac{1}{\mu\tau}X_{ij}(t) + \frac{1}{\tau\sqrt{N}}\eta_{ij}(t). \tag{3.25}$$

This differential equation can be formally integrated from time $t = 0$ and squared:

$$\begin{aligned} X_{ij}(t) &= X_{ij}(0)\exp-\frac{1}{\mu\tau}t + \frac{1}{\tau\sqrt{N}}\int_0^t dt'\,\eta_{ij}(t')\exp-\frac{1}{\mu\tau}(t-t'), \\ X_{ij}(t)^2 &= X_{ij}(0)^2\exp-\frac{2}{\mu\tau}t + 2X_{ij}(0)\frac{1}{\tau\sqrt{N}}\int_0^t dt'\,\eta_{ij}(t')\exp-\frac{1}{\mu\tau}(t-t') \\ &\quad + \frac{1}{\tau^2 N}\int_0^t dt'\int_0^t dt''\,\eta_{ij}(t')\eta_{ij}(t'')\exp-\frac{1}{\mu\tau}(2t-t'-t''). \end{aligned}$$

We can average these equations over the possible realizations of the noise. From the definition of the unbiased noise, we find

$$\begin{aligned} \overline{X_{ij}(t)} &= X_{ij}(0)\exp-\frac{1}{\mu\tau}t, \qquad (3.26) \\ \overline{X_{ij}(t)^2} &= X_{ij}(0)^2\exp-\frac{2}{\mu\tau}t + \frac{2}{\tau\tilde{\beta}N}\int_0^t dt'\int_0^t dt''\,\exp-\frac{1}{\mu\tau}(2t-t'-t'')\delta(t'-t'') \\ &= X_{ij}(0)^2\exp-\frac{2}{\mu\tau}t + \frac{\mu}{\tilde{\beta}N}\left[1-\exp-\frac{2}{\mu\tau}t\right]. \end{aligned}$$

After a long time the variance will be $\mu/\tilde{\beta}N$. Is it possible that a spin is flipped? The flipping of a spin at site i occurs when the sum of its weights is negative, i.e. $\sum_j J_{ij} < 0$. In the transformed weight variables this is equivalent with $\sum_j X_{ij} < -\mu/(1-A)$. The mean of $\sum_j X_{ij}$ is of course zero until a spin flips, but the variance will be

$$\overline{\left(\sum_j X_{ij}\right)^2} = \sum_j \overline{X_{ij}^2} = \frac{\mu}{\tilde{\beta}} \tag{3.27}$$

As this variance is independent of N, it will not vanish in the thermodynamic limit. Therefore, if $\tilde{\beta} < \infty$, there will undeniable a time at which a spin is flipped and the original ferro-magnetic configuration is altered. When the spin is flipped its weights will quickly move away from $(1-A)/N$ to a new equilibrium around $-(A+1)/N$. When A is positive, the new equilibrium position will be farther away from the turning point, and the process will thus be asymmetric as the variance of the fluctuations is not altered.

Important is that this analysis also enables us to conclude that the specific choice made for the scaling of the terms in the Langevin equation (1.45) have made *the weight dynamics ergodic in the thermodynamic limit*. If there is no bias present and $\tilde{\beta}$ is very large, it can take a very long time before the sum of the weights of a spin changes sign and the spin flips. On short time scales, the system might be trapped in a certain Mattis-magnet configuration. On a longer time scale the system is not bound to one region of configuration space, eventually a spin will flip even in a spin zero temperature.

3.2 Continuum Limit

From considering just one cluster we go on to the other extreme: infinitely many clusters. We perform the limit of Λ to infinity, after we have taking the thermodynamic limit. In doing so we replace the discrete cluster labels λ, κ and ρ with the real vectors $\mathbf{x}$,$\mathbf{y}$ and $\mathbf{z}$. The set $\{1, 2, \ldots, \Lambda\}$ is replaced by $D \subset \mathbb{R}^d$. The solutions of this equation depend strongly on the choice of the dimension d and topology of D. Therefore we postpone the choice of D until we are trying to find explicit solutions. The capital Q and capital M expressions become integrals:

$$\begin{aligned} Q(\mathbf{x}) &= \int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) q(\mathbf{y}), \\ M(\mathbf{y}) &= -\int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) A(\mathbf{x},\mathbf{y}) m(\mathbf{y}). \end{aligned}$$

We consider the continuum limit only under the assumption of replica symmetry. The structure of the saddle-point equations (2.42) has not changed

$$\begin{aligned} q(\mathbf{x}) &= \left[\int \mathcal{D}\xi\, \cosh^n \Xi \tanh^2 \Xi\right] \left[\int \mathcal{D}\xi\, \cosh^n \Xi\right]^{-1}, \\ m(\mathbf{x}) &= \left[\int \mathcal{D}\xi\, \cosh^n \Xi \tanh \Xi\right] \left[\int \mathcal{D}\xi\, \cosh^n \Xi\right]^{-1}, \\ \Xi &= \xi\sqrt{\beta/n}\sqrt{Q(\mathbf{x})} + \beta M(\mathbf{x}) + \beta\theta(\mathbf{x}). \end{aligned} \tag{3.28}$$

From now on we work with the normalization condition on the decay μ:

$$\int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) = 1. \tag{3.29}$$

Using this normalization in the last equality we find

$$\begin{aligned} 0 &\leq \int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) \left(m(\mathbf{y}) - \int_D d\mathbf{z}\, \mu(\mathbf{x},\mathbf{z}) m(\mathbf{z})\right)^2 \\ &= \int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) m(\mathbf{y})^2 - \left(\int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) m(\mathbf{y})\right)^2 \end{aligned} \tag{3.30}$$

If $A(\mathbf{x},\mathbf{y}) = A$ we can use this inequality to see that

$$Q(\mathbf{x}) \geq \int_D d\mathbf{y}\, \mu(\mathbf{x},\mathbf{y}) m(\mathbf{y})^2 \geq \frac{1}{A^2} M(\mathbf{x})^2 \tag{3.31}$$

3.3 Translation Invariance and Convolutions

Systems that are translation invariant are much easier to study analytically than systems that lack this symmetry. Restricting the analysis to translation invariant hardware means that $A(\mathbf{x}, \mathbf{y})$ en $\mu(\mathbf{x}, \mathbf{y})$ will become functions of the relative position $\mathbf{x} - \mathbf{y}$ only. Also the dimensions of D need to be either infinite or closed into themselves. More unambigously stated: D is isomorphic to the Cartesian product of m real lines $\mathbb{R}$ with $p = d - m$ circles S^1. We assume D not only to be isomorphic but identical to $\mathbb{R}^m \times (S^1)^p$. For any two scalar functions f and g on D the *convolution* $f * g$ is defined by

$$(f * g)(\mathbf{x}) = \int_D d\mathbf{y}\, f(\mathbf{x} - \mathbf{y}) g(\mathbf{y}). \tag{3.32}$$

The defining equations of $Q(\mathbf{x})$ and $M(\mathbf{x})$ are convolutions and can be rewritten as

$$Q \equiv \mu * q, \qquad M \equiv -(\mu A) * m. \tag{3.33}$$

Now we define the appropriate multi-dimensional Fourier transform of the function $u \in L^1_D$:

$$\hat{u}(\mathbf{k}) = (2\pi)^{-d/2} \int_D d\mathbf{x}\, u(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} \quad \forall \mathbf{k} \in \hat{D}, \tag{3.34}$$

$$\hat{D} = \mathbb{R}^m \times \mathbb{Z}^p,$$

along with its inverse

$$u(\mathbf{x}) = (2\pi)^{-d/2} \int_{\mathbb{R}} dk^1 \cdots \int_{\mathbb{R}} dk^m \sum_{k^{m+1} \in \mathbb{Z}} \cdots \sum_{k^d \in \mathbb{Z}} \hat{u}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} \quad \forall \mathbf{x} \in D. \tag{3.35}$$

Applying the Fourier transform to our convolution equations yields

$$\begin{array}{rcll} (2\pi)^{-d/2} \hat{Q}(\mathbf{k}) & = & \hat{\mu}(\mathbf{k}) \hat{q}(\mathbf{k}) & \forall \mathbf{k} \in \hat{D}, \\ (2\pi)^{-d/2} \hat{M}(\mathbf{k}) & = & -\widehat{\mu A}(\mathbf{k}) \hat{m}(\mathbf{k}) & \forall \mathbf{k} \in \hat{D}. \end{array} \tag{3.36}$$

The necessary and sufficient conditions for the operations $\mu*$ and $(\mu A)*$ having an inverse are

$$\forall \mathbf{k} \in \hat{D} \qquad \hat{\mu}(\mathbf{k}) \neq 0, \quad \widehat{\mu A}(\mathbf{k}) \neq 0. \tag{3.37}$$

Note that the inverse decay be non-negative for all connections as a negative decay is effectively a growth factor and causes the weights to grow unlimited. As a result of this the zero Fourier component of μ is the largest component, $\mu(0) \geq \mu(\mathbf{k})$.

3.4 Bifurcations

If the temperature of the neurons is infinite, while the quotient of the neuron and weights temperatures is non zero, there is only one solution of the saddle point equations. Inserting $\beta = 0$ into (2.42) yields directly the trivial saddle point,

$$Q(\mathbf{x}) = 0, \quad M(\mathbf{x}) = 0.$$

In fact this solution solves the saddle point equations for all finite temperatures as well, but it need not be a stable solution of the system.

We will now use this solution as a starting point from which more interesting ones bifurcate as the two temperatures of the system are varied. Calculating the free energy of the found solutions will answer the question which saddle point is the actual equilibrium state of the network. By looking at bifurcations only, we will not be able to spot first order transitions.

The argument of the hyperbolic tangent suggests a variation around the trivial solution of the form:

$$M(\mathbf{x}) = \mathcal{O}(\varepsilon), \quad Q(\mathbf{x}) = \mathcal{O}(\varepsilon^2).$$

The replica symmetric saddle point equations assume the form

$$\begin{aligned} Q(\mathbf{x}) &= \beta^2 \int_D d\mathbf{y}\, \mu(\mathbf{x}-\mathbf{y}) \left[\frac{1}{\tilde{\beta}} Q(\mathbf{y}) + M(\mathbf{y})^2 \right] + \mathcal{O}(\varepsilon^4), \\ M(\mathbf{x}) &= -\beta \int_D d\mathbf{y}\, \mu(\mathbf{x}-\mathbf{y}) A(\mathbf{x}-\mathbf{y}) M(\mathbf{y}) + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{3.38}$$

Considering the previous remarks about convolutions, we can remove the integral by performing a Fourier transform:

$$\begin{aligned} (2\pi)^{-1} \hat{Q}(\mathbf{k}) &= \frac{\beta^2}{\tilde{\beta}} \hat{\mu}(\mathbf{k}) \hat{Q}(\mathbf{k}) + \beta^2 \hat{\mu}(\mathbf{k}) \widehat{M^2}(\mathbf{k}), \\ (2\pi)^{-1} \hat{M}(\mathbf{k}) &= -\beta \widehat{\mu A}(\mathbf{k}) \hat{M}(\mathbf{k}). \end{aligned} \tag{3.39}$$

Second order transitions can happen as soon as these equations allow non-trivial solutions.

3.4.1 An example in one dimension

Consider the following one dimensional example. For the domain of the clusters we take $D = S^1$. A natural function to consider for the inverse decay is high for small distances and small for distances near π. As a toy model we consider $\mu(x) = 1/2\pi + (a/2\pi)\cos x - (b/2\pi)\cos 2x$, $a + b \leq 1$ and $A > 0$ a constant. The Fourier components of μ that are non-zero are:

$$\hat{\mu}(k) = \begin{cases} \sqrt{2\pi}^{-1}, & k = 0 \\ \frac{1}{2} a \sqrt{2\pi}^{-1}, & k = \pm 1 \\ -\frac{1}{2} b \sqrt{2\pi}^{-1}, & k = \pm 2 \end{cases} \tag{3.40}$$

If $A = 0$, then the first bifurcation one will encounter when raising β is in the $\hat{Q}(0)$ component. This is because the $\hat{\mu}(0)$ component is the largest. If $A > 0$, then a bifurcation is possible in which the $k = \pm 2$ component of M becomes non-zero. This happens at $\beta = 2/bA$. If $\beta^2/\tilde{\beta} < 1$, or equivalently $2/bA < \tilde{\beta}^2$, this bifurcation occurs before the $Q = c$, $M = 0$ bifurcation. For β just after this point, a solution of (3.28) is possible with a small ϵ such that $\hat{M}(\pm 2) = \epsilon\sqrt{2\pi}$ or equivalently,

$$M(x) = \epsilon \cos 2x \qquad \text{and} \qquad m(x) = \frac{2\epsilon}{bA} \cos 2x \tag{3.41}$$

up to a translational constant. The bifurcation in the magnetization is accompanied with a bifurcation in the Edwards-Anderson order parameter. We know that this happens, because $Q(x) \geq M(x)^2/A^2$. How $Q(x)$ relates to the value of ϵ of the magnetization, we can calculate from (3.39). The $M(x)$ above yields for the zero Fourier-component of its square $\sqrt{2\pi}\epsilon^2/2$. The $k = \pm 1$ and $k = \pm 2$ components are zero. This means for the first Fourier-components of $Q(x)$:

$$\left.\begin{aligned} \hat{Q}(0) &= \tfrac{\beta}{n}\hat{Q}(0) + \tfrac{\beta^2}{2\sqrt{2\pi}}\epsilon^2 \\ \hat{Q}(\pm 1) &= \tfrac{\beta a}{2n}\hat{Q}(\pm 1) \\ \hat{Q}(\pm 2) &= -\tfrac{\beta b}{2n}\hat{Q}(\pm 2) \end{aligned}\right\} \Rightarrow \begin{aligned} \hat{Q}(0) &= \tfrac{\beta^2\sqrt{2\pi}}{2(1-\beta/n)}\epsilon^2 \\ \hat{Q}(\pm 1) &= 0 \quad \text{if} \quad \beta \neq \tfrac{2n}{a} \quad \text{else} \quad \hat{Q}(\pm 1) \propto \epsilon^2 \\ \hat{Q}(\pm 2) &= 0 \end{aligned} \tag{3.42}$$

The other Fourier components must be zero, because the higher modes of μ are chosen zero. The lines where bifurcations are possible are drawn in figure (3.6). To check whether the non-trivial solution with π-periodicity prevails over the $q = 0$, $m = 0$ solution we substitute both solutions in the continuum version of the free energy precursor $\phi(q, \hat{q}, m, \hat{m})$. In a saddle-point the hatted variables $\hat{q}$ and $\hat{m}$ are closely related to Q and M:

$$\hat{q}(\mathbf{x}) = i\frac{\beta^2}{\tilde{\beta}} Q(\mathbf{x}), \qquad \hat{m}(\mathbf{x}) = i\beta M(\mathbf{x}) + i\beta\theta(\mathbf{x}). \tag{3.43}$$

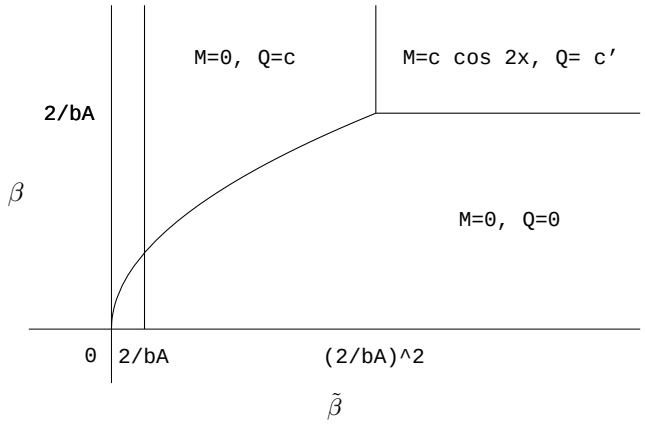


Figure 3.6: Phase diagram for the one dimensional system (3.40) showing the second order transitions from the trivial state $Q = 0, M = 0$ to states with some order.

To compare the values of ϕ in two saddle-points we substitute these hatted variables into ϕ. After some simple algebra similar to the steps we performed to find originally the replica symmetric saddle-point equations, we find

$$\begin{aligned}\tilde{\phi}(q,m) &= c + \frac{\beta^2}{2\tilde{\beta}^2}\int d\mathbf{x}\, q(\mathbf{x})Q(\mathbf{x}) + \frac{\beta}{2\tilde{\beta}}\int d\mathbf{x}\, M(\mathbf{x})m(\mathbf{x}) + \frac{\beta}{2\tilde{\beta}}\int d\mathbf{x}\, Q(\mathbf{x}) \\ &\quad -\frac{1}{\tilde{\beta}}\log\int_{-\infty}^{\infty}\mathcal{D}\xi\, \cosh^n\left(\sqrt{\frac{\beta^2}{\tilde{\beta}}Q(\mathbf{x})}\xi + \beta(M(\mathbf{x}) + \theta(\mathbf{x}))\right). \end{aligned} \tag{3.44}$$

Expanding this ϕ in powers of ϵ gives:

$$\tilde{\phi}(q,m) = c + \frac{\beta}{2\tilde{\beta}}\int_D d\mathbf{x}\, M(\mathbf{x})m(\mathbf{x}) - \frac{1}{2}\beta\int_D d\mathbf{x}\, M(\mathbf{x})^2 + \mathcal{O}(\epsilon^4). \tag{3.45}$$

The trivial saddle-points gives the value, $\tilde{\phi} = c$. The saddle-point $M(x) = \epsilon\cos 2\epsilon$ yields:

$$\tilde{\phi}(q,m) = c\beta\pi\left(\frac{1}{\tilde{\beta}bA} - \frac{1}{2}\right)\epsilon^2 + \mathcal{O}(\epsilon^4). \tag{3.46}$$

When the second term is negative, i.e. $\tilde{\beta} > 2/bA$, the 'free energy' of the solution showing some spatial structure is lower than the trivial solution. If we mildly ignore the possibility of a earlier first order transition to a state where $m = c \neq 0$, we see that this simple one-dimensional example system with a negative bias has an equilibrium state featuring spatial structure in the magnetization. Such a state will be present in a broad range of decay functions, however it depends crucially on the presence of negative Fourier-components. Not all functions have such components. For example, if we had taken the real line for the domain and a Gaussian function for the inverse decay, there would be no negative Fourier components as the Fourier transform of a Gaussian is just another Gaussian.

Choosing $M(x) = \epsilon\cos 2x$ we have broken the symmetry of the system. The symmetry-breaking did not occur only in the magnetization, but it also occurs in the weights. We can see this by looking at the continuum limit of the analysis in section 2.5 for the Gibbs-Boltzmann average of the coupling between two neurons i and j in clusters at x and y. This analysis yields:

$$\overline{NJ_{ij}} = \mu(x-y)\left(-A + m(x)m(y)\right). \tag{3.47}$$

The term $m(x)m(y)$ is not just a function of the distance between x and y and therefore the translation invariance of the weights is broken even in the Gibbs-Boltzmann average.

3.5 Linsker's structure

The setup of Linsker is easily embedded in the clusters. We will do this here in the continuum limit for the clusters. The entire network domain D is decomposed into L two-dimensional layers $\{D_1, D_2, \ldots D_L\}$. A point $\mathbf{x} \in \mathbb{R}^2$ in layer $l \in \{1, \ldots, L\}$ will be denoted as $(l, \mathbf{x})$. The connections between the clusters are governed by the inverse decay function μ. This function can be split into an inter-layer part $\hat{\mu}$ and a intra-layer or lateral part $\tilde{\mu}$ (note that the hat here has nothing to do with Fourier transforms):

$$\mu((l, \mathbf{x}), (l', \mathbf{y})) = \hat{\mu}(\mathbf{x} - \mathbf{y}) f(l + l')(\delta_{l,l'+1} + \delta_{l,l'-1}) + \tilde{\mu}(\mathbf{x} - \mathbf{y})(2l - 1)\delta_{l,l'} \tag{3.48}$$

Clearly this μ is symmetric if we choose both $\hat{\mu}$ and $\tilde{\mu}$ symmetric. The function $f : \mathbb{N} \to \mathbb{R}^+$ can provide a feed-forward structure if we choose it such that:

$$f(3) \gg f(5) \gg f(7) \gg \ldots \tag{3.49}$$

The volume of the layers D_l is layer independent, i.e. $\int_{D_l} d\mathbf{y} = \int_{D_l} d\mathbf{y}$ for all layers l, l'. And $\tilde{\mu}$ and $\hat{\mu}$ are of the same order of magnitude. The magnetization and Edwards-Anderson order in a cluster point $(l, \mathbf{x})$ only depend on the order in other points through $Q(l, \mathbf{x})$ and $M(l, \mathbf{x})$ as we can see from equations (3.28). With this choice of f only the cluster directly below l and the lateral connections will influence the clusters in l:

$$\begin{aligned} Q(l, \mathbf{x}) &\approx \int_{D_{l-1}} d\mathbf{y}\, f(2l-1)\hat{\mu}(\mathbf{x} - \mathbf{y}) + \int_{D_l} d\mathbf{y}\, f(2l-1)\tilde{\mu}(\mathbf{x} - \mathbf{y}), \\ M(l, \mathbf{x}) &\approx -\int_{D_{l-1}} d\mathbf{y}\, f(2l-1)\hat{A}(\mathbf{x} - \mathbf{y})\hat{\mu}(\mathbf{x} - \mathbf{y}) \int_{D_l} d\mathbf{y}\, f(2l-1)\tilde{A}(\mathbf{x} - \mathbf{y})\tilde{\mu}(\mathbf{x} - \mathbf{y}), \end{aligned} \tag{3.50}$$

where $\hat{A}$ and $\tilde{A}$ are of the same order of magnitude. The presence of the function factor $f(2l-1)$ will cause different layers to have different 'effective' temperatures. We see that the Linsker structure can be dealt with, but what about the results. With a $\hat{\mu}$ that is only distance dependent and that has a bell-shaped form combined with a negative intercluster bias, we hope to find at least a more pronounced structure in the averaged weights $\overline{J_{ij}}$. However, at the same time we would like the magnetization and Edwards-Anderson order to be constant within each layer. The Gibbs-Boltzmann average of the coupling between neurons i and j in respectively clusters $(l, \mathbf{x})$ and $(l-1, \mathbf{y})$ is (analogous to (2.5)):

$$\begin{aligned} \overline{NJ_{ij}} &\approx -\hat{A}(\mathbf{x}, \mathbf{y})\hat{\mu}(\mathbf{x} - \mathbf{y}) f(2l-1) + \hat{\mu}(\mathbf{x} - \mathbf{y}) f(2l-1) m(l, \mathbf{x}) m(l-1, \mathbf{y}) \\ &= \hat{\mu}(\mathbf{x} - \mathbf{y}) f(2l-1) \left(-\hat{A}(\mathbf{x}, \mathbf{y}) + m_l m_{l-1} \right). \end{aligned} \tag{3.51}$$

If we take a uniform bias $\hat{A}(\mathbf{x}, \mathbf{y}) = A$, we see that in the full equilibrium state of the network, the average value of the weights between two clusters will then exactly mirror the structure of the inverse decay function. Between two layers the sign of the average weights will be the same for all connections. No neurons with excitatory connections near the center of their receptive fields and inhibitory connections near the border (on-center neurons) are anticipated in this theory. Of course, Linsker fed his bottom layer with random patterns. We have not developed a method to produce macroscopic fluctuations in the magnetization of the clusters. However, this cannot explain the difference between Linsker's results and ours. As the weight dynamics are ergodic, the history of the first layer is irrelevant and cannot be the reason that we do not see more spatial structure in the network that we have put in.

We thus cannot reproduce the self-organization as seen in Linsker's model, when we look at the Gibbs-Boltzmann equilibrium of the network. On shorter time scale the history of the first layer is relevant and it may still be possible to get the weights close to a on-center configuration. To check this we should develop a method that deals with the dynamics of the weights, rather than with their equilibrium. For a system without any noise on the weight dynamics, work in this direction is done by Jonker and Coolen [17].

Chapter 4

Numerical Simulations

Although we have tried to include many properties of biological neural networks in our model, we have made too many concessions for calculability (e.g. symmetric interactions, fully connectedness), to compare the analytic results with experiments performed on real life neural networks. But we have questions we need to answer. Are we right when we use the replica symmetric ansatz if we are on the stable side of the AT-instability? Is the analytical calculation done correctly, i.e. is it justified to use the saddle-point approximation? Does the system indeed resemble a Mattis-glass most of the time? To answer these questions and check whether we have done the mathematical analysis right, one has to simulate the model.

For the one-cluster system analytical results have already been confirmed by Penney *et al.* [35]. When I wrote the computer program, I intended to use it for simulation of systems with more than one cluster. At present, however, the calculations consume so much time, that a multiple cluster simulation of a system with a size that would yield results, that could reliably be compared to the analytical $N = \infty$ results, is out of the question. Therefore I have done a new series of simulations of the one-cluster system, giving more evidence that the analysis of the previous chapters has been correct.

In this short chapter, we will describe how the neuron dynamics are implemented with a common Monte Carlo technique, the *Metropolis* algorithm. After the algorithm for the neural dynamics, the discretizing of the weight dynamics is shown. In the last two sections the setup and the results of some experiments are discussed.

4.1 Neural Dynamics

The way we have modeled the neuron dynamics in section (1.2) is a straightforward recipe for simulation on a computer.

1. Start at time $t = 0$ with a state $\sigma(t)$.
2. Select a neuron i at random.
3. Calculate local field $h_i(\sigma(t))$.
   The change in energy when neuron i is flipped, is $\Delta E \equiv 2\sigma_i(t)h_i(\sigma(t))$.
4. Flip neuron with chance $W_i = \frac{1}{2}(1 - \sigma_i f(h_i(\sigma))) = \frac{\exp -\beta\Delta E}{1 + \exp -\beta\Delta E}$.
5. Increase time and return to line 2.

This recipe is called the *Glauber* algorithm. If the neuron is not flipped in step four, then the time for calculation of the local field in step three is wasted. Metropolis *et al.* [26] devised a different Markov process which reduces this waste of computation time. For the Ising-model their process is described by an algorithm very similar to the one above. Only the fourth line is changed:

4. If $\Delta E \leq 0$, then flip neuron unconditionally. If $\Delta E > 0$, flip neuron with probability $\exp -\beta \Delta E$.

The Metropolis algorithm describes a Markov-process, just like the Glauber dynamics of chapter 1. The transition probability matrix for one time step is non-negative, but contains many zeros. Starting in a certain state, it is only possible to reach N out of 2^N other states in one time step. However, the transition matrix of N subsequent time steps is strictly positive if β is finite. Identical to what we have done in chapter 1 for the Glauber process, we can conclude by using Perron's theorem that the Metropolis process converges to unique stationary state.

This stationary state will be the same stationary state as the Glauber dynamics yield, i.e. the Gibbs-Boltzmann measure with energy H and inverse temperature β. One can check this easily, because the Gibbs-Boltzmann distribution also satisfies detailed balance in the Metropolis dynamics: consider a neural state vector σ and a neuron i selected for update. Define $\Delta E = H(F_i\sigma) - H(\sigma)$, the energy difference between the flipped and the original state. If $\Delta E \leq 0$, then for the probability of a transition to $F_i\sigma$ the Metropolis algorithm gives $W_{Met}(F_i\sigma \leftarrow \sigma) = 1$ and for the probability of neuron i to flip being in state $F_i\sigma$ it gives $W_{Met}(\sigma \leftarrow F_i\sigma) = \exp \beta \Delta E$. Substitution of the Gibbs-Boltzmann distribution for the stationary distribution $p(\sigma)$ yields detailed balance with respect to the transitions between $F_i\sigma$ and σ:

$$W_{Met}(F_i\sigma \leftarrow \sigma)p(\sigma) - W_{Met}(\sigma \leftarrow F_i\sigma) \propto 1.e^{-\beta H(\sigma)} - e^{\beta \Delta E} e^{-\beta H(F_i\sigma)} = 0 \tag{4.1}$$

If $\Delta E > 0$, the argument gives the same result. The condition that a state is stationary, is for these single-neuron flip processes the sum of the above detailed balance condition over all possible flips.

The Metropolis algorithm is less likely to reject a flip in step 5 compared to the Glauber algorithm, because both for $\delta E \leq 0$ and $\delta E > 0$ the flip probability is higher in the Metropolis recipe. The conclusion is that the Metropolis algorithm shares the equilibrium state with the Glauber dynamics, but that it might converge faster.

4.2 Weight Dynamics

The Langevin equation (1.45) that defines the time evolution of the weights is a continuous time differential equation and numerical simulation will be numerical integration. We use the most intuitive (and crudest) way of doing this, namely the *Euler feed-forward* method. Time is discretized in steps of $\Delta t \equiv \tau \delta$. The change in a weight in the time step from t to $t + \Delta t$ is calculated using the values of the weights at time t only. For each time step we calculate for all weights (sequently and in a fixed order)

$$J_{ij}(t + \Delta t) = J_{ij}(t) + \delta \left(\frac{1}{N} \langle \sigma_i \sigma_j \rangle_{J(t)} - \frac{1}{N} A_{\lambda_i \lambda_j} - \frac{1}{\mu_{\lambda_i \lambda_j}} J_{ij} \right) + \sqrt{\frac{2\delta}{\tilde{\beta} N}} \mathcal{N}(0,1), \qquad i < j, \tag{4.2}$$

where $\mathcal{N}(0,1)$ is a random variable chosen from a normal distribution for each weight at each time step. The derivation of the scale of the noise is done in appendix A. For numerical integration there are methods that use the same computation time but get results within a smaller error margin. These methods require more effort to implement and because a noise is already part of the system we do not except computational errors to alter the behavior of the solution.

4.3 Simulations

The simulation of the coupled system of weight and neural dynamics starts with choosing a set of couplings. Subsequently let the neurons converge towards equilibrium, update the weights, let the neurons reach near equilibrium again, update weights, etc. until the weights have reached equilibrium. The number of iterations of Monte Carlo steps (randomly selecting a neuron for

Figure 4.1: The Edwards-Anderson order parameter q during a simulation run for a 40 neuron system. In the first $R_3 = 500$ weight updates the system is heated to $\beta = 1$ and $\tilde{\beta} = 3$. In the last $R_4/4 = 375$ weight updates the equilibrium value for q is measured.

update and deciding whether or not to flip its state) to let the neurons converge to equilibrium is $N \times R_1$. The parameter R_1 should scale exponentially with N, but we will just choose a value for R_1 and compare the results of one of the simulations with a much larger $R1$ to check if the results agree and R_1 was not chosen too small. To calculate the equilibrium values for the variance of two spins (needed for the update of the weights) another $N \times R_2$ Monte Carlo steps are performed. During these steps also the magnetization and the Edwards-Anderson order parameter are calculated.

After one session of simulating the neural dynamics the weights are updated. The computation time needed for the a single change in all weights scales as N^2. Then the Monte Carlo steps are repeated. This whole procedure of neural dynamics and weight update is first done in a low temperature regime (a high β as well as a high $\tilde{\beta}$). This to ensure that the system tends to start in an ordered state. During the next iterations of the whole procedure the system is heated to reach after R_3 iterations the noise regime for which we want to measure the order in the system. Then the weights are updated another R_4 times. It is during the last quarter of these iterations that the system is supposed to be near equilibrium and the order parameters are calculated. R_4 needs to grow with system size and $R_4 \Delta t$ has to be much larger than the largest decay time $\mu_{\kappa\lambda}\tau$, i.e. $R_4\delta \gg \mu_{\kappa\lambda}$ for all κ, λ. A sample run is shown in figure 4.1.

4.4 Experiments

For doing the simulations, one needs to choose the iteration numbers R_1, R_2, R_3, R_4 and the time step δ. The time step should be chosen so small that the discretizatized weight update rule yields a good approximation of the continuous rule for the time we want to simulate the system. I have used $\delta = 0.01$ which is the same value used in [35]. The number of iterations should be chosen large enough for the system to (almost) reach equilibrium. It is difficult to calculate analyticly how many neuron and weight updates are necessary to reach equilibrium. Nor is it easy to see in a simulation when equilibrium is reached. Measuring the order parameters during the simulations gives some indication. In figure 4.1, for example, one sees that after the temperature has reached its final value, the order first drops and than slowly climbs to reach a stable level at about 0.9. Although by looking at the figure, we cannot ascertain that the system has reached equilibrium after 1500 weight updates, we can be sure that it did not reach it long before that time. By looking at the order during the neuron and weight updates I have taken the smallest number of iterations that I felt save with. The numbers I used are

$$R1 = 250, \quad R2 = 250, \quad R3 = 500, \quad R4 = 1500. \tag{4.3}$$

For comparison, Penney *et al.* used $R1, R2 \approx 250$, $R3, R4 \approx 500$. They have found that by doubling the equilibrium and averaging times the results do not alter qualitatively. Yet by looking at the evolution of the order parameter q, I did not felt safe with smaller values of $R3$ and $R4$.

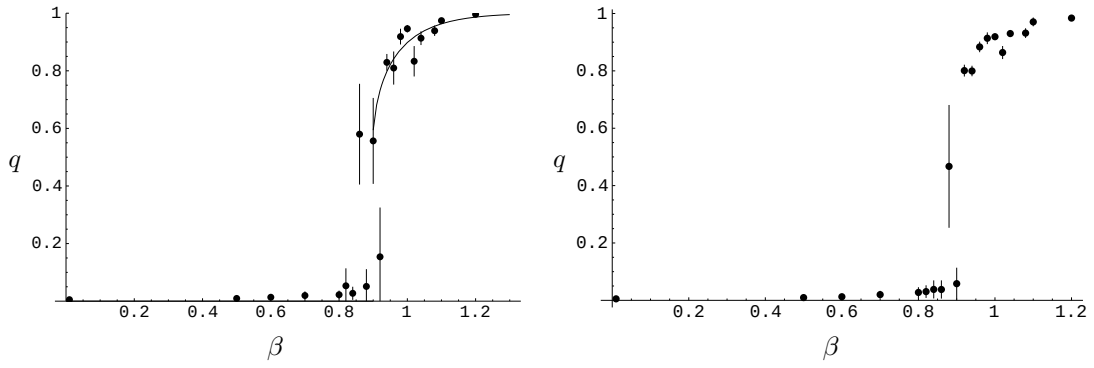


Figure 4.2: A first order transition between $q = 0$ and $q \neq 0$ for $n = 3$. On the left the measurements of a fully connected 40 neuron system, on the right the results of a network with 80 neurons. Each dot is the average of a single run. The error bar is the standard deviation of q during that run. The full line is the theoretical replica symmetric value.

With the above choice of parameters, a simulation run of an eighty neuron system takes in the order of an hour on a workstation. The experiments I have done are limited to forty and eighty neuron system sizes. These are also the systems sizes studied earlier by Penney. To make it easier to compare the results with the earlier work, I have chosen to simulate the system in the SK-parametrization (3.4), where β and n are used for parametrization the noise levels and instead of the inverse decay μ the variance $\tilde{J} = \mu/\beta n$ is fixed when changing temperature. Simulations have been done on unbiased systems for $n = 1$ and $n = 3$. In figures 4.2 and 4.3 the mean and standard deviation of the order parameters of single runs are shown for the $n = 3$ and $n = 1$ systems respectively. The full lines indicates the theoretical value predicted by the replica symmetric approach. As predicted, see the phase diagram in figure 3.2, the $n = 1$ has a second order phase transition from the $q = 0$ state to a $q \neq 0$ state and $n = 3$ has a first order transition. The averaged order during the trials of the $n = 1$ system is consistently too low. That finite size effects can account for this discrepancy of result and theory is clear from looking at an individual trial, such as shown on the right in figure 4.3. For large periods of time the system's order is near the theoretically predicted value. This corresponds to the neuron system being in a certain valley of the free energy. Because the finite size system is ergodic, the system might climb a barrier between two valleys. At the top of the barrier the order has almost vanished, only to quickly return again when the system falls back into a valley. The two graphs in figure 4.2 give another indication of the finite-size effects as the left graph is the result of a $N = 40$ simulation and the right graph shows the result of $N = 80$ system.

I have also done experiments to see whether the name *Mattis glass* can be justified. For various $\tilde{\beta}$ I have searched for stable states in an $N = 200$ system. After a certain time simulating without neural noise I froze the weights, and let the neurons cool from a very high neuron temperature to zero temperature. Preliminary results show that for low weight temperatures, $\tilde{\beta} > 6$, the ground states are the only two final states of the system, indicating that the ground states are the only two very stable states. For $\tilde{\beta} < 4$ the cooled neurons can be trapped in many different states. These results do not conflict with the interpretation of the system's behavior given in section 2.4.1, but the system is interpreted as a Mattis glass only in the thermodynamic limit where the ergodicity of the neural dynamics is broken. It is difficult to predict the number of ergodic components in the $N = \infty$ system by looking at the stable states of considerably smaller system.

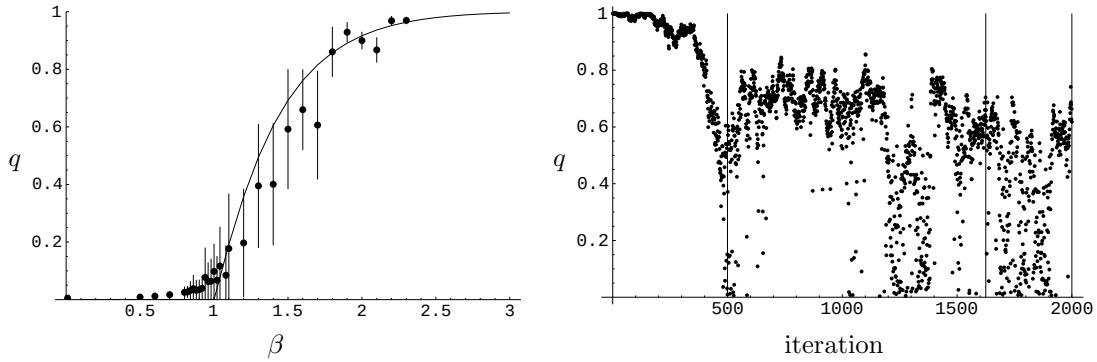


Figure 4.3: The system exhibits a second order transition from $q = 0$ to $q \neq 0$ for $n = 1$. The experiments shown on the left are done with a fully connected network of 40 neurons. Each dot is the average of a single run. The error bar is the standard deviation of q during that run. On the right q during a single run with $\beta = 1.40$ is shown.

Chapter 5

Conclusion

The work on this thesis started with the goal to apply the analytical work of Penney, Coolen and Sherrington [5] to network configurations that can model the development of orientation sensitive neurons. The articles of Penney *et al.* deal with Hebbian learning Ising-type neurons, whereas the previous analytical work explaining the growth of orientation sensitive neurons by Hebbian learning is done on linear response neurons. It is interesting to see if the onset of orientation sensitive neurons is also possible in the stochastic Ising-type neural networks. The main conclusion of this thesis is that I have not been able to extend the work of Penney *et al.* in such a way that these special cells are present in equilibrium. The ergodicity in the weight space of the investigated system prevents a lasting presence of such cells.

5.1 Discussion

What is shown in this work is that one can extend the PCS-model with a clustering of the neurons to make an initial spatial structure in the connections between neurons possible. The number of clusters is hold fixed when taking the thermodynamic limit. The presence of the cluster dependence of the systems parameters as the weight decay and the weight bias does at first sight not alter the analysis very much. We find that instead of needing two order parameters, the averaged magnetization and the averaged squared magnetization per site, to describe the system, we need to find these two order parameters for each of the clusters. The number of self-consistency equations that we need to solve grows linearly with the number of clusters, but the structure of the equations remains the same.

To derive the self-consistency equations we need to assume that using the saddle-point method for analyzing the model in the thermodynamic limit (the limit where the number of neurons goes to infinity) yields the correct answer. For the one cluster PCS-model with a non-negative bias the equations can also be derived in another way, avoiding the saddle-point method (see Appendix C). Under the condition that the inverse decay matrix and the matrix composed of the inverse decay and the weight bias are both positive definite, we can use this alternative route for a general number of clusters. The biases, that would best fit the Linsker configuration, are negative and if we consider this type of systems, I have not been able to prove the validity of the results found by using the saddle-point method. Another assumption that is shown to be correct for a large temperature region in the one-cluster system is the stability of the replica symmetric solution. I did not succeed in extendeding the proof to a multiple-cluster system. The most solid results in this thesis are therefore found in the analysis of the model of Penney, Coolen and Sherrington. Below I will first discuss the conclusions we can draw about the PCS-model, before making any further remarks on the multiple cluster systems.

5.1.1 One cluster system

The main part of this thesis has been reproducing the results found by Penney *et al.* in [35]. Their results have been confirmed. The behavior of the unbiased system can be categorized in three classes: paramagnet, Mattis glass and spin glass. Whether the system behaves according to one of these types, depends on the two noise levels. If both the noise in the neural dynamics and the noise in the weight dynamics are relatively low, the system will be in a Mattis glass state, i.e. the magnetization of the entire system is zero ($m = 0$), while individual neurons have a preferred state ($q \neq 0$). If the neural dynamics is very noisy, the neurons will not have a preferred state and the system is in a paramagnetic state ($m = 0, q = 0$). If not the neural dynamics but the weight dynamics are very noise, the system is a spin glass state. This spin glass state is also characterized by $q \neq 0$, $m = 0$. Therefore the spin glass and the Mattis glass phase are not easy to tell apart by measurement. The spin glass phase is defined as the region where replica symmetry is broken. The physical distinction between spin glass and Mattis glass is that in the spin glass phase there are at each time during measurements on the system, many (probably a number scaling like α^N for some positive α) (meta)stable neuron states in the spin glass phase whereas in the Mattis glass phase we expect a much smaller number of stable states. The number of stable states could be the subject of a follow-up research. Introduction of a positive bias leads to the appearance of a ferromagnetic phase for relatively low noise levels. In this phase the system has a non-zero magnetic moment ($m \neq 0, q \neq 0$). The transitions between the four phases are identical to the transitions shown in [35], except for a small discrepancy in the first order transition from Mattis glass to the paramagnetic phase.

The computer simulations that have been published in [35] and the simulations done by me are in agreement with the theoretically calculated phase diagram.

An important feature that has not been shown explicitly in earlier studies of this neural net model is that the weight dynamics remain ergodic. This indicates not only that the system can learn anything by chance, but also that it will certainly forget in time everything that has been learned. The model can be adjusted by introducing a time dependent learning rate or changing the scale of the noise in the weight dynamics, but in both cases the system would no longer reach a state of thermal equilibrium. The analysis of such a model will be very different from the analysis done is this work.

A question that was asked me is, what does the PCS neural network learn? The answer is that it learns only noise, noise in the weight dynamics and noise in the neural dynamics. The system could be fed with input from the outside world by applying locally external fields to the neurons (comparable to light hitting cells in the retina). The effect of a cluster dependent constant external field can be determined well within the analytical framework of this thesis, but as we have only done an analysis on the equilibrium behavior of the system, the analysis can not extended easily to a time dependent external field. The ergodicity of the system tells us that the equilibrium of the system will not reflect the varying external fields to which the network was subjected earlier.

5.1.2 Many cluster system

The self-consistency equations derived in chapter two could be used for any number of clusters (assuming that the equations are correct). I have found no reason for studying any number of clusters in particular. For an attempt to find organization of the receptive fields, a study of a system with a large number of clusters is needed. The surprising step to reduce complexity is to consider not just a large, but an infinite number of clusters. In this way the order parameters of several different clusters become one order parameter function and under the replica symmetry ansatz the many coupled non-linear saddle-point equations reduce to two functional equations.

I have not done extensive analysis of this system. Many cluster-cluster interactions (decays and biases) could be evaluated on many different network topologies. In chapter three I have shown that even in a very simple neural network where the clusters are lying on a circle, spatial order in the magnetization can arise. I have not tried to get a confirmation of theoretical result by means of computer simulations. The computer program I have made is capable of a multiple

cluster simulation, but simulating the system means simulating a large number of clusters. All this clusters contain a large number of neurons. To get any result that could be trusted takes far to long.

To model the spontaneous onset of orientation selective neurons or clusters, we are interested in spatial ordering of the couplings instead of spatial order in the order parameters. We would have liked to see ordering in the receptive fields of neurons, a part of the couplings of a particular neuron becoming negative, another part becoming positive. The average value of the weights in equilibrium is $N\langle J_{ij}\rangle = \mu_{\lambda\kappa}(-A_{\lambda\kappa} + m_\lambda m_\kappa)$, where $i \in I_\lambda$ and $j \in I_\kappa$, showing that the sign of a weight average is completely determined by $-A_{\lambda\kappa} + m_\lambda m_\kappa$ as the inverse decay $\mu_{\kappa\lambda}$ is non negative. We do not want the magnetization to reflect any spatial ordering (within a layer). This leaves only the bias to make average weights of different sign possible. One could choose the biases such that on-center neurons or orientation selective neurons directly arise. However, this would be cheating, as one needs to set the signs of the biases precisely the way one wants sign of the average weights to be. The conclusion is that by looking at the equilibrium of the system one will not find orientation selective neurons if one does not put them in. A method to describe and calculate the organization of the network (including the order parameters and the weight configuration) away from equilibrium is needed.

Appendix A

Equilibrium of a Langevin System

In this appendix it is proven that a Langevin equation with a conservative force converges in time to equilibrium. In order to use the proof of Coolen [6], the Langevin equation is first transformed into a Fokker-Planck or generalized diffusion equation. Along the way we will see how to discretize the time. This will be necessary for doing any computer simulations of a Langevin equation.

A.1 Langevin Equation

Originally to describe a particle in Brownian motion, Langevin (1908) used a linear deterministic equation of motion to which he added some external noise. In this appendix, we derive the stationary state of a system, parametrized by the N-vector $\mathbf{y}$, subject to a generalized Langevin equation of motion

$$\tau \frac{d}{dt}\mathbf{y}(t) = \mathbf{f}(\mathbf{y}) + \boldsymbol{\eta}(t), \tag{A.1}$$

where τ defines the time-scale of the process. The first term in the r.h.s. is a deterministic force on the system, the second term is a *Gaussian white noise*. It's behavior is completely determined by its first two moments in the noise sample average $\langle\cdot\rangle$:

$$\langle \eta_i(t) \rangle = 0, \qquad \langle \eta_i(t)\eta_j(t') \rangle = 2\tau\beta^{-1}\delta_{ij}\delta(t-t'). \tag{A.2}$$

The higher moments follow from the noise being Gaussian. For one sample of the noise, the difference $\Delta\mathbf{y}$ of the state of the system between time t_0 and $t_0 + \Delta t$ is exactly:

$$\tau\Delta\mathbf{y} = \int_{t_0}^{t_0+\Delta t} \mathbf{f}(\mathbf{y}(t'))dt' + \hat{\boldsymbol{\eta}}(t_0), \tag{A.3}$$

where we have defined

$$\hat{\boldsymbol{\eta}}(t_0) \equiv \int_{t_0}^{t_0+\Delta t} \boldsymbol{\eta}(t')dt'.$$

When $\mathbf{f}(\mathbf{y})$) is a smooth function, a sample average of $\Delta\mathbf{y}$ becomes for small Δt:

$$\tau\langle\Delta y_i\rangle = f_i(\mathbf{y}(t_0))\Delta t + \mathcal{O}(\Delta t^2), \qquad (\Delta t \downarrow 0) \tag{A.4}$$

where y_i and f_i mean the ith-component of $\mathbf{y}$ and $\mathbf{f}$. The second moments of the deviation from $\mathbf{y}(t_0)$ are given by

$$\begin{aligned} \tau^2\langle\Delta y_i \Delta y_j\rangle &= \int_{t_0}^{t_0+\Delta t} dt \int_{t_0}^{t_0+\Delta t} dt' \langle f_i(\mathbf{y}(t)) f_j(\mathbf{y}(t'))\rangle \\ &+ \int_{t_0}^{t_0+\Delta t} dt \langle f_i(\mathbf{y}(t))\hat{\eta}_j(t_0)\rangle + \int_{t_0}^{t_0+\Delta t} dt \langle f_j(\mathbf{y}(t))\hat{\eta}_i(t_0)\rangle \\ &+ \int_{t_0}^{t_0+\Delta t} dt \int_{t_0}^{t_0+\Delta t} dt' \, \langle \hat{\eta}_i(t)\hat{\eta}_j(t')\rangle. \end{aligned} \tag{A.5}$$

The first term is $\mathcal{O}(\Delta t^2)$. Taylor expansion of $\mathbf{f}$ around $\mathbf{y}(t_0)$ shows the terms on the second line to be of $o(\Delta t^2)$. The third term is calculated easily:

$$\langle \hat{\eta}_i(t_0)\hat{\eta}_j(t_0)\rangle = 2\Delta t\tau\beta^{-1}\delta_{ij} + o(\Delta t^2) \qquad (\Delta t \downarrow 0). \tag{A.6}$$

The second moment of the deviation of the starting position will contain in the lowest order of Δt only this contribution:

$$\langle \Delta y_i \Delta y_j\rangle = \frac{2\Delta t}{\beta\tau}\delta_{ij} + o(\Delta t^2) \qquad (\Delta t \downarrow 0). \tag{A.7}$$

Higher moments of the deviation, $\langle \Delta y_{i_1} \cdots \Delta y_{i_n}\rangle$ are for $n > 2$ all of order $o(\Delta t)$ by the specific choice of the noise. In the next section we will use these first two moments to describe the evolution of the system in the form of an evolution of a probability function. This will prove to be a more fruitful approach to calculate the long time behavior of the system.

Numerical integration

The differential equation (A.1) can be solved numerically by a number of methods. Perhaps the easiest way of doing this, is Euler-forward numerical integration of the differential equation. This procedure is very simple. We start with a known position $\mathbf{y}$ at time t, we calculate for that position the derivative according to (A.1) and multiply this derivative with a chosen time step Δt. For the noise we use (A.6). If we parametrize the time step by $\Delta t = \tau\delta$, the integration is equal to calculating and summing the steps:

$$y_i(t + \Delta t) - y_i(t) = \delta f_i(\mathbf{y}(t)) + \sqrt{\frac{2\delta}{\beta}} N(0,1)_{i,t}, \tag{A.8}$$

where $N(0,1)_{i,t}$ is normally distributed stochastic variable. This is only correct for $\delta \to 0$. For finite δ this is only a numerical approximation. The smaller δ, the smaller the error margin in the approximation.

A.2 Fokker-Planck Equation

Consider the following generalized diffusion or Fokker-Planck equation:

$$\frac{\partial P(\mathbf{y},t)}{\partial t} = -\sum_j \frac{\partial}{\partial y_j} f_j(\mathbf{y}) P(\mathbf{y},t) + \beta^{-1} \sum_j \frac{\partial^2}{\partial y_j^2} P(\mathbf{y},t). \tag{A.9}$$

This equation describes the evolution of the probability $P(\mathbf{y},t)$ of a system being in a certain state $\mathbf{y}$. Assume that we know that at time t_0, the system is in state $\mathbf{y}_0$. This knowledge is made explicit by defining $P(\mathbf{y},t) \equiv P(\mathbf{y},t|\mathbf{y}_0,t_0)$. At time t_0, $P(\mathbf{y},t_0)$ is a delta function peaked around $\mathbf{y}_0$. A short time later, say at $t_0 + \Delta t$, $P(\mathbf{y},t_0+\Delta t)$ is no longer a delta function. We cannot calculate $y(t)$ anymore, but the Fokker-Planck equation gives us just enough information to calculate the moments of $\Delta\mathbf{y} \equiv \mathbf{y}(t) - \mathbf{y}_0$. For instance, the average performed with $P(\mathbf{y},t_0+\Delta t)$ yields:

$$\begin{aligned} \langle \Delta y_i\rangle &= \int d\mathbf{y}(y_i - y_{0,i}) P(\mathbf{y},t_0+\Delta t) \\ &= \int d\mathbf{y}(y_i - y_{0,i}) P(\mathbf{y},t_0) + \Delta t \int d\mathbf{y} \frac{\partial P(\mathbf{y},t_0}{\partial t}(y_i - y_{i,0}). \end{aligned} \tag{A.10}$$

The first term vanishes, due to the delta character of $P(\mathbf{y},t_0)$. Next we insert the Fokker-Planck equation (A.9) and perform some partial integrations.

$$\begin{aligned} \langle \Delta y_i\rangle &= -\Delta t \sum_j \int d\mathbf{y}(y_i - y_{0,i}) \frac{\partial}{\partial y_j} f_j(\mathbf{y}) P(\mathbf{y},t_0) + \Delta t\beta^{-1} \sum_j \int d\mathbf{y}(y_i - y_{0,i}) \frac{\partial^2}{\partial y_j^2} P(\mathbf{y},t_0) \\ &= \Delta t \int d\mathbf{y} f_i(\mathbf{y}) P(\mathbf{y},t_0) = \Delta t f_i(y_0) + o(\Delta t) \qquad (\Delta t \downarrow 0). \end{aligned} \tag{A.11}$$

The second moments can be calculated in a similar way:

$$\langle \Delta y_i \Delta y_j \rangle = 2\beta^{-1}\delta_{ij}\Delta t + o(\Delta t) \qquad (\Delta t \downarrow 0). \tag{A.12}$$

All higher moments are $o(\Delta t)$ for small Δt. The moments of Δy_i of this Fokker-Planck equation are exactly equal to the moments found for the Langevin process. We conclude (see Van Kampen [18]) that the two equations (A.1) and (A.9) are different descriptions of one and the same (idealized) physical process. In the next section it will become apparent how the new way of describing the Langevin process can be very useful.

A.3 Conservative Forces

Assume the force $\mathbf{f}(\mathbf{y})$ to be conservative, i.e, there exists a scalar function $H(\mathbf{y})$ such that

$$f_i(\mathbf{y}) = -\frac{\partial}{\partial y_i} H(\mathbf{y}) \qquad \text{for all } i \in \{1, \ldots, N\}. \tag{A.13}$$

If such an $H(\mathbf{y})$ exist we call it a Hamiltonian of the system. The function $\exp -\beta H(\mathbf{y})$ is called the *Gibbs-Boltzmann measure*. Suppose there exists a number Z such that the integral of the Gibbs-Boltzmann measure can be normalized, i.e.

$$1 = Z^{-1} \int d\mathbf{y} e^{-\beta H(\mathbf{y})}. \tag{A.14}$$

It is then easy to show that the evolution of the Gibbs-Boltzmann probability distribution $P(\mathbf{y}) = Z^{-1} \exp -\beta H(\mathbf{y})$ is stationary, by just substituting this $P(y)$ into the Fokker-Planck equation (A.9) with the force (A.13).

Following Coolen [6], we set out to prove that the Gibbs-Boltzmann distribution, provided that is normalizable, is the unique stationary probability function and find the condition which the initial probability distribution $P(\mathbf{y}, t_0)$ must obey to converge to this distribution.

We start by transforming the time-dependent probability function in the following way

$$\psi(\mathbf{y}, t) = P(\mathbf{y}, t) e^{\frac{1}{2}\beta H(\mathbf{y})}. \tag{A.15}$$

The time evolution of this function can be found using the Fokker-Planck equation. Some simple algebra gives

$$\frac{\partial \psi(\mathbf{y}, t)}{\partial t} = \mathcal{L}\psi(\mathbf{y}, t), \tag{A.16}$$

with the differential operator

$$\mathcal{L} = -\frac{1}{4}\beta \sum_i \left[\frac{\partial H(\mathbf{y})}{\partial y_i} \right]^2 + \frac{1}{2} \sum_i \frac{\partial^2 H(\mathbf{y})}{\partial y_i^2} + \beta^{-1} \sum_i \frac{\partial^2}{\partial y_i^2}. \tag{A.17}$$

In the Hilbert space $L_2(\mathbb{R}^N)$, which has the inner product $\langle \psi | \phi \rangle = \int_{\mathbb{R}^N} d\mathbf{y} \psi(\mathbf{y}) \phi(\mathbf{y})$, this operator is clearly Hermitian. What is more, this knowledge can be made explicit by writing $\mathcal{L}$ as a sum of 'squares'

$$\mathcal{L} = -\sum_i A_i^\dagger A_i, \qquad A_i = \beta^{-\frac{1}{2}} \frac{\partial}{\partial y_i} + \frac{1}{2}\beta^{\frac{1}{2}} \frac{\partial H}{\partial y_i}. \tag{A.18}$$

Being Hermitian, $\mathcal{L}$ has a complete set of eigenfunctions. Consider an eigenfunction ϕ. The corresponding eigenvalue λ will be non-negative because

$$\langle \phi | \mathcal{L} \phi \rangle = -\sum_i \langle \phi | A_i^\dagger A_i \phi \rangle \leq 0. \tag{A.19}$$

The function ϕ_0 can be an eigenvector of $\mathcal{L}$ with eigenvalue 0, if for all i, $A_i\phi_0(\mathbf{y}) = 0$. These equations have a unique solution (up to a multiplicative constant):

$$\phi_0(\mathbf{y}) = ce^{-\frac{1}{2}\beta H(\mathbf{y})}. \tag{A.20}$$

The eigenvalue 0 exists if $\phi_0 \in L_2$ and it will be non-degenerate. Let the remainder of the complete set of eigenvectors of $\mathcal{L}$ be written as $\{\phi_1, \phi_2, \ldots\}$ and the corresponding eigenvalues as $\{\lambda_1, \lambda_2, \ldots\}$ (if the spectrum is continous, the notation should be adjusted, but the following conclusion remains valid).

If the starting point of the evolution, $\psi_{t_0}(\mathbf{y}) \equiv \psi(\mathbf{y}, t_0)$ is an L_2-function, we can formally write the solution $\psi_t(\mathbf{y}) \equiv \psi(\mathbf{y}, t)$ of (A.16) in Dirac bra-ket notation

$$|\psi_t\rangle = |e^{t\mathcal{L}}\psi_{t_0}\rangle. \tag{A.21}$$

Using completeness, this solution can be decomposed into eigenfunctions of $\mathcal{L}$

$$\begin{aligned} |\psi_t\rangle &= \sum_{i=1} |\phi_i\rangle\langle\phi_i|\psi_t\rangle + |\phi_0\rangle\langle\phi_0|\psi_t\rangle \\ &= \sum_{i=1} |\phi_i\rangle e^{\lambda_i t}\langle\phi_i|\psi_{t_0}\rangle + |\phi_0\rangle\langle\phi_0|\psi_{t_0}\rangle. \end{aligned} \tag{A.22}$$

After a very long time, only the last term has not vanished, as all $\lambda_i < 0$ for $i > 0$. If we apply the transformation (A.15) in reverse to the functions ψ_0 and ψ_{t_0}, we have proved the following theorem.

Theorem 1 *Let the function $H : \mathbb{R}^N \to \mathbb{R}$, $\beta \in \mathbb{R}$ larger than zero, and $P_{t_0} \in L_1(\mathbb{R}^N)$. Further assume:*

$$\int_{\mathbb{R}^N} d\mathbf{y} e^{-\beta H(\mathbf{y})} < \infty, \quad \textit{and} \quad \int_{\mathbb{R}^N} d\mathbf{y} P_{t_0}(\mathbf{y})^2 e^{\beta H(\mathbf{y})} < \infty. \tag{A.23}$$

And assume that $P(\mathbf{y}, t)$ is a solution of the Fokker-Planck equation (A.9) with the boundary condition $P(\mathbf{y}, t_0) = P_(t_0)$. Then

$$\lim_{t\to\infty} P(\mathbf{y}, t) \propto e^{-\beta H(\mathbf{y})}. \tag{A.24}$$

Appendix B

Saddle-Point Method

In this appendix an approximation for a special type of integrals is described. The method is due to Laplace (1820), but the precise conditions and proof are of a later date. First we treat Laplace's original method on the real space. Secondly we treat the saddle-point method, the complex extension of Laplace's method. Before stating the methods, we give a short heuristic derivation. Although the proofs are not really lengthy (about two pages each) an introduction to the theory of asymptotics would be necessary. It would go too far to give such an introduction here. The proofs of the theorems can be found in [30], the authorative text about saddle-point method and asymptotics in general. The theorems taken from Olver can not be applied directly to our problem. In an extra corollary and an extra theorem, Laplace's method is made to fit our precise needs.

B.1 Laplace's Method

Consider the integral given by

$$I(x) = \int_a^b e^{-xp(t)} dt \tag{B.1}$$

If the function $p(t)$ has a single global minimum in the interior point t_0, the function $e^{-xp(t)}$ will be very sharply peaked around that point when x is really large. The dominant contribution to the integral will be from that peak (see figure (B.1)). This leads us to consider the Taylor expansion of $p(t)$ around the point t_0 and change the domain of the integral to the entire real line. If $p''(t_0) \neq 0$, the resulting integral is a simple Gaussian integral and can be evaluated easily:

$$I(x) \approx e^{-xp(t_0)} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} xp''(t_0)(t-t_0)^2\right] dt \tag{B.2}$$

$$= e^{-xp(t_0)} \left[\frac{2\pi}{xp''(t_0)}\right]^{1/2} \tag{B.3}$$

The validity of this procedure is of course subject to certain conditions. The proof is based on the theory of asymptotics and therefore the conditions use this language, in particular the relation $f(x)$ *is asymptotic to* $\phi(x)$:

Definition: if $\lim_{x\to\infty} f(x)/\phi(x) = 1$, we write:

$$f(x) \sim \phi(x) \qquad (x \to \infty)$$

If there is no risk of ambiguity, we simply write $f \sim \phi$.

Using this notation the basis for the Laplace method is formalized in the following theorem.

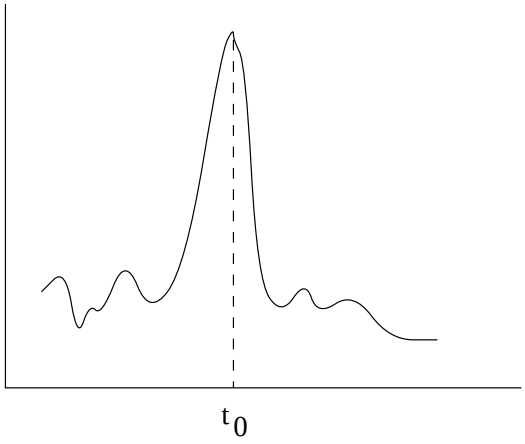


Figure B.1: Laplace's method. The dominant contribution to the integral of the function will be from around the peak at t_0.

Theorem 2 (Laplace's Method [30]) *Consider the integral*

$$I(x) = \int_a^b e^{-xp(t)} q(t) dt, \tag{B.4}$$

where the limits a, b and the functions $p(t)$ and $q(t)$ all are independent of x. The limits a and b are real, a is finite and b is larger than a and either finite or infinite. The function $p(t)$ is a real valued function and $q(t)$ is real or complex. If in addition the following conditions hold:

(i) *$p(t) > p(a)$ when $t \in (a, b)$, and the minimum $p(a)$ is only approached at a.*

(ii) *$p'(t)$ and $q(t)$ are continous in a neighborhood of a, except possibly at a.*

(iii) *There are positive real constants P,μ and λ and a real or complex constant Q such that*

$$p(t) - p(a) \sim P(t-a)^{\mu}, \qquad q(t) \sim Q(t-a)^{\lambda-1}, \qquad (t \downarrow a)$$

(iv) *$I(x)$ converges absolutely throughout its range for all sufficiently large x.*

then the following relation holds

$$I(x) \sim \frac{Q}{\mu} \Gamma\left(\frac{\lambda}{\mu}\right) \frac{e^{-xp(a)}}{(Px)^{\lambda/\mu}} \qquad (x \to \infty). \tag{B.5}$$

The $\Gamma(x)$ is the Euler Gamma function equal to $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ for $x > 0$.

For a general continuous function $p(t)$, that does not satisfy condition (i), one can always make the integral over $e^{p(t)}$ meet this condition by splitting the integration domain at the minima and maxima of $p(t)$ and changing the direction of the integration where necessary. Under stronger conditions, this is expressed in the following corrolary.

Corrolary 1 *Consider the integral $I(x) = \int_a^b e^{-xp(t)} dt$, where the limits $a, b \in]\infty, \infty[$ and $p(t) \in C^2(\mathbb{R})$ are independent of x. Assume that*

(i) *$I(x)$ converges for all sufficiently large x. This implies among other things that there is a lower bound p_0 to $p(t)$.*

(ii) *The set $S = \{t_0 | p(t_0) = p_0 \wedge p'(t_0) = 0 \wedge p''(t_0) \neq 0\}$ is discrete and finite.*

(iii) *The lower bound p_0 is approached only at points in S.*

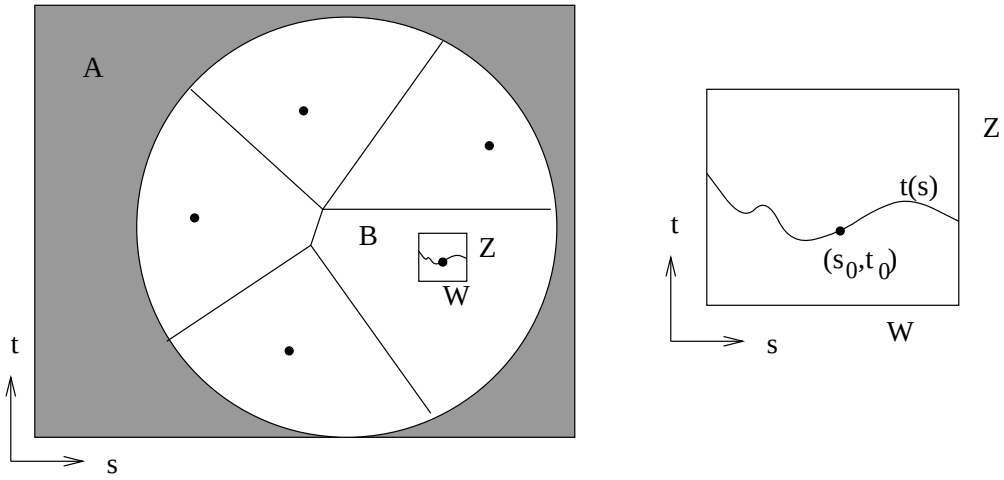


Figure B.2: The structure of the proof of the two-dimensional Laplace method. On the left the division of $\mathbb{R}^2$ into an outer part A and an inner part subdivided into neighborhoods of global minima, is shown. On the right the path $t(s)$ through one of the global minima is shown enlarged.

Then

$$\lim_{x\to\infty} -\frac{1}{x}\log\int_{-\infty}^{\infty} e^{-xp(t)}dt = p_0 \tag{B.6}$$

Proof The elements of S divide the real line. The interval between two elements can be cut into two parts. Each part satisfies the conditions for theorem 2. The integration over this part is therefore asymptoticly equal to a constant times $x^{-1/2}e^{-xp_0}$. The asymptotic relations of all the parts can be added together. Finally, the logarithm and subsequent division by x single out the exponential factor. ■

The extension of this corrolary to more dimensional integrals is not as straightforward as might be suspected at first sight. I succeeded by introducing a strong extra constraint. I think that the next theorem can be proven without this constraint, but for our goal (that is, the usage in appendix C), this formulation of the theorem suffices.

Theorem 3 *Consider the integrals*

$$I(x,s) = \int_{-\infty}^{+\infty} e^{-xp(s,t)}dt, \qquad I(x) = \int_{-\infty}^{+\infty} I(x,s)ds, \tag{B.7}$$

where $p(s,t) \in C^3(\mathbb{R}^2 \to \mathbb{R})$ is independent of x.] Assume that

(i) *For sufficiently large x, $I(x)$ converges and $I(x,s)$ converges for all s. This implies the existence of a lower bound p_0 to $p(s,t)$.*

(ii) *The set $S = \{(s_0,t_0)|p(s_0,t_0) = p_0$ and in (s_0,t_0), $\frac{\partial p}{\partial s} = 0 \wedge \frac{\partial p}{\partial t} = 0 \wedge \frac{\partial^2 p}{\partial s^2} \neq 0 \wedge \frac{\partial^2 p}{\partial t^2} \neq 0\Big\}$ is discrete and its cardinality $|S|$ finite.*

(iii) *The lower bound p_0 is approached only at points in S.*

(iv) *For all ϵ, there is an r such that if $||(s,t)|| > r$ then $p(s,t) > \epsilon$.*

Then

$$\lim_{x\to\infty} -\frac{1}{x}\log I(x) = p_0. \tag{B.8}$$

Proof Choose an $\epsilon > 0$. According to assumption (iv), there will be a region A, see figure (B.2), where $p(s,t) > \epsilon$. Define $A(x) = \int_A ds\, dt\, e^{-x(p(s,t)-\epsilon)}$, then for sufficiently large x, its derivative converges and is given by

$$A'(x) = -\int_A ds\, dt\, (p(s,t) - \epsilon) e^{-x(p(s,t)-\epsilon)} < 0. \tag{B.9}$$

In the limit of x to infinity the contribution of the region A to the integral $I(x)$, will become negligible:

$$\lim_{x\to\infty} \int_A ds\, dt\, e^{-xp(s,t)} = \lim_{x\to\infty} e^{-x\epsilon} A(x) = 0. \tag{B.10}$$

The area enclosed by A can be separated into $|S|$ parts, each a neighborhood of one element of S. Consider the neighborhood B (see figure B.2) of the point (s_0, t_0) of S. By assumption, in the point (s_0, t_0) is $\partial p/\partial t = 0$ and $\partial^2 p/\partial t^2 \neq 0$. The implicit function theorem ensures the existence of a neighborhood U of s_0 and a neighborhood V of t_0 and a unique function $t(s)$ from U to V, such that in $(s, t(s))$, $\partial p/\partial t = 0$ and $\partial^2 p/\partial t^2 \neq 0$. The value p_0 is approached nowhere else in B. Therefore it is possible to find a neighborhood $W \times Z$ of (s_0, t_0), which is a subset of $U \times V$, and a number δ, such that if (s,t) is element of B but not of $W \times Z$, $|p(s,t) - p_0| > \delta$. For every $s \in W$, theorem 2 tells us that

$$\int_Z dt\, e^{-xp(s,t)} \sim e^{-xp(s,t(s))} \left[\frac{x}{2\pi} \frac{\partial^2}{\partial t^2} p(s, t(s))\right]^{-1/2} \qquad (x \to \infty). \tag{B.11}$$

Asymptotic relations can be integrated and theorem 2 can be applied again (where is used that $p \in C^3$).

$$\int_W ds \int_Z dt\, e^{-xp(s,t)} \sim c x^{-1} e^{-xp_0} \qquad (x \to \infty), \tag{B.12}$$

where c is a certain constant.

The work is done, the theorem can now be proven without any more effort. If the elements of S and their corresponding regions B, W and Z are labeled i, with i running from 1 to $|S|$, we can write

$$\begin{aligned}
&\lim_{x\to\infty} -\frac{1}{x} \log \int_{-\infty}^{+\infty} ds \int_{-\infty}^{+\infty} dt e^{-xp(s,t)} = \\
&= \lim_{x\to\infty} -\frac{1}{x} \log \sum_i \left[\int_{B_i \setminus W_i \times Z_i} e^{-xp(s,t)} + \int_{W_i \times Z_i} e^{-xp(s,t)}\right] \\
&= p_0 - \lim_{x\to\infty} \frac{1}{x} \log \sum_i \left[\int_{B_i \setminus W_i \times Z_i} e^{-x(p(s,t)-p_0)} + cx^{-1} + \mathcal{O}(x^{-2})\right] = p_0.
\end{aligned}$$

The second equality is due to the vanishing of the contribution from A. The fourth equality is caused by:

$$\lim_{x\to\infty} \int_{B_i \setminus W_i \times Z_i} ds dt\, e^{-x(p(s,t)-p_0)} < \lim_{x\to\infty} \int_{B_i \setminus W_i \times Z_i} ds dt\, e^{-x\delta} = 0. \tag{B.13}$$

This completes the proof. ■

The proof essentialy uses the implicit function theorem to evaluate the integrals one at the time by ways of the Laplace method. The extension of this theorem to more dimensional integrals is very straightforward.

B.2 Saddle-point Method

Consider now the integral over the path $\mathcal{P}$ in the complex plane

$$I(x) = \int_a^b e^{-xp(t)} dt, \tag{B.14}$$

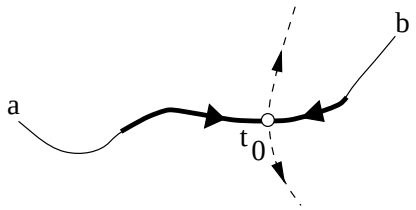


Figure B.3: The steepest descents notion of the saddle-point method. In t_0 lies a saddle-point with the smallest real part of $p(t)$ along the path. The thick path around the saddle-point is defined by having a constant imaginary value. The arrows point downhill.

where $p(t)$ is a complex analytic function and x a large positive real number. One might expect that one can apply Laplace's method directly the real part of $p(t)$, such that the main contribution to the integral is coming from the neighborhood of the point $t = t_0$ where $\Re(p(t))$ takes its minimum on $\mathcal{P}$. The behaviour of the imaginary part of $p(t)$ can destroy the simple argument. One can do some repair work, when $p(t)$ does not posses any poles. In that case one is free to change the path $\mathcal{P}$ in the complex plane. The functions $p(t)$ and $e^{-xp(t)}$ are holomorphic, so by the maximum principle the function $e^{\Re(-xp(t))} = |e^{-xp(t)}|$ has no maxima or minima in the complex plane. This prevents the positioning of the path through a global minimum of $\Re(p(t))$. But let's assume $\Re(p(a)) < \Re(p(b))$ and that the path can be deformed and split into two parts such that at the first part of the new path $\mathcal{P}_1$ the imaginary part of $p(t)$ is a constant

$$\Im(p(t)) = \Im(p(a)) \qquad \text{for } t \in \mathcal{P}_1 \tag{B.15}$$

and that for the second part $\mathcal{P}_2$, there is an ϵ such that $\Re(p(a)) + \epsilon < \Re(p(t))$ for all $t \in \mathcal{P}_2$. For x very large, the main contribution is expected to come from the first part of the path. This part can be evaluated by means of the real Laplace method. The point $t_0 = t$ in which $\Re(p(t))$ takes its minimum is either the border point a or a solution of the equation $p'(t) = 0$. Solutions of the latter equation are saddle-points. The outcome of the integral is equal to (B.3).

The method outlined above is well known under the name of *steepest descents*. The name is coming from a geometrical feature of the paths defined by (B.15). This can be seen if we split $p(t)$ in a real and an imaginary part:

$$p(t) = p_R(t) + ip_I(t). \tag{B.16}$$

While $p(t)$ is an analytic function and has a complex derivative independent of the direction, the function $p_R(t)$ and $p_I(t)$ do not have complex derivatives in general. One can only speak of derivatives of $p_R(t)$ and $p_I(t)$ with respect to a certain path through the complex plane. Let $t(\tau)$ be a path with τ the arc parameter and $|t'(\tau)| = 1$. Let $\rho_R(\tau) \equiv p_R(t(\tau))$ and $\rho_I(\tau) \equiv p_I(t(\tau))$, then we can differentiate (B.16) with respect to τ:

$$p'\left(t(\tau)\right) t'(\tau) = \rho'_R(\tau) + i\rho'_I(\tau). \tag{B.17}$$

The norm of the path derivative of the real part of $p(t)$ is thus linked to the path derivative of the imaginary part

$$\rho'_R(\tau)^2 = |p'\left(t(\tau)\right)|^2 - \rho'_I(\tau)^2. \tag{B.18}$$

A path defined by (B.15) is thus at each point the steepest path in the $\Re(p(t)) = p_R(t)$ landscape. This explains the name *method of steepest descents*. Most of the formal proofs for the Laplace method for contour integrals explicitly use the deforming of the path into a steepest descents path. Olver has proved that this is not really necessary, making *saddle-point method* a more appropriate name. For the estimation of error terms the steepest path are very convenient. If we introduce the notation $(a, b)_{\mathcal{P}}$ to denote the part of $\mathcal{P}$ between a and b with a and b both excluded, then under the following assumptions, Olver was able to prove a formalisation of the saddle point procedure:

(i) *$p(t)$ and $q(t)$ are independent of x, and single valued and holomorphic in a domain T.*

(ii) *The integration path $\mathcal{P}$ is independent of z. The endpoints a and b of $\mathcal{P}$ are finite or infinite, and $(a,b)_{\mathcal{P}}$ lies within T.*

(iii) *$p'(t)$ has a simple zero at an interior point t_0 of $\mathcal{P}$.*

(iv) *$I(x)$ converges at a and b absolutely.*

(v) *$\Re\left(p(t)-p(t_0)\right)$ is positive on $(a,b)_{\mathcal{P}}$, except at t_0, and is bounded away from zero as $t \to a$ or b along $\mathcal{P}$.*

Actually, the theorem and assumptions Olver proved were slightly more general, because he considered x to be complex. As this complicates the assumptions unnecesseraly we have chosen to restrict x to the real line.

Theorem 4 (Saddle point integration [30]) *With the above assumptions,*

$$\int_a^b e^{-xp(t)}dt \sim e^{-xp(t_0)}\left(\frac{2\pi q(t_0)}{p''(t_0)}\right)^{\frac{1}{2}} \qquad (x\to\infty). \tag{B.19}$$

The second derivative $p''(t_0)$ is taken is such a way that the $\omega_0 \equiv$phase of $p''(t_0)$ satisfies

$$|\omega_0 + 2\omega| \le \frac{1}{2}\pi, \tag{B.20}$$

were ω is the limiting value of the phase of $t-t_0$ as $t\to t_0$ along $(t_0,b)_{\mathcal{P}}$.

In the integrals we encounter in this thesis, we start with an integration path over (a part of) the real line. In general, this straight path will not pass a saddle-point. To estimate the integral with the theorem above we have to adjust the path so that it moves through at least one saddle point. The condition that the real part of the function $p(t)$ needs to reach its minimum in the saddle-point, might make it impossible to just travel through one saddle-point. The path then has to pass through perhaps several mountain passes (when we see $\Re(p(t))$ as a mountain range) with identical height. For the function we are considering it is not easy to see if this is the case.

We will not even try to formulate the conditions for a theorem for multi-dimensional saddle-point integrals here, but if we ignore all possible hazards we can hope that in case of the two-dimensional integral

$$\int ds \int dt\, e^{-xp(s,t)},$$

we can first apply the saddle-point method to the t integration. The equation $\partial p/\partial t = 0$ and the condition $\partial^2 p/\partial t^2/not = 0$ yield by the implicit function theorem a function $t(s)$, such that the line $(s,t(s))$ consist of saddle-points of p with respect to t. The saddle-points s_0 of te function with respect to s, are defined by:

$$0 = \frac{d}{ds}p(s,t(s)) = \frac{\partial}{\partial s}p(s,t(s)) + \frac{\partial t}{\partial s}\frac{\partial}{\partial t}p(s,t(s)). \tag{B.21}$$

As the second term vanishes automaticaly by choice of $t(s)$, the conditions for finding the point in which to evaluate the function are simply:

$$\frac{\partial}{\partial s}p(s_0,t_0) = 0, \qquad \frac{\partial}{\partial t}p(s_0,t_0) = 0. \tag{B.22}$$

Extending this sloppy argument to more dimensions is straighforward.

Appendix C

Gaussian Derivation of Free Energy

In chapter 2, we have derived the saddle-point equations for the order parameters $q^{\alpha\beta}$ and m^{α} by first introducing the complex conjugate parameters $\hat{q}^{\alpha\beta}$ and $\hat{m}^{\alpha}$. Doing so we lost the ability to check if a given critical point found by solving the saddle-point equations is exactly a minimum of the free energy. If the partition function could have been written in the form

$$\tilde{Z} \propto \int dq \int dm \exp N\phi(q,m),$$

then the eigenvalues of the Hessian of $\phi(q,m)$ at the critical point would answer the question whether the point is a minimum or a maximum. The ϕ we found was a complex harmonic function of q, $\hat{q}$, m and $\hat{m}$ and therefore has no minima or maxima in the complex plane. In this appendix, the partition function is analyzed without introducing Dirac-delta functions and subsequent use of the complex plane. A prize will be paid in the form of an extra condition on the structure of the network.

C.1 Gaussian Integrals

To avoid the use of any complex numbers we have to start with a positive bias K_{ij} instead of the negative bias A_{ij}. For clarity:

$$\tau \frac{d}{dt} J_{ij} = \frac{1}{N} \langle \sigma_i \sigma_j \rangle_J + \frac{1}{N} K_{ij} - \frac{1}{\mu_{ij}} J_{ij} + \frac{1}{\sqrt{N}} \eta_{ij}(t) \quad \text{for} \quad i < j, \tag{C.1}$$

where $K_{ij} = K_{ji} \geq 0$. No external field is assumed to be present. This leads to the equivalent of (2.10), where the cluster structure has been already put in:

$$\tilde{Z} \;\propto\; \sum_{\vec{\sigma}} \exp \sum_{\kappa,\lambda} \left[\frac{\beta^2}{4N\tilde{\beta}} \mu_{\kappa\lambda} \sum_{\alpha,\beta} \sum_{i \in I_\kappa} \sum_{j \in I_\lambda} \sigma_i^\alpha \sigma_i^\beta \sigma_j^\alpha \sigma_j^\beta + \frac{\beta}{2N} \mu_{\kappa\lambda} K_{\kappa\lambda} \sum_{\alpha} \sum_{i \in I_\kappa} \sum_{j \in I_\lambda} \sigma_i^\alpha \sigma_j^\alpha \right].$$

Instead of the Dirac delta functions we are going to use Gaussian integrals to linearize the terms over which the trace should be taken. To mold the above expression into a form which can be easily linearized this way, we introduce the abbreviations:

$$m_\lambda^\alpha(\vec{\sigma}) \equiv \frac{1}{V} \sum_{i \in I_\lambda} \sigma_i^\alpha, \qquad q_\lambda^{\alpha\beta}(\vec{\sigma}) \equiv \frac{1}{V} \sum_{i \in I_\lambda} \sigma_i^\alpha \sigma_i^\beta, \tag{C.2}$$

and get a partition function where the spin dependencies are all part of these two functions.

$$\tilde{Z} \;\propto\; \sum_{\vec{\sigma}} \exp \sum_{\kappa,\lambda} \left[\frac{\beta^2 N}{4\tilde{\beta}\Lambda^2} \mu_{\kappa\lambda} \sum_{\alpha,\beta} q_\lambda^{\alpha\beta}(\vec{\sigma}) q_\kappa^{\alpha\beta}(\vec{\sigma}) + \frac{\beta N}{2\Lambda^2} \mu_{\kappa\lambda} K_{\kappa\lambda} \sum_{\alpha} m_\lambda^\alpha(\vec{\sigma}) m_\kappa^\alpha(\vec{\sigma}) \right].$$

A generalized Gaussian identity can now be put to practice:

$$\exp\frac{1}{2}\mathbf{b}\cdot\mathbf{A}\mathbf{b} = \sqrt{\det\mathbf{A}}(2\pi)^{N/2}\int d\mathbf{x}\exp-\frac{1}{2}\left[\mathbf{x}\cdot\mathbf{A}\mathbf{x}-\mathbf{b}\left(\mathbf{A}+\mathbf{A}^T\right)\mathbf{x}\right]. \tag{C.3}$$

This identity is only valid if all eigenvalues of A are positive. In that case $\mathbf{A}$ is certainly invertible. The strong extra condition necessary for the proof is therefore:

Condition 1 *The $\Lambda\times\Lambda$ matrices $\mu_{\kappa\lambda}$ and $\mu_{\kappa\lambda}K_{\kappa\lambda}$ are positive definite.*
Note that this implies that $\mu_{\kappa\lambda}$ and $\mu_{\kappa\lambda}K_{\kappa\lambda}$ are invertable.

Under this condition we can use the Gaussian identity and write:

$$\begin{aligned}
\tilde{Z} &\propto \int\left[\prod_{\substack{\alpha<\beta\\ \lambda}} dq_\lambda^{\alpha\beta}\right]\left[\prod_{\substack{\alpha\\ \lambda}} dm_\lambda^\alpha\right]\exp -N\phi(q,m), \qquad \text{(C.4)}\\
\phi(q,m) &= \sum_{\kappa,\lambda}\left[\frac{\beta^2}{2\tilde{\beta}\Lambda^2}\sum_{\alpha<\beta} q_\kappa^{\alpha\beta}\mu_{\kappa\lambda}q_\lambda^{\alpha\beta}+\frac{\beta}{2\Lambda^2}\sum_\alpha m_\kappa^\alpha\mu_{\kappa\lambda}K_{\kappa\lambda}m_\lambda^\alpha\right]-\log B(q,m),\\
B(q,m)^N &= \sum_{\vec{\sigma}}\exp\sum_{\kappa,\lambda}\left[\frac{\beta^2}{\tilde{\beta}\Lambda}\sum_{\alpha<\beta} q_\kappa^{\alpha\beta}\mu_{\kappa\lambda}\sum_{i\in I_\lambda}\sigma_i^\alpha\sigma_i^\beta+\frac{\beta}{\Lambda}\sum_\alpha m_\kappa^\alpha\mu_{\kappa\lambda}K_{\kappa\lambda}\sum_{i\in I_\lambda}\sigma_i^\alpha\right].
\end{aligned}$$

The trace term B can be reduced to a sum over a single replicated spin

$$B(q,m)=\prod_\lambda\left[\sum_{\vec{\sigma_1}}\exp\sum_\kappa\left[\frac{\beta^2}{\tilde{\beta}\Lambda}\sum_{\alpha<\beta} q_\kappa^{\alpha\beta}\mu_{\kappa\lambda}\sigma_1^\alpha\sigma_1^\beta+\frac{\beta}{\Lambda}\sum_\alpha m_\kappa^\alpha\mu_{\kappa\lambda}K_{\kappa\lambda}\sigma_1^\alpha\right]\right]^{1/\Lambda}. \tag{C.5}$$

The partition function has assumed the form of a Laplace transform and can in the thermodynamic limit be calculated exactly by looking at the minima of ϕ. This procedure is called Laplace's method. In appendix B, theorem 3 states precise conditions to which the integral must comply and the result of the calculation. We check the conditions of the (multi-dimensional version of the) theorem one by one. First it is easy to see that $\phi(q,m)$ is C^3 in all its variables. To ensure the convergence of the integral we needed to state a second condition.

Condition 2 *For all clusters, the self-interaction terms of the decay matrix are positive and the self-interaction terms of the bias matrix are non-negative, i.e. $\mu_{\lambda\lambda}>0$ and $K_{\lambda\lambda}\geq 0$ for all λ.*

If this condition is satisfied, we have for ϕ as a function of one specific variable that

$$\begin{aligned}
\phi(q_\lambda^{\alpha\beta}) &= \mathcal{O}((q_\lambda^{\alpha\beta})^2) \qquad q_\lambda^{\alpha\beta}\to\infty,\\
\phi(m_\lambda^\alpha) &= \mathcal{O}((m_\lambda^\alpha)^2) \qquad m_\lambda^\alpha\to\infty.
\end{aligned} \tag{C.6}$$

This settles, along with the convergence issue, the question of meeting the fourth condition of theorem 3. For the elements of the set S, as defined in the conditions of the theorem, we come again to equations for the replica order parameters q and m. By condition 1, we can invert the cluster matrix multiplications and find

$$\begin{aligned}
q_\lambda^{\alpha\beta} &= \left[\sum_{\vec{\sigma_1}}\sigma_1^\alpha\sigma_1^\beta\exp H_\lambda(\vec{\sigma_1})\right]\left[\sum_{\vec{\sigma_1}}\exp H_\lambda(\vec{\sigma_1})\right]^{-1},\\
m_\lambda^\alpha &= \left[\sum_{\vec{\sigma_1}}\sigma_1^\alpha\exp H_\lambda(\vec{\sigma_1})\right]\left[\sum_{\vec{\sigma_1}}\exp H_\lambda(\vec{\sigma_1})\right]^{-1},
\end{aligned} \tag{C.7}$$

where the abbreviations of

$$H_\lambda(\vec{\sigma}) = \frac{\beta^2}{\tilde{\beta}} \sum_{\alpha<\beta} Q_\lambda^{\alpha\beta} \sigma_1^\alpha \sigma_1^\beta + \beta \sum_\alpha M_\lambda^\alpha \sigma_1^\alpha \tag{C.8}$$

and

$$M_\lambda^\alpha \equiv \frac{1}{\Lambda} \sum_\kappa \mu_{\lambda\kappa} K_{\lambda\kappa} m_\kappa^\alpha, \qquad Q_\lambda^{\alpha\beta} \equiv \frac{1}{\Lambda} \sum_\kappa \mu_{\lambda\kappa} q_\kappa^{\alpha\beta} \tag{C.9}$$

were introduced. Solutions of equations (2.16) found with the saddle-point method also solve these equations and vice versa. If one has found a solution of these equations, one should check if it is indeed a global minimum and if the number of solutions which give the same value for ϕ is finite. The check whether one has found the global minimum is the hardest task. See the proof of Elliot Lieb in appendix D for an attempt to solve this task for the replica symmetric solutions of (C.7).

Appendix D

Replica Symmetry Proof for integer n

Lieb[1] was the first to prove that the replica symmetric solution is the solution that minimizes the free energy for integer and positive replica number n. With the arrival of models with finite replica number his proof has found new use. The proof is not written for use in the complex plane and therefore our starting point should be the expression for the free energy derived in appendix C. For this formula to be applicable we remind the reader that the following condition should be satisfied.

Condition 1 *The $\Lambda \times \Lambda$-matrices $\mu_{\kappa\lambda}$ and $\mu_{\kappa\lambda}K_{\kappa\lambda} \equiv -\mu_{\kappa\lambda}A_{\kappa\lambda}$ are positive definite.*

D.1 Replica Symmetry Theorem

Replica symmetry for positive integer n was proven for a one cluster system by considering the following expression for the partition function:

$$Z \propto \int \left[\prod_{\substack{\alpha<\beta \\ \lambda}} dq_\lambda^{\alpha\beta}\right] \left[\sum_{\vec{\sigma}} \exp \frac{\beta}{n} \sum_{\alpha<\beta} \left(q^{\alpha\beta}\sigma^\alpha\sigma^\beta - \frac{1}{2}(q^{\alpha\beta})^2\right)\right]^N . \tag{D.1}$$

In trying to extend the original prove we want the multiple cluster partition function to have a similar structure. The multiple cluster partition function derived in appendix C can be molded into something similar, but it includes a product of cluster related terms.

$$Z \propto \int \left[\prod_{\substack{\alpha<\beta \\ \lambda}} dq_\lambda^{\alpha\beta}\right] \left[\prod_{\substack{\alpha \\ \lambda}} dm_\lambda^\alpha\right] \prod_\lambda P_\lambda(q,m)^{N/\Lambda},$$

where the factors are

$$P_\lambda(q,m) = \sum_{\vec{\sigma}} \exp\left[\gamma \sum_{\alpha<\beta} \left(Q_\lambda^{\alpha\beta}\sigma^\alpha\sigma^\beta - \frac{1}{2}q_\lambda^{\alpha\beta}Q_\lambda^{\alpha\beta}\right) + \beta\sum_\alpha \left(M_\lambda^\alpha\sigma^\alpha - \frac{1}{2}m_\lambda^\alpha M_\lambda^\alpha\right)\right], \tag{D.2}$$

where $\gamma = \beta^2/\tilde{\beta} > 0$. The σ^α is just a single spin here. Although it is not made explicit in the notation, $Q_\lambda^{\alpha\beta}$ and M_λ^α are treated as functions of small q and m. The element $P_\lambda(q,m)$ is therefore not a function of the local parameters q_λ and m_λ only, but depends on the order parameters of all clusters. As we will see in this appendix, this will prevent a simple extension of the prove. However, we will continue with this expression because our other attempts have failed and this expression at least leads to a theorem which proves replica symmetry for the one-cluster case.

[1] His proof was published by Van Hemmen and Palmer in [14]

We would like to prove that $m_\lambda^\alpha = m_\lambda$ and $q_\lambda^{\alpha\beta} = q_\lambda$ for all α, β and for all clusters. The way this was proved by Lieb for an unbiased one-cluster system is not by trying to prove everything at once, but to start singleing out just two replicas. Following him, we decompose $q_\lambda^{\alpha\beta}$ in four parts: q_λ^{12}, $A_\lambda = \{q_\lambda^{1\alpha} | 3 \leq \alpha \leq n\}$, $B_\lambda = \{q_\lambda^{2\alpha} | 3 \leq \alpha \leq n\}$ and $C_\lambda = \{q_\lambda^{\alpha\beta} | 3 \leq \alpha < \beta \leq n\}$. Likewise we decompose m_λ^α in m_λ^1, m_λ^2 and $D_\lambda = \{m_\lambda^\alpha | \alpha \geq 3\}$. We now prove the following lemma.

Lemma 1 *if $n \in \mathbb{N}$, $\beta > 0$ and for all clusters ρ:*

$$\begin{array}{rclcrcl} \vec{q}_\rho &=& (q_\rho^{12}, A_\rho, B_\rho, C_\rho), & & \vec{m}_\rho &=& (m_\rho^1, m_\rho^2, D_\rho), \\ \hat{q}_\rho &=& (q_\rho^{12}, A_\rho, A_\rho, C_\rho), & & \hat{m}_\rho &=& (m_\rho^1, m_\rho^1, D_\rho), \\ \tilde{q}_\rho &=& (q_\rho^{12}, B_\rho, B_\rho, C_\rho), & & \tilde{m}_\rho &=& (m_\rho^2, m_\rho^2, D_\rho) \end{array} \tag{D.3}$$

and $Q_\lambda^{12} \geq 0$ for a certain cluster λ, then

$$P_\lambda(\vec{q}, \vec{m}) \leq \left[P_\lambda(\hat{q}, \hat{m}) P_\lambda(\tilde{q}, \tilde{m}) \right]^{1/2}. \tag{D.4}$$

Further, if $Q_\lambda^{12} > 0$, equality holds in (D.4) if and only if $Q_\lambda^{1\alpha} = Q_\lambda^{2\alpha}$ and $M_\lambda^1 = M_\lambda^2$.

Proof To separate the sum over the first two replicas from the remainder consider

$$K_\lambda(g, h) \equiv \sum_{\sigma^1, \sigma^2 = \pm 1} g(\sigma^1) \exp\left(\gamma Q_\lambda^{12} \sigma^1 \sigma^2\right) h(\sigma^2). \tag{D.5}$$

The center part, $\exp\left(\gamma Q_\lambda^{12} \sigma^1 \sigma^2\right)$, can be seen as the matrix with $\exp \gamma Q_\lambda^{12}$ on the diagonal and $\exp -\gamma Q_\lambda^{12}$ off the diagonal. As the trace and determinant of this 2×2 matrix are non-negative, the matrix is positive semi-definite. By the Cauchy-Schwarz inequality we have

$$|K_\lambda(g, h)|^2 \leq K_\lambda(g, g) K_\lambda(h, h). \tag{D.6}$$

When $Q_\lambda^{12} > 0$ the matrix is positive definite and the inequality above will only be an equality when g and h are linearly dependent.

Letting $X = \left\{\sigma^3, \ldots, \sigma^n\right\}$, we succeed in the separation by defining:

$$\begin{aligned} g_\lambda^1(\sigma^1 | X) &= \exp\left[\gamma \sum_{\alpha \geq 3} \left(Q_\lambda^{1\alpha} \sigma^1 \sigma^\alpha - \frac{1}{2} q_\lambda^{1\alpha} Q_\lambda^{1\alpha} \right) + \beta \left(M_\lambda^1 \sigma^1 - \frac{1}{2} m_\lambda^1 M_\lambda^1 \right) \right], \\ g_\lambda^2(\sigma^2 | X) &= \exp\left[\gamma \sum_{\alpha \geq 3} \left(Q_\lambda^{2\alpha} \sigma^2 \sigma^\alpha - \frac{1}{2} q_\lambda^{2\alpha} Q_\lambda^{2\alpha} \right) + \beta \left(M_\lambda^2 \sigma^2 - \frac{1}{2} m_\lambda^2 M_\lambda^2 \right) \right], \\ V_\lambda X &= \exp\left[\gamma \sum_{3 \leq \alpha < \beta \leq n} \left(Q_\lambda^{\alpha\beta} \sigma^\alpha \sigma^\beta - \frac{1}{2} q_\lambda^{\alpha\beta} Q_\lambda^{\alpha\beta} \right) - \frac{\gamma}{2} q_\lambda^{12} Q_\lambda^{12} + \beta \sum_{\alpha=3} \left(M_\lambda^\alpha \sigma^\alpha - \frac{1}{2} m_\lambda^\alpha M_\lambda^\alpha \right) \right]. \end{aligned} \tag{D.7}$$

We are now able to prove the lemma:

$$\begin{aligned} P_\lambda(\vec{q}, \vec{m}) &= \sum_{\sigma^3, \ldots, \sigma^n} K_\lambda(g_\lambda^1, g_\lambda^2 | X) V_\lambda(X) && \text{(D.8)} \\ &\leq \sum_{\sigma^3, \ldots, \sigma^n} \left[K_\lambda(g_\lambda^1, g_\lambda^1 | X) K_\lambda(g_\lambda^2, g_\lambda^2 | X) \right]^{1/2} V_\lambda(X) \\ &\leq \left[\sum_{\sigma^3, \ldots, \sigma^n} K_\lambda(g_\lambda^1, g_\lambda^1 | X) V_\lambda(X) \right]^{1/2} \left[\sum_{\sigma^3, \ldots, \sigma^n} K_\lambda(g_\lambda^2, g_\lambda^2 | X) V_\lambda(X) \right]^{1/2} \\ &= \left[P_\lambda(\hat{q}) P_\lambda(\tilde{q}) \right]^{1/2}. && \text{(D.9)} \end{aligned}$$

Both inequalities are applications of the Cauchy-Schwarz inequality. If $Q_\lambda^{12} > 0$ the first inequality is strict unless $g_\lambda^1(\sigma|X)$ and $g_\lambda^2(\sigma|X)$ are linearly dependent. This is the case when $g_\lambda^1(+1|X)g_\lambda^2(-1|X) = g_\lambda^1(-1|X)g_\lambda^2(+1|X)$ for all X, which is equivalent to $Q_\lambda^{1\alpha} = Q_\lambda^{2\alpha}$ for all α and $M_\lambda^1 = M_\lambda^2$. Sufficiency is obvious. ■

This lemma provides the core of the next theorem. The theorem shows that to maximize any one of the functions P_λ, one can have replica symmetric Q_λ and M_λ.

Theorem 5 (Local Replica Symmetry Theorem) *If $\beta > 0$ and condition (1) is satisfied, then there are for each cluster λ, values for $q_\rho^{\alpha\beta}$ and m_ρ^α which absolutely maximize the function $P_\lambda(q, m)$ and for which there exist numbers $Q_\lambda \geq 0$ and $M_\lambda \geq 0$, such that $Q_\lambda^{\alpha\beta}(q) = Q_\lambda$ and $M_\lambda^\alpha = M_\lambda$.*

Proof

The $P_\rho(q, m)$'s are smooth functions of q and m. If condition (1) is satisfied, $P_\lambda(q, m)$ will go to zero if $|q^{\alpha\beta}|$ or $|m^\alpha|$ tend to infinity. Therefore $P_\lambda(q, m)$ will have an absolute maximum and by equations (C.7) we see that the maximum will be in the region bounded by $|q_\lambda^{\alpha\beta}| \leq 1$ and $|m_\lambda^\alpha| \leq 1$ for all replicas α, β and all clusters λ.

It is easy to see that $P_\lambda(q, m)$ is invariant under the *global* sign transformation

$$\vec{\mu} \in \{\pm 1\}^n \qquad T_\mu : \begin{cases} q^{\alpha\beta} \to \mu^\alpha \mu^\beta q^{\alpha\beta} \\ m^\alpha \to \mu^\alpha m^\alpha \end{cases} \tag{D.10}$$

Suppose $(q_\rho^{\alpha\beta}, m_\rho^\alpha)$ maximize P_λ. An appropriate sign transformation can be used to make sure that $Q_\lambda^{12} \geq 0$. Let $\hat{q}, \hat{m}, \tilde{q}, \tilde{m}$ be defined as in the lemma. We have $P_\lambda(\vec{q}, \vec{m}) \geq P_\lambda(\hat{q}, \hat{m})$ and $P_\lambda(\vec{q}, \vec{m}) \geq P_\lambda(\tilde{q}, \tilde{m})$. The lemma on the other hand tells us $P_\lambda(\vec{q}, \vec{m}) \leq [P_\lambda(\hat{q}, \hat{m})P_\lambda(\tilde{q}, \tilde{m})]^{1/2}$. We have to conclude that $P_\lambda(\vec{q}, \vec{m}) = P_\lambda(\hat{q}, \hat{m}) = P_\lambda(\tilde{q}, \tilde{m})$. This implies $M_\lambda^1 = M_\lambda^2$ and $Q_\lambda^{1\alpha} = Q_\lambda^{1\beta}$ for all $\alpha \geq 0$.

Here and in the lemma the replicas 1 and 2 are singled out, but because P_λ is symmetric in the replicas the lemma is valid for any two replicas. Now consider the three replicas δ, ϵ, ζ and their associated order parameters $M_\lambda^\delta, M_\lambda^\epsilon, M_\lambda^\zeta$ and $Q_\lambda^{\delta\epsilon}, Q_\lambda^{\delta\zeta}, Q_\lambda^{\epsilon\zeta}$. By an appropriate sign transformation (Q, M) can be placed in one of the four possibilities for $Q_\lambda^{\delta\epsilon}, Q_\lambda^{\delta\zeta}, Q_\lambda^{\epsilon\zeta}$:

1. all three positive
2. two positive, one non-positive
3. one positive, two zero
4. all three zero

For case 1, the application of the lemma yields $Q_\lambda^{\delta\epsilon} = Q_\lambda^{\delta\zeta} = Q_\lambda^{\epsilon\zeta}$ and $M_\lambda^\delta = M_\lambda^\epsilon = M_\lambda^\zeta$. By a replica independent sign transformation, it can be assured that there exists a solution where the M_λ's are all non-negative. Case 2 is ruled out by the lemma. If the lemma is applied to case 3 with one of the zero Q's in the role of Q^{12}, we get a $\hat{Q}$ with one zero Q and two positive Q's and therefore $\hat{Q}$ finds itself in case 2 and cannot maximize P_λ. By the lemma however we have $P_\lambda(\vec{q}, \vec{m}) = P_\lambda(\hat{q}, \hat{m})$ and thus $\vec{q}$ cannot maximize P_λ either. The remaining cases are therefore number 1 and 4. In case 4, where all $Q_\lambda^{\alpha\beta}$ are equal to zero, we conclude from equations (C.7), that $M_\lambda^\alpha M_\lambda^\beta = Q_\lambda^{\alpha\beta} = 0$ for all α, β. This means that M_λ^α is zero for at least all except one of the replicas. The presence of the replica intertwining Q-terms can no longer account for a replica asymmetric maximum and therefore all replica magnetizations should be equal and identical to zero.

Since the above result applies to any three replicas δ, ϵ, ζ, we can conclude that there are, possibly cluster dependent, numbers M_λ and Q_λ for which the n-dimensional vector $M_\lambda^\alpha = M_\lambda$ and the $n(n-1)/2$-dimensional vector $Q_\lambda^{\alpha\beta} = Q_\lambda$ absolutely maximize the functions $P_\lambda(q, m)$. ■

If the system that we are studying contains just one cluster, then this theorem has proved replica symmetry. For the multiple-cluster case, this has been a nice exercise, but has little practical use.

We made a error of judgement when we implicitly assumed that at maximum of the integrand of the partition function is, all integrand factors P_λ are at their maximum. This will not be true, except when the maximum of the integrand is at the origin of the parameter space. We can see that if we examine the extrema of P_λ:

$$0 = \frac{\partial P_\lambda}{\partial q_\rho^{\alpha\beta}} = \sum_{\vec{\sigma}} \gamma \left\{ \frac{1}{\Lambda} \mu_{\rho\lambda} \sigma^\alpha \sigma^\beta - \frac{1}{2} \delta_{\lambda\rho} Q_\lambda^{\alpha\beta} - \frac{1}{2\Lambda} \mu_{\rho\lambda} q_\lambda^{\alpha\beta} \right\} \exp \ldots$$

$$\Rightarrow \frac{1}{\Lambda} \mu_{\rho\lambda} \frac{\sum_{\vec{\sigma}} \sigma^\alpha \sigma^\beta \exp H_\lambda}{\sum_{\vec{\sigma}} \exp H_\lambda} = \frac{1}{2} \delta_{\lambda\rho} Q_\lambda^{\alpha\beta} + \frac{1}{2\Lambda} \mu_{\rho\lambda} q_\lambda^{\alpha\beta}. \tag{D.11}$$

If we sum this last equation over all clusters λ, we indeed find the conditions (C.7) as should be expected as the summed equation is gives the extrema of the entire integrand of the partition function. However, the requirement that all P_λ must be at a maximum restricts the order parameters much more. For a cluster $\rho \neq \lambda$, the substitution of (C.7) in the above equation yields:

$$\mu_{\lambda\rho} q_\lambda^{\alpha\beta} = \frac{1}{2} \mu_{\lambda\rho} q_\lambda^{\alpha\beta} \Rightarrow q_\lambda^{\alpha\beta} = 0 \,\text{or}\, \mu_{\lambda\rho} = 0. \tag{D.12}$$

For systems to be interesting, we do not want to demand either of these two conditions to hold.

We must conclude that theorem 5 has not helped us in proving replica symmetry for a multiple cluster system. For these systems, replica symmetry remains an ansatz.

Bibliography

[1] J.R.L. de Almeida and D.J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glas model. *Journal of Physics A*, 11(5):983–990, 1978.

[2] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

[3] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infite number of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530–1533, 1985.

[4] J. Anemüller. Coupled synaptic and neuronal dynamics for oscillator networks. Master's thesis, King's College London, 1996.

[5] A. C. C. Coolen, R. W. Penny, and D. Sherrington. Coupled dynamics of fast spins and slow interactions. *Physical Review B*, 48(21):16116–16118, December 1993.

[6] A.C.C. Coolen. Langevin equations and spherical spin models. unpublished, 1996.

[7] A.C.C. Coolen. Statistical mechanics of neural networks. Lecture notes, King's College London, 1997.

[8] A.C.C. Coolen and D. Sherrington. Equilibrium distributions of stochastic networks without detailed balance. *Physica A*, 200:602–607, 1993.

[9] V. Dotsenko, S. Franz, and M. Mézard. Partial annealing and overfrustration in disordered systems. *Journal of Physics A*, 27:2351–2365, 1994.

[10] S. F. Edwards and P. W. Anderson. Theory of spin glasses. *Journal of Physics F*, 5:965–974, 1975.

[11] K. H. Fischer and J. A. Hertz. *Spin Glasses*. Cambridge University Press, 1991.

[12] F.R. Gantmacher. *Applications of the theory of matrices*. Interscience, 1959.

[13] D. O. Hebb. *The organization of behaviour*. Wiley, New York, 1949.

[14] J. L. van Hemmen and R. G. Palmer. The replica method and a solvable spin glass model. *Journal of Physics A*, 12(4):563–580, 1979.

[15] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558, 1982.

[16] G. Jongen, A.C.C. Coolen, and D. Bollé. private communication, 1997.

[17] H.J.J. Jonker and A.C.C. Coolen. Unsupervised dynamic learning in layered neural networks. *Journal of Physics A*, 24:4219, 1991.

[18] N.G. van Kampen. *Stochastic processes in physics and chemistry*. North-Holland, 1981.

[19] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Essentials of neural science and behavior*. Prentice Hall, 1995.

[20] E.L. King and C. Altman. *J.Phys.Chem.*, 60:1375, 1956.

[21] S. Kirkpatrick and D. Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384–4403, June 1978.

[22] I. Kondor. Parisi's mean-field solution for spin glasses as an analytic continuation in the replica number. *Journal of Physics A*, 16:L127–L131, 1983.

[23] R. Linsker. From basic network principles to neural architecture. *Proc. Natl. Acad. Sci.*, 83:7508–7512, 8390–8394, 8779–8783, 1986.

[24] D. J. C. MacKay and K. D. Miller. Analysis of Linsker's application of Hebbian rules to linear networks. *Network*, 1:257–298, 1990.

[25] P. C. Mattis. Solvable spin systems with random interactions. *J. Phys. Letters A*, 56:421–422, 1976.

[26] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[27] M. Mezard, G. Parisi, and M. A. Virasoro, editors. *Spin glass theory and beyond*. World Scientific, 1987.

[28] K.D. Miller, J.B. Keller, and M.P. Stryker. Ocular dominance column development: analysis and simulation. *Science*, 245:605–615, 1989.

[29] K. Mogi. Multiple-valued energy function in neural networks with asymmetric connections. *Physical Review E*, 49(5):4616–4626, 1994.

[30] F.W.J. Olver. *Asymptotics and Special Functions*. Academic Press, New York, 1974.

[31] G. Parisi. Toward a mean field theory for spin glasses. *Physics Letters*, 73A(3):203–205, 1979.

[32] G. Parisi. The order paramtere for spin glasses. *Journal of Physics A*, 13:1101–1112, 1980.

[33] G. Parisi. A sequence of approximated solutions to the S-K model for spin glasses. *Journal of Physics A*, 13:L115–L121, 1980.

[34] G. Parisi. On the probabilistic formulation of the replica approach to spin glasses. *cond-mat/9801081*, 1998.

[35] R. W. Penney, A. C. C. Coolen, and D. Sherrington. Coupled dynamics of fast spins and slow interactions in neural networks and spin systems. *Journal of Physics A*, 26:3681–3695, 1993.

[36] R. W. Penney and D. Sherrington. Slow interaction dynamics in spin-glass models. *Journal of Physics A*, 27:4027–4041, 1994.

[37] P. Peretto. *An Introduction to the Modeling of Neural Networks*. Cambridge University Press, 1992.

[38] D. Sherrington. Ising replica magnets. *Journal of Physics A*, 13:637, 1980.

[39] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1792–1796, December 1975.

[40] D. Sherrington, R. W. Penney, and A. C. C. Coolen. Complexity in the coupled dynamics of fast neurons and slow synapses. preprint OUTP-93-25S, Oxford University, 1993.

[41] S. Shinomoto. Memory maintenance in neural networks. *Journal of Physics A*, 20:L1305–L1309, 1987.

[42] D.J. Thouless, P.W. Anderson, and R.G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1976.

[43] T.N. Wiesel and D.H. Hubel. *J. Comp. Neurol.*, 158:307–318, 1974.