University of Helsinki
Department of Computer Science
Series of Publications C, No. C-1994-15

Serge Santos

Phase Transitions in Sparsely Connected Boltzmann Machines

Department of Computer Science
P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, Finland

Helsinki, April 1994

Phase Transitions in Sparsely Connected Boltzmann Machines

Serge Santos

Diploma work
Department of Mathematics and Physics
Eidgenössische Technische Hochschule Zürich
Helsinki, April 1994

Abstract

The primary motivation of this work has been the study of phase transitions in the simulated annealing schedules of Boltzmann machines.

At first, we introduce and focus on the harmonium Boltzmann machine, which can be used to implement a Bayesian reasoning scheme. These sparsely connected Boltzmann machines are composed of two connected clusters of independent nodes, allowing a massive parallel updating scheme.

The study of critical phenomena in simulated annealing presupposes a deeper understanding of Monte Carlo methods, especially in connection with statistical physics. We discuss various aspects of the influence of the correlation between the elements of Monte Carlo samples on the determination of the relevant physical quantities.

The close analogy between Boltzmann machines and spin glass models is brought to the fore in our investigations, by computational simulations, of the thermodynamical properties of the harmonium Boltzmann machine. The study of the specific characteristics of this model contributes to a closer comprehension of the emergence of critical phenomena, corroborated by the results of our simulations.

For the larger framework of sparsely connected Boltzmann machines, we extensively present the theoretical methods of statistical physics applied to strongly diluted spin glasses. Finally, the predicted equilibrium properties are compared with the empirical results.

Acknowledgments

I am greatly indebted to my advisor, Assoc. Prof. Pekka Orponen, for his scientific support and for providing me with good working conditions during my stay at the University of Helsinki. I am grateful to Prof. Dr. Erwin Engeler for the confidence he showed in me by authorising my visit to the University of Helsinki.

I also wish to thank Kari Laasonen for his LaTeX support and informal discussions, and Teemu Leisti for his help in proofreading this work.

Contents

Chapter 1

Introduction

The human nervous system is composed of about 10^{11} neurons of many different types. Despite the relative slowness of the individual neurons, the brain is capable of efficiently handling computationally complicated tasks, such as image processing. Moreover, the brain exhibits robustness and fault tolerance to a degree that would be most desirable in artificial systems. It can deal with incomplete and inconsistent information, and can learn from experience.

The attempt of improving the performance of the computers in complex tasks by means of artificial neural networks has been inspired by the computational power of the brain. An artificial neural network is composed of relatively simple processors connected with each other and processing information collectively and simultaneously. The high connectivity of such networks means that small errors remain mostly inconsequential and that all the computations are performed in a highly parallel manner. The necessary processing information is encoded in the connections between the processing units, called neurons by biological analogy. The computational neuron, however, is not comparable with their biological counterparts.

McCulloch and Pitts (1943) proposed a simple model of the neuron as a binary threshold unit. Specifically, a weighted sum of the inputs from the other units connected to a neuron is compared with a certain threshold value, and the result of the comparison determines the output of this neuron. However, the investigations of Minsky and Papert (1969) gave evidence of the limitations of some neural network models based on such simple processing units.

There is a close analogy between the binary neurons of an artificial network and some simple models of random magnetic materials, the so-called spin glasses, in statistical physics. Hinton and Sejnowski (1983) and Peretto (1984) have formulated the idea of adding new helpful physical insight to artificial networks by introducing a stochastic component into the determining of the output of a neuron, by analogy with the thermal agitation of a physical system. The computational artifact controlling the stochastics of the processing units can be seen to be analogous with the temperature of a physical system.

On the other hand, Metropolis et al. (1953) proposed an algorithm for the efficient simulation of the evolution of a solid to thermal equilibrium. As late as 1982, Kirkpatrick et al. realised the profound analogy between the simulation of the cooling of a solid and optimisation problems. As the temperature of a solid is slowly decreased, the system spontaneously achieves a state of minimal potential energy. In this manner, the simulation of the cooling process can be seen as a solution to a computationally costly optimisation task. Simulated annealing was born.

By and large, the natural idea of joining simulated annealing and artificial networks gives rise to a kind of network called a Boltzmann machine, introduced by Hinton and Sejnowski (1983).

In view of the aforementioned analogy with physics, the understanding of the collective behaviour of artificial networks draws its methods, by means of statistical physics, from the study of systems with a large number of degrees of freedom.

The specific study of spin glasses developed after the middle of the 1970's. A spin glass is a magnetic system composed of magnetic moments, the spins. The interactions between the spins can be considered to be in some frozen-in structural disorder which induces conflicts between the interacting spins. After a freezing transition, these systems exhibit no conventional long-range order, but instead will be in a state that is determined by a kind of order in which the spins are aligned in random directions. The phenomenon of spin ordering were originally observed in the diluted metallic alloy CuMn and in the concentrated insulator $\mathrm{Eu}_x\mathrm{Sr}_{1-x}\mathrm{S}$.

Statistical physics handles macroscopical phenomena occurring in systems with many degrees of freedom. For instance, spin glasses typically exhibits, besides the many spin variables, in the order of 10^{23} random interactions between the magnetic moments. Among the most remarkable macroscopic events in nature are transformations between the various states of a physical system. These phase transitions are characterised by abrupt changes which signify a spontaneous breaking of the symmetry of the system. In the case of decreasing temperature a new phase of lower symmetry, which can consequently be considered of higher order, emerges from a disordered phase.

The purpose of this diploma work is the study of phase transitions in the schedules of simulated annealing of sparsely connected Boltzmann machines. Specifically, the studied networks are generated for the primary purpose of Bayesian reasoning as explained in detail in Chapter 2, where the harmonium Boltzmann machine is also introduced. In Chapter 3, Monte Carlo methods and simulated annealing are discussed. Furthermore, some theoretical foundations of the Monte Carlo method and its applications in statistical physics are presented. In Chapter 4, we study the statistical characteristics of the harmonium Boltzmann machine, embracing some typical behaviours of the networks. Chapter 5 is dedicated to the theoretical derivation of the mean field theory applied to sparsely connected Boltzmann machines. The theoretical predictions on the harmonium Boltzmann machine are verified by means of simulation, and the results are discussed.

Chapter 2

Bayesian reasoning by stochastic neural networks

One of the fundamental problems in the field of expert systems is the incapability of traditional rule-based systems of handling either imprecise, incomplete or inconsistent data. Bayesian reasoning [Pear88], based on axioms of probability theory, offers a solid and unifying framework that has been shown to be well adapted to reasoning that contains degrees of uncertainty. Unfortunately, an efficient implementation of such a computational mechanism, capable of handling uncertain data, seems to require recourse to an approximation algorithm involving some stochastic processes. However, the judicious suggestion [OFMT90] of applying an appropriate mapping from a given Bayesian reasoning framework to a certain class of neural networks, combined with a stochastic method called *simulated annealing*, appears to offer an elegant approach to this problem.

The *Boltzmann machine*, introduced by Hinton and Sejnowski [HiSe83], can be considered as a model for a massively parallel implementation of the simulated annealing algorithm. By providing these particular neural networks with both a stochastic dynamics and an updating scheme, one aims at performing an optimisation task of a well-defined function over all the possible configurations one allows the networks to explore. This function, which can be associated with an energy by analogy with physical systems, is intimately related to the original Bayesian reasoning framework, for which the required stochastic process has now been introduced.

The primary purpose of this chapter is to introduce the fundamental features of the neural networks underlying our further investigation of the method of simulating annealing.

2.1 The Maximum Likelihood Estimation task

Our main interest lies in developing methods for uncertain reasoning using discrete symbolic attributes. One of these methods, *Maximum Likelihood Estimation* (MLE) allows to express the likelihood of attribute-value combinations in terms of numerical probabilities.

Consider a set $\mathbf{X}$ of N discrete random variables, $\mathbf{X} = \{X_1, \ldots, X_N\}$, whose values are denoted by the state vector $\mathbf{x} = \{x_1, \ldots, x_N\}$, where $x_i \in \{0, 1\}$. The set of all possible configuration vectors form the configuration space Ω. We additionally introduce a probability measure $\mathcal{P}$ defined on the entire Ω.

The general problem we have to solve is to determine for a given partial value assignment $\{\mathbf{E} = \mathbf{e}\}$, on a subset $\mathbf{E} \subset \mathbf{X}$, the possible values of the remaining random variables $X_i \notin \mathbf{E}$.

The chosen approach to this problem is the MLE. Let $\mathcal{P}_{max}$ denote the maximal marginal probability compatible with the given value assignment

$$\mathcal{P}_{max} \equiv \max_{\mathbf{x} \in \{\mathbf{E}=\mathbf{e}\}} \mathcal{P}\{\mathbf{x}\}.$$

If we define Ω_{opt} as the set of all configurations $\mathbf{x}_i \in \{\mathbf{E} = \mathbf{e}\}$ satisfying $\mathcal{P}\{\mathbf{x}_i\} = \mathcal{P}_{max}$, the MLE task simply consists of giving one configuration vector $\mathbf{x}_i$ with the probability

$$\mathcal{P}\{\mathbf{x}_i\} = \begin{cases} \frac{1}{|\Omega_{opt}|} \,, & \text{if } \mathbf{x}_i \in \Omega_{opt}; \\ 0 \quad , & \text{otherwise,} \end{cases} \tag{2.1}$$

where $|\Omega_{opt}|$ defines the cardinality of the subspace $\Omega_{opt} \subseteq \Omega$.

Unfortunately, solving this task exactly requires the exploration of the 2^N configurations of Ω, which is infeasible for real applications. Therefore, we shall consider a stochastic simulation method that provides approximate solution to the MLE task.

2.2 The Bayesian network

Intuitively speaking, a Bayesian network [Pear88] can be used to represent any probability distribution by showing the direct causal dependences among the random variables of a problem domain. In such a network, each node corresponds to one of the random variables, and each arc between two of the nodes corresponds to a direct dependence.

Formally, a Bayesian network for a probability distribution $\mathcal{P}$ is defined as a directed acyclic graph $\mathcal{G}$ whose nodes correspond to the variables $\mathbf{U} = \{U_1, \ldots, U_N\}$, and whose topology satisfies the following conditions: first, each variable $U \in \mathbf{U}$ is conditionally independent of all its non-descendants in $\mathcal{G}$, and second, no proper subset of the parent set $\mathbf{F}_U$ of U satisfies the first condition. In this definition, the set of variables $\mathbf{X}$ is said to be conditionally independent of the set of variables $\mathbf{Y}$, given $\mathbf{Z}$, if

$$\mathcal{P}\{\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}\} = \mathcal{P}\{\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}\}$$

holds for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$ satisfying $\mathcal{P}\{\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}\} > 0$.

It has been shown in [Pear88] that the joint probability distribution $\mathcal{P}$ can be represented as the product of the conditional probabilities

$$\mathcal{P}\{\mathbf{U} = \mathbf{u}\} = \prod_{i=1}^{N} \mathcal{P}\{U_i = u_i \mid \bigwedge_{U_j \in \mathbf{F}_{U_j}} U_j = u_j\}. \tag{2.2}$$

For later use, we may introduce the so-called moralisation process that transforms a given Bayesian network by making all the existing arcs undirected, and by adding new undirected arcs between any nodes that have a common child. In such moralised networks, a subset $\mathbf{C} \subseteq \mathbf{U}$ is called a clique if all nodes of $\mathbf{C}$ are connected to each other. As the conditional probability $\mathcal{P}\{U_i = u_i \mid \bigwedge_{U_j \in \mathbf{F}_{U_j}} U_j = u_j\}$ depends only on the variable U_i and its parents, we obtain N cliques, one for each variable U_i.

2.3 The harmony network

Smolensky first introduced in [RuMc86, p. 194–281] his theory of harmony networks. A clever application of his model [OFMT90, MyOr91] aims at transforming a Bayesian network into a two-layer neural network. Each harmony network constitutes a particular kind of Boltzmann machine, on which a stochastic process can be implemented advantageously. Such networks are composed of two clusters of non-connected nodes, which consequently allows a massively parallel updating scheme for the dynamics of the stochastic process. Thus, these networks provide an efficient solution to the MLE task discussed above.

A harmony network consists of two layers of nodes: a set of F *feature nodes* corresponding to the random variables $\{U_1, \ldots, U_F\}$ with $U_i \in \{-1, +1\}$, and a set of P *pattern nodes* $\{A_1, \ldots, A_P\}$ with $A_i \in \{0, 1\}$. The arcs connecting the pattern nodes and the feature nodes are defined by a set of P pattern vectors $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iF})$ where $\omega_{ij} \in \{-1, 0, +1\}$ (see Figure 2.1 on the following page) , and a set of corresponding real-valued parameters λ_i. The sign of ω_{ij} and the value of λ_i determine the strength of the corresponding arc between the feature node U_j and the pattern node A_i.

A major advantage of the harmony network [Myll93] rests on the possibility of a direct mapping from Bayesian networks to harmony networks. We can achieve this identification of the two network structures by introducing a potential function defined as

$$V(\mathbf{u}) \equiv \sum_{i=1}^{P} \lambda_i \chi(\mathbf{u}, \boldsymbol{\omega}_i).$$

The characteristic function χ expresses the consistency of the two vectors $\mathbf{u}$ and $\boldsymbol{\omega}_i$, that is

$$\chi(\mathbf{u}, \boldsymbol{\omega}_i) \equiv \begin{cases} 1, & \text{if } \mathbf{u} \cdot \boldsymbol{\omega}_i = |\boldsymbol{\omega}_i|; \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

where

$$|\boldsymbol{\omega}_i| \equiv \sum_{j=1}^{F} |\omega_{ij}| \tag{2.4}$$

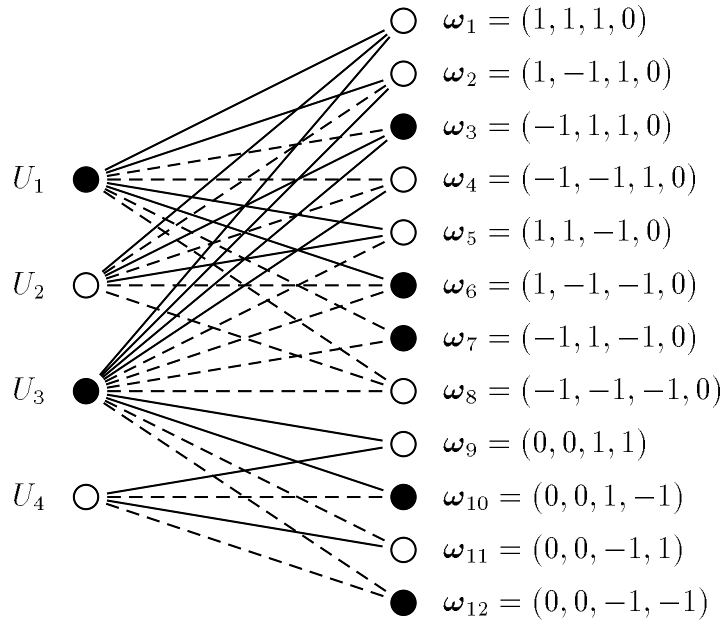


Figure 2.1: A simple harmony network with the corresponding pattern vectors. The current feature vector $\mathbf{u}$ is $(+1, -1, +1, -1)$, and the corresponding pattern activation vector $\mathbf{a}$ is $(0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1)$.

denotes the size of a pattern vector, that also expresses the number of nonzero arcs connecting the pattern node A_i to the feature nodes. The link with the joint probability distribution $\mathcal{P}$ of a given Bayesian network can now be constructed. As shown in Figure 2.2, we identify the parameters λ_i with the conditional probabilities of the causal Bayesian network by the following equation:

$$\lambda_i = \ln \mathcal{P}\{U_i = u_i \mid \bigwedge_{U_j \in \mathbf{F}_{U_j}} U_j = u_j\}.$$

We aim at relating the joint probability $\mathcal{P}$ in (2.2) to the potential function V of the harmony network by

$$\mathcal{P}\{\mathbf{U} = \mathbf{u}\} = \frac{1}{Z} \exp\bigl(V(\mathbf{u})\bigr), \tag{2.5}$$

where $Z = \sum_{\mathbf{u} \in \Omega} \exp\bigl(V(\mathbf{u})\bigr)$ is the normalising function. Considering that the characteristic function $\chi(\mathbf{u}, \boldsymbol{\omega}_i)$ takes the values 1 or 0 according to whether the value assignment $\boldsymbol{\omega}_i$ matches the state vector $\mathbf{u}$ or not, the value of $\exp\bigl(\lambda_i \chi(\mathbf{u}, \boldsymbol{\omega}_i)\bigr)$ will in

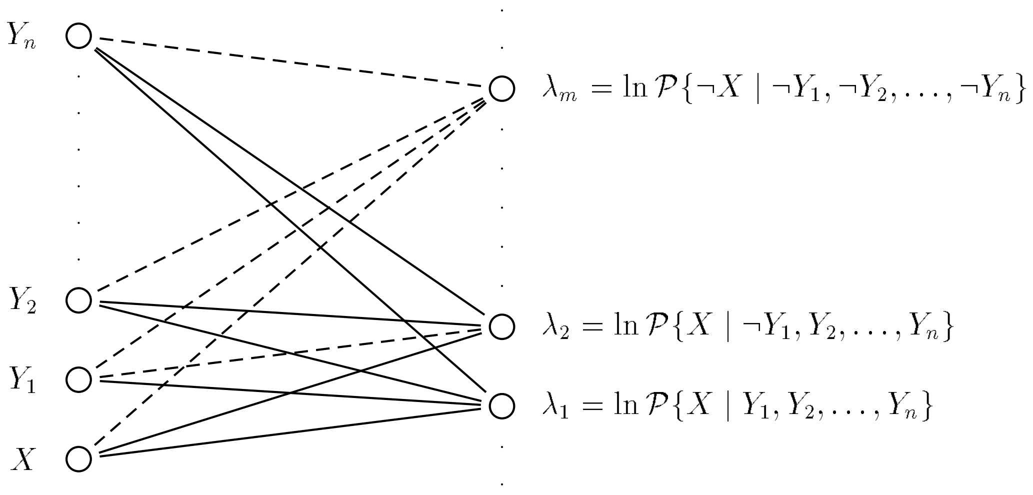


Figure 2.2: The parameters λ_i and the corresponding possible conditional probabilities of a Bayesian network in a clique formed by the nodes $X, Y_1, Y_2, \ldots, Y_n$.

each clique correspond to $\mathcal{P}\{U_i = u_i \mid \bigwedge_{U_j \in \mathbf{F}_{U_j}} U_j = u_j\}$. Thus, by virtue of (2.2), the equation (2.5) holds.

Moreover, we note from (2.5) that finding the maximum of the joint probability $\mathcal{P}$ corresponds to finding the maximum of the potential function V.

At this point, it is meaningful to emphasise that corresponding to a clique of n state variables there are 2^n conditional probabilities. Consequently, the number of pattern nodes of the transformed harmony network increases exponentially with the number of feature nodes as confirmed in Figure 2.2.

2.4 The Boltzmann machine

On many occasions, we have mentioned the possible use of stochastic process for solving the difficult MLE task. The neural network implementation of such dynamical stochastic processes leads naturally to one of the most common of the feedback neural models, the Boltzmann machine [HeKP91, AcHS85].

A *Boltzmann machine* (BM) is a neural network consisting of a set of stochastic binary nodes of states $\{S_1, \ldots, S_n\}$ where $S_i \in \{-1, +1\}$. In BM models, the nodes are connected and all these connections are assumed to be symmetric, i.e., $J_{ij} = J_{ji}$. Because of the symmetrical connections, we can define an energy function mapping the configurations of the nodes to real values as

$$\mathcal{H}(\mathbf{S}) \equiv - \sum_{i=1}^{n} \sum_{j=1}^{n} J_{ij} S_i S_j. \tag{2.6}$$

The nodes of a BM network are updated stochastically according to the following probabilistic rule, usually called Glauber dynamics [Glau63]:

$$P\{S_i = \pm 1\} = \frac{1}{1+\exp(\mp 2\beta h_i)} = \frac{1}{2}\big(1 \pm \tanh(2\beta h_i)\big), \tag{2.7}$$

where the so-called molecular field is defined as

$$h_i \equiv \sum_j J_{ij} S_j. \tag{2.8}$$

We have introduced $\beta = 1/T$ as the control parameter of the random process. However, the stochastic dynamics (2.7) can, in a natural way, be associated with a sort of thermal randomisation in which T plays the role of the temperature.

If we consider two configurations $\mathbf{S}$ and $\mathbf{S}'$, different from each other only by the state S_k, then the energy difference between them can be conveniently rewritten as

$$\Delta\mathcal{H}_k = \mathcal{H}(\mathbf{S}) - \mathcal{H}(\mathbf{S}') = (S_k - S'_k)\sum_{j=1}^{n} J_{kj} S_j = (S_k - S'_k)\, h_k. \tag{2.9}$$

Thus, the probability of flipping of the node S_k becomes

$$P\{S_k \to -S_k\} = \frac{1}{1+\exp(-\beta\Delta\mathcal{H}_k)}. \tag{2.10}$$

We will see in Chapter 3 how stochastic processes satisfying (2.10) can be expected to converge at a fixed temperature T towards an equilibrium. This equilibrium is characterised by the probability $P_{equ}\{\mathbf{S}\}$, called the *Gibbs distribution*, of finding the BM network in the state $\mathbf{S}$ during the stochastic process. The Gibbs distribution is

$$P_{equ}\{\mathbf{S}\} = \frac{1}{Z_T}\exp\big(-\beta\mathcal{H}(\mathbf{S})\big), \tag{2.11}$$

where Z_T is the normalising constant.

The dependence of the Gibbs distribution on T gives us the key to solve the MLE task. If the temperature T decreases slowly towards zero, the Gibbs distribution (2.11) degenerates to the final state distribution (2.1) of the MLE task, where, this time, $\Omega_{opt} \subset \Omega$ corresponds to the subspace minimising the energy function $\mathcal{H}$.

2.5 The Harmonium Boltzmann machine

We have seen in Section 2.3 that the task of maximising the joint probability $\mathcal{P}$ is reducible to the one of maximising the potential function V. In Smolensky's theory of harmony networks, the potential function has been conveniently reformulated to introduce into the harmony network the dynamics of the Boltzmann machine. The resulting neural network model is called the *Harmonium Boltzmann machine* (HBM).

Using a clever trick presented by Smolensky, we can rewrite the characteristic function (2.3) as

$$\chi(\mathbf{u},\boldsymbol{\omega}_i) = \max_{a_i\in\{0,1\}}\left\{ a_i \frac{\frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} - \kappa_i}{1-\kappa_i}\right\} = \begin{cases} 1, & \frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} > \kappa_i; \\ 0, & \frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} \le \kappa_i, \end{cases}$$

where the constants $\kappa_i < 1$ have to satisfy

$$\frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} > \kappa_i \quad \Longleftrightarrow \quad \frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} = 1. \tag{2.12}$$

This condition can be rewritten in terms of the number m_i of consistent components between the state $\mathbf{u}$ and the pattern vector $\boldsymbol{\omega}_i$ as

$$\frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} = \frac{2m_i - |\boldsymbol{\omega}_i|}{|\boldsymbol{\omega}_i|} > \kappa_i \quad \Longleftrightarrow \quad m_i = |\boldsymbol{\omega}_i|.$$

Therefore, in our future use of the harmony network, we can make the following choice:

$$\kappa_i \equiv 1 - \frac{2}{1+|\boldsymbol{\omega}_i|} \; < \; 1, \tag{2.13}$$

ensuring that the condition (2.12) is always met.

Finally the potential function comes down to

$$V(\mathbf{u}) = \sum_i \frac{\lambda_i}{1-\kappa_i} \max_{a_i\in\{0,1\}} \left\{ a_i \left(\frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} - \kappa_i \right) \right\}.$$

The last step to the HBM can now be achieved by introducing the *harmony function* H as

$$H(\mathbf{u},\mathbf{a}) = \sum_i \sigma_i a_i h(\mathbf{u},\boldsymbol{\omega}_i), \tag{2.14}$$

where

$$\sigma_i = \frac{\lambda_i}{1-\kappa_i} \qquad \text{and} \qquad h(\mathbf{u},\boldsymbol{\omega}_i) = \frac{\mathbf{u}\cdot\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} - \kappa_i. \tag{2.15}$$

Assuming that all λ_i are positive, it is easy to see that

$$V(\mathbf{u}) = \max_{\mathbf{a}} H(\mathbf{u},\mathbf{a}). \tag{2.16}$$

This equation implies that the maximum of the potential function V corresponds to the maximum on the entire set of nodes of the harmony network of the harmony function H:

$$\max_{\mathbf{u}} V(\mathbf{u}) = \max_{\mathbf{u}} \max_{\mathbf{a}} H(\mathbf{u},\mathbf{a}).$$

Hence, we have constructed a Boltzmann machine for which H plays the role of the energy $\mathcal{H}$ in (2.6). The interactions between the nodes are, thus,

$$J_{ij} = J_{ji} = \frac{\sigma_i \omega_{ij}}{|\boldsymbol{\omega}_i|}. \tag{2.17}$$

The molecular field (2.8) inducing the stochastic Glauber dynamics (2.7) turns out to be for the i^{th} pattern node

$$I_i^p = \sum_{j=1}^{F} J_{ij} u_j - \sigma_i \kappa_i = (a_i - a_i')\,\Delta H, \tag{2.18}$$

where $\Delta H = H(\mathbf{u}, \mathbf{a}) - H(\mathbf{u}, \mathbf{a}')$ for two pattern vectors $\mathbf{a}$ and $\mathbf{a}'$, differing only by the value of the node $a_i \in \{0, 1\}$. In this case, this leads to the probabilistic rule

$$P\{a_i = 1\} = \frac{1}{1 + \exp(-\beta I_i^p)}.$$

On the other hand, considering the difference of energy, $\Delta H = H(\mathbf{u}, \mathbf{a}) - H(\mathbf{u}', \mathbf{a})$, of two configurations differing only by the value of the node $u_i \in \{-1, +1\}$, the input field of the feature node U_i appears to be

$$I_i^f = \sum_{j=1}^{P} J_{ij} a_j = (u_i - u_i')\,\Delta H, \tag{2.19}$$

taking us back to the dynamics defined in (2.7).

2.6 NEULA: a hybrid neural-symbolic system for Bayesian reasoning

As a practical application of the HBM, a hybrid neural-symbolic expert system shell, the so-called *NEULA system* [OFMT90, MyOr91, OFMT92], has been implemented at the Department of Computer Science of the University of Helsinki. The NEULA system consists of a high-level knowledge representation language, NEULA (NEUral LAnguage); a compiler that is capable of realising the symbolic syntax as a neural network; a neural network simulator; and a user interface.

A NEULA program [Myll93] consists of a hierarchically organised set of concept descriptions. The described concepts consist of references to their immediate ancestors, together with lists of attributes. Associated with each concept's attributes are the distribution of attribute values for objects belonging to this concept's class. These kinds of dependences determine the Bayesian structure incorporating the knowledge given in a NEULA description code. First, the NEULA compiler creates a description of the code as a Bayesian network. Then, the moralised network is transformed into a harmony network as previously discussed. Unfortunately, this procedure still yields highly connected networks. However, at this stage, the compiler exploits the description code in order to leave out a large number of irrelevant nodes arising in the complete model, but describing impossible situations. Moreover, many of the links remain insignificant, mainly because some of the probabilities described in the NEULA code lead to mutually exclusive situations. This way, the number of nodes and arcs of the network can be drastically reduced. After this pruning, we finally obtain a Harmonium Boltzmann machine.

However, an implementation problem still remains. Indeed, as the parameters λ_i in our model are defined as logarithms of probabilities, they are undefined for zero-value probabilities. To avoid this difficulty, these zero-value probabilities are approximated by a small constant $\varepsilon \in [10^{-15}, 10^{-1}]$. One can expect that the accuracy of the results is not significantly affected by this approximation.

Until now, we have disregarded the restricting condition of the positivity of the parameters λ_i. Because of their logarithmic dependence on the conditional probabilities, they appear to be negative. Therefore, we need to find an appropriate rescaling of the λ_i that does not modify the original problem. It has been suggested in [MyOr91, Myll93] that the λ_i could be rescaled according to

$$\lambda_i^R \equiv \lambda_i + \lambda^* \max_i |\lambda_i|, \tag{2.20}$$

where $\lambda^* > 1$ is an arbitrary constant allowing, conceivably, a fine-tuning of the performance of the networks. In order to justify this choice, the resulting rescaled harmony function H^R must be shown to differ from the harmony function H by at most a constant. This happens at low temperatures, where the probability distribution of the configurations approaches the solution of the MLE task. Considering the net input I_i^p to the pattern node a_i,

$$I_i^p = \sum_j J_{ij} u_j - \sigma_i^R \kappa_i = \sigma_i^R \left(\frac{\mathbf{u} \cdot \boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} - \kappa_i \right),$$

the deliberate choice of the parameters κ_i implies that $\mathbf{u} \cdot \boldsymbol{\omega}_i = |\boldsymbol{\omega}_i|$ only if $I_i^P > 0$. This situation corresponds to the perfect matching of the two vectors $\mathbf{u}$ and $\boldsymbol{\omega}_i$, and consequently occurs for exactly one element in each clique. At zero temperature, the update probability $P\{a_i = +1\} = \left(1 + \exp(-\beta I_i^p)\right)^{-1}$ of a_i is determined by the sign of I_i^p to be either 0 or 1. Hence, for sufficiently low temperatures T, the final configuration of the HBM has the same number of pattern nodes in state 1 as the number of cliques. The rescaled harmony function can now be rewritten as

$$H(\mathbf{u}, \mathbf{a}) = H^R(\mathbf{u}, \mathbf{a}) + \lambda^* \max_i |\lambda_i| \sum_{\text{clique}} (1 - \kappa_i)^{-1},$$

showing the judiciousness of the suggested choice.

Chapter 3

Monte Carlo methods

In *Monte Carlo methods* a sequence of states is generated by a stochastic process combined with a freezing scheme on the randomness of this process. An appropriate choice of the *simulated annealing* algorithm for the Monte Carlo method is a prerequisite to ensuring that the stochastic sequence converges asymptotically towards a stationary distribution. We shall discuss the *Gibbs distribution* that results from a suitable stochastic updating rule, as well as the annealing schedule in general.

However, these concepts are based on a strong analogy with the physical annealing process. We will pursue this analogy in the study of the statistical physics of the networks that are subject to simulated annealing. The singular behaviour of such systems during their cooling down is intimately related to a cooperative phenomenon and is reflected in their average properties. Thus, we shall investigate the reliability of the measurement of these average properties.

3.1 Markov processes

3.1.1 Markov chains

The algorithm to be simulated can be modelled mathematically by means of the theory of Markov chains [Fell68]. A *Markov chain* $\{\mathbf{x}(t) \in \Omega \mid t = 0, 1, \ldots\}$, defined on the space Ω of states is the result of a process whose transition probability from the state $\mathbf{x}(t)$ at time t to the state $\mathbf{x}(t+1)$ satisfies the condition

$$P\{\mathbf{x}(t+1) \mid \mathbf{x}(t), \ldots, \mathbf{x}(0)\} = P\{\mathbf{x}(t+1) \mid \mathbf{x}(t)\} \equiv W\big(\mathbf{x}(t) \to \mathbf{x}(t+1)\big).$$

The Markov chain is finite if it is defined on the finite phase space $\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. If the transition probability $W(\mathbf{x}_r \to \mathbf{x}_s)$ is independent of the time t, the Markov chain is said to be homogeneous. For Markov chains, the stationary distribution P_{eq} is defined as

$$P_{eq}(\mathbf{x}) = \lim_{t\to\infty} P\{\mathbf{x}(t) = \mathbf{x} \mid \mathbf{x}(0)\}.$$

If such a stationary distribution exists, we have

$$\lim_{t\to\infty} P\{\mathbf{x}(t)=\mathbf{x}\} = \lim_{t\to\infty} \sum_{l=1}^{N} P\{\mathbf{x}(t)=\mathbf{x} \mid \mathbf{x}(0)=\mathbf{x}_l\} P\{\mathbf{x}(0)=\mathbf{x}_l\} = P_{eq}(\mathbf{x}),$$

following trivially from $\sum_{l=1}^{N} P\{\mathbf{x}(0) = \mathbf{x}_l\} = 1$. Furthermore, it is easy to show [AaKo89, p. 38] that P_{eq} satisfies the *principle of detailed balance*

$$P_{eq}(\mathbf{x}_r)W(\mathbf{x}_r \to \mathbf{x}_s) = P_{eq}(\mathbf{x}_s)W(\mathbf{x}_s \to \mathbf{x}_r), \tag{3.1}$$

which states that, in the stationary case, the number of transitions from the state $\mathbf{x}_r$ to the state $\mathbf{x}_s$ is equal to the number of transitions from $\mathbf{x}_s$ to $\mathbf{x}_r$.

In a stochastic simulation, the total transition probability W is generally determined by both the *generation probability* G_{rs}, i.e., the probability of generating a state $\mathbf{x}_r$ from a state $\mathbf{x}_s$, and the *acceptance probability* A_{rs}, i.e., the probability of accepting such a transition. Then W is defined as

$$W(\mathbf{x}_s \to \mathbf{x}_r) = \begin{cases} G_{rs}A_{rs} & \text{if } r \neq s; \\ 1 - \sum_{l\neq r} G_{ls}A_{ls} & \text{if } r = s. \end{cases}$$

The asymptotic convergence of Markov chains can be proved for a general class of generation and acceptance probabilities. The following theorem states the convergence of a stochastic simulation towards a given distribution $P(\mathbf{x})$.

Theorem 1 *Let $P(\mathbf{x})$ be a probability distribution over Ω, and consider a stochastic simulation process, with acceptance probability A_{rs} and generation probability G_{rs} that fulfil the following requirements for all $r, s \in \{1, \dots, N\}$:*

(G_1) $\exists\{l_0 = r, l_1, \dots, l_p = s\}$ *with* $p \geq 1$ *so that* $G_{l_k l_{k+1}} > 0$ *for* $k = 0, 1, \dots, p-1$;

(G_2) $G_{rs} = G_{sr}$;

(A_1) $A_{rs} = 1$ *if* $P(\mathbf{x}_r) \leq P(\mathbf{x}_s)$, *and otherwise* $A_{rs} \in (0, 1)$;

(A_2) $A_{rl} = A_{rs}A_{sl}$ *for all* $l \in \{1, \dots, N\}$ *with* $P(\mathbf{x}_l) \leq P(\mathbf{x}_s) \leq P(\mathbf{x}_r)$.

Then there exists a stationary distribution P_{eq}. Furthermore $P_{eq}(\mathbf{x}) = P(\mathbf{x})$.

Proof See [AaKo89, p. 42]. □

The simplest alternative for the generation probability satisfying (G_1) and (G_2) is a uniform distribution over the entire phase space Ω.

3.1.2 The Gibbs distribution

From the physical point of view, stochastic processes similar to the ones discussed above are widely observed in nature. The stationary state of a physical system, closed and isolated, but in contact with a thermal bath of temperature T, will depend on the many interactions among its particles. These interactions are described in terms of an energy function $\mathcal{H}$, called in physics a *Hamiltonian* and in optimisation problems a cost function. The *thermodynamic equilibrium* state is the state that maximises the entropy of the physical system (in this particular case, the state that minimises the free energy). The state probability density at the thermodynamic equilibrium, called the *Gibbs distribution*, can be shown [Reic91] to be

$$P_{eq}(\mathbf{x}) = \frac{1}{Z}\exp(-\beta\mathcal{H}(\mathbf{x})) \qquad \text{with} \qquad \beta \equiv (k_B T)^{-1}, \tag{3.2}$$

where the *partition function* Z is a normalisation constant

$$Z \equiv \sum_{\mathbf{x}\in\Omega} \exp(-\beta\mathcal{H}(\mathbf{x})). \tag{3.3}$$

If we want to consider a Markov process leading to such a distribution, the detailed balance principle (3.1) implies that

$$\frac{W(\mathbf{x}_r \to \mathbf{x}_s)}{W(\mathbf{x}_s \to \mathbf{x}_r)} = \exp(-\beta\Delta\mathcal{H}).$$

In other words, the ratio of the transition probability W depends only on the energy difference $\Delta\mathcal{H} \equiv \mathcal{H}(\mathbf{x}_s) - \mathcal{H}(\mathbf{x}_r)$ between two configurations. This equation obviously does not specify the transition probability uniquely, and some arbitrariness in the explicit choice of W remains. One possible alternative fulfilling this requirement is the *Glauber dynamics*

$$W(\mathbf{x}_s \to \mathbf{x}_r) \equiv \frac{1}{2\tau_s}\big(1 + \tanh\frac{\beta}{2}\Delta\mathcal{H}\big) = \frac{1}{\tau_s}\frac{1}{1+\exp(-\beta\Delta\mathcal{H})}, \tag{3.4}$$

where τ_s is an arbitrary factor. For the moment, we can set $\tau_s = 1$, but when we interpret the Monte Carlo process dynamically, we shall choose τ_s to be the unit of Monte Carlo time.

A well-known argument giving evidence of the convergence towards the Gibbs probability distribution P_{eq} consists of considering a large number of Markov chains in parallel. At a given step of the process, there are N_r systems in state $\mathbf{x}_r$ and N_s systems in state $\mathbf{x}_s$. Disregarding the energy change $\Delta\mathcal{H}$, the transition probability is symmetric, i.e., $W_{\Delta\mathcal{H}=0}(\mathbf{x}_r \to \mathbf{x}_s) = W_{\Delta\mathcal{H}=0}(\mathbf{x}_s \to \mathbf{x}_r)$. At this step, the total number $N_{r\to s}$ of transitions from $\mathbf{x}_r$ to $\mathbf{x}_s$ can be written

$$N_{r\to s} = N_r W_{\Delta\mathcal{H}=0}(\mathbf{x}_r \to \mathbf{x}_s)\exp(-\beta\Delta\mathcal{H}).$$

Consequently, the total number of transitions becomes

$$\Delta N_{r\to s} \equiv N_{r\to s} - N_{s\to r} = N_r W_{\Delta\mathcal{H}=0}(\mathbf{x}_r \to \mathbf{x}_s)\left(\frac{\exp\bigl(-\beta\mathcal{H}(\mathbf{x}_s)\bigr)}{\exp\bigl(-\beta\mathcal{H}(\mathbf{x}_r)\bigr)} - \frac{N_s}{N_r}\right).$$

Thus, as long as the ratio N_s/N_r remains smaller than the Gibbs distribution P_{eq}, we have $\Delta N_{r\to s} > 0$. This means that N_s/N_r increases. On the other hand, the same argument applies to the case $N_s/N_r > P_{eq}$, and in this case, N_s/N_r decreases. In this manner, it is certain that the process converges towards configurations distributed according to the Gibbs distribution.

However, the correctness of this argument is restricted to the case of an update scheme involving only a few configurations at a time. Indeed, the simultaneous updating of $N \gg 1$ degrees of freedom leads to a total change of energy $\Delta\mathcal{H}$ of order $\beta N'\varepsilon$, where ε sets the energy scale for the updating of one degree of freedom. The temperatures of interest are such that $\beta\varepsilon$ is of order unity. Hence, for nearly every transition step with $N' \sim N \gg 1$, the Glauber dynamics provides only an extremely small transition probability, and it follows that the system would stay in its previous configuration most of the time.

3.2 Simulated annealing

According to the Gibbs distribution (3.2), the lowest energy states are favoured at equilibrium. The Gibbs distribution dependence on the temperature induces the configurations of low energy to be macroscopically populated at low temperature. In the limit $T \to 0$, solely the configurations of minimal energy can be occupied; this is the *ground state* of the system. However, such a final distribution can only be achieved by cooling the system down carefully keeping it in equilibrium at each step. This process is called *simulated annealing*.

3.2.1 The cooling schedule

In practice, we are not able to run an algorithm for infinitely long at infinitely many temperatures in order to ensure a Gibbs distribution at the end of the annealing. Therefore, we must define a set of rules that determine how the temperature is to be decreased in a finite-time simulation. This set of rules, or the *cooling schedule* specifies:

- the initial value of the temperature,
- the final value of the temperature (a convergence criterion),
- a decrement rule for the temperature,
- the number of iterations to be performed at each stage of the cooling process (a quasi-equilibrium criterion).

The simulated annealing process can be regarded as a sequence of homogeneous Markov chains at each temperature stage. In this context, a relevant theorem due to Geman and Geman [GeGe84] provides a lower bound for the decrement rule for the temperature, guaranteeing convergence in a finite number of steps.

Theorem 2 *Consider a Markov process of state $\mathbf{x}(t)$, at time t, of N components, resulting in the case where, at each iteration step t, only one component of index $I(t) \in \{1, \ldots, N\}$ is updated. Then the global minimum of the energy function is achieved with probability 1 if the following requirements are fulfilled for every $t > 1$:*

1. *there exists an integer $\tau \geq N$ so that $\{1, \ldots, N\} \subseteq \{I(t+1), \ldots, I(t+\tau)\}$;*
2. *$T(t) \to 0$ as $t \to \infty$, where $T(t)$ is the temperature at the time t;*
3. *$T(t) \geq \Gamma / \ln t$, where $\Gamma \equiv N\Delta E$ with $\Delta E \equiv E_{max} - E_{min}$.*

Proof See [GeGe84] $\square$

It is instructive to look informally at the reason for this lower bound. If we consider the case where the temperature $T(t) = \Gamma/\ln t$, and where, at this temperature, all N components are updated according to Glauber dynamics, then the probability of moving to any state in iteration t will be at least

$$\prod^{N} \frac{1}{1+\exp(\beta\Delta E)} = \left(1 + t^{\frac{\Delta E}{\Gamma}}\right)^{-N},$$

where $\beta(t) \equiv T'(t)^{-1}$ with the rescaled temperature $T' = k_\beta T$. As the number of iterations goes to infinity, the number of times that any particular state will be visited is at least $\sum_t \left(1 + t^{\Delta E/\Gamma}\right)^{-N}$. The time spent in any state becomes infinite if we choose $\Gamma \geq N\Delta E$, ensuring in this way that it is impossible for the system to be trapped in a bad local minimum of the energy. As the temperature decreases, the system will spend an increasing fraction of the time in the global minimum (or minima) of the energy.

From the implementational point of view, this schedule has two major drawbacks. First, although the process is finite, the logarithmic decrement of the temperature remains extremely slow and renders the simulated annealing impracticable. Second, one resorts to a conservative estimation of the value of Γ [LaAa87] for which no estimations of ΔE are even available. Nonetheless, an interesting approach [ChHo93] to the determination of an optimal cooling schedule shows that, for small systems, a logarithmic time decrement is both necessary and sufficient. Indeed, a system with a very small number of distinct states has a sufficiently small number of possible transitions so that it is possible to compute numerically the entire transition matrix of the Markov process. By this method, the optimal annealing schedule, minimising the mean final energy, can be determined numerically. The obtained schedule behaves logarithmically as proposed by Geman and Geman. In addition, the effects of large energy barriers can be explicitly investigated.

3.2.2 Quasi-equilibrium

In the construction of cooling schedules, the concept of *quasi-equilibrium* plays a central role. Let us define L_k as the length of the k^{th} Markov chain at the corresponding temperature T_k. The simulated annealing algorithm is said to achieve a quasi-equilibrium if the probability distribution $\mathbf{P} = \{P(\mathbf{x})\}_{\mathbf{x}\in\Omega}$ of the system, after L_k transitions, is sufficiently close to the stationary distribution $\mathbf{P}_{eq}(T_k) = \{P_{eq}(\mathbf{x})\}_{\mathbf{x}\in\Omega}$ at temperature T_k. Formally, this requires

$$\|\mathbf{P}(L_k, T_k) - \mathbf{P}_{\text{eq}}(T_k)\| < \varepsilon$$

for some specified positive value of ε. Therefore, for small steps in the decrement of the temperature T_k, the subsequent stationary distributions $\mathbf{P}(L_k, T_k)$ of the homogeneous Markov chains remain close to each other. Intuitively, it is clear that a large decrement of the temperature T_k makes necessary a larger number L_{k+1} of transitions at T_{k+1} to restore the quasi-equilibrium. There is a trade-off between a fast decrement of T_k and small values of L_k, and one usually opts for small decrements of the temperature to avoid extremely long Markov chains. In this case, a small number of transitions is likely to be sufficient to restore the quasi-equilibrium prevailing at the previous temperature T_k. This condition of quasi-equilibrium can be formulated as

$$\|\mathbf{P}(T_k) - \mathbf{P}(T_{k+1})\| < \varepsilon \quad \text{for } k = 0, 1, \dots\ . \tag{3.5}$$

In this manner, the quasi-equilibrium is maintained during the entire cooling process. Evidently, this is based on the assumption that the initial probability distribution at T_0 is close to the stationary distribution $\mathbf{P}_{eq}(T_0)$.

It is now possible to find a sufficient condition for the decrement rate of the temperature ensuring quasi-equilibrium by the following two steps. First, there exists a $\delta(\varepsilon) > 0$ so that equation (3.5) follows from the inequalities

$$\frac{1}{1+\delta} < \frac{P_l(T_k)}{P_l(T_{k+1})} < 1 + \delta \quad \text{for } k = 0, 1, \dots, \tag{3.6}$$

for all components $P_l \equiv P(\mathbf{x}_l)$ of the distribution $\mathbf{P}$. Second, the desired sufficient condition for the decrement rule of the temperature is provided by the following theorem:

Theorem 3 *Let $\mathbf{P}(T_k)$ be the stationary distribution of the homogeneous Markov chain of length L_k, and let two successive values of the temperature of the cooling schedule satisfy $T_{k+1} \leq T_k$. Then the quasi-equilibrium condition (3.6) holds if the following requirement is fulfilled for all $\mathbf{x}_l \in \Omega$:*

$$\frac{\exp(-\beta_k \Delta\mathcal{H}_l)}{\exp(-\beta_{k+1}\Delta\mathcal{H}_l)} < 1 + \delta \quad \textit{for } k = 0, 1, \dots, \tag{3.7}$$

where $\beta_k = T_k^{-1}$ and $\Delta\mathcal{H}_l \equiv \mathcal{H}(\mathbf{x}_l) - \min_{\Omega} \mathcal{H}$.

Proof Provided we consider Glauber dynamics, the components of the stationary distribution at temperature T_k can be rewritten as

$$P_l(T_k) = \frac{1}{Z_0(T_k)} \exp(-\beta_k \Delta\mathcal{H}_l),$$

where the normalisation factor Z_0 is not the partition function Z, but

$$Z_0(T_k) = \sum_{\mathbf{x}_l \in \Omega} \exp(-\beta_k \Delta\mathcal{H}_l).$$

If $T_{k+1} \leq T_k$, it follows from equation (3.7) that

$$\begin{aligned} P_l(T_k) = \frac{1}{Z_0(T_k)} \exp(-\beta_k \Delta\mathcal{H}_l) &\leq \frac{1}{Z_0(T_{k+1})} \exp(-\beta_k \Delta\mathcal{H}_l) \\ &< (1+\delta)\frac{1}{Z_0(T_{k+1})} \exp(-\beta_{k+1} \Delta\mathcal{H}_l) = (1+\delta) P_l(T_{k+1}) \end{aligned}$$

for all configurations $\mathbf{x}_l \in \Omega$.

On the other hand, equation (3.7) also implies that

$$\begin{aligned} P_l(T_k) &\geq \frac{1}{Z_0(T_k)} \exp(-\beta_{k+1} \Delta\mathcal{H}_l) \\ &> \frac{\exp(-\beta_{k+1}\Delta\mathcal{H}_l)}{\sum_{\mathbf{x}_m \in \Omega} (1+\delta)\exp(-\beta_{k+1}\Delta\mathcal{H}_m)} = \frac{1}{1+\delta} P_l(T_{k+1}), \end{aligned}$$

which completes the proof. □

Equation (3.7) can be conveniently rewritten to yield the following lower bound for T_k:

$$T_{k+1} > \frac{T_k}{1 + T_k \frac{\ln(1+\delta)}{\Delta\mathcal{H}_l}}. \tag{3.8}$$

This way, the frequently used decrement rule $T_{k+1} = \alpha T_k$, where $\alpha < 1$, can be shown to be clearly too steep at low temperatures. Indeed, by inserting this annealing schedule into equation (3.8), for each $\mathbf{x}_l \in \Omega$, we obtain

$$T_k > \frac{1-\alpha}{\alpha} \frac{\Delta\mathcal{H}_l}{\ln(1+\delta)}. \tag{3.9}$$

As can be seen, the proposed lower bound in Theorem 2 also confirms this result. However, it is instructive to note that this expression contains an implicit reference, through the term $\Delta\mathcal{H}_l$, to the size of the space to be explored by the Markov chain. Consequently, if exterior constraints, such as the local fields in a Boltzmann machine, limit the set of configurations to be explored by making some configurations totally disadvantageous from the energetical point of view, then the lower bound for maintaining quasi-equilibrium is lowered. In such a case, the apparently rapid annealing schedule $T_{k+1} = \alpha T_k$, whose reliability has been noted for harmony networks in [Myll93], constitutes a possibly appropriate choice.

3.3 Study of phase transitions and critical phenomena

3.3.1 Statistical mechanics

Statistical physics studies the relationship between the microscopical and macroscopical descriptions of physical systems. A physical system can exhibit a singular behaviour, called a *phase transition*. These abrupt changes or discontinuities are caused by a cooperative phenomenon among the many constituents of the system. It has been shown that if a phase transition occurs, it occurs at a well-defined temperature (Lee and Yang (1952), see [Reic91]), and it coincides with a long-ranged reorganisation phenomenon among the particles. The phase transitions are classified according to the singular behaviour of the *free energy* of the system [Mour84]. If we consider a closed and isolated system of Hamiltonian $\mathcal{H}$ in thermodynamical equilibrium with a thermal bath of temperature T, we know that the probability of finding the system in a given state obeys the Gibbs distribution. Therefore, the expectation value for the energy of a system with a continuous set of states $\mathbf{x}$ is

$$\langle\mathcal{H}\rangle_{\mathrm{T}} \equiv \frac{1}{Z}\int \mathrm{d}\mathbf{x}\,\mathcal{H}(\mathbf{x})\exp\bigl(-\beta\mathcal{H}(\mathbf{x})\bigr), \tag{3.10}$$

where Z denotes the partition function and $\beta \equiv (k_\beta T)^{-1}$. Equation (3.10) defines the *ensemble average* $\langle\,\cdot\,\rangle_{\mathrm{T}}$ which constitutes the formal connection between the microscopical and the macroscopical physical worlds. The free energy F, defined as

$$F \equiv -k_\beta T \ln Z, \tag{3.11}$$

appears to be convenient in expressing the *heat capacity* C, itself defined as

$$C \equiv \left(\frac{\partial\langle\mathcal{H}\rangle_{\mathrm{T}}}{\partial T}\right) = -T\left(\frac{\partial^2 F}{\partial T^2}\right) = \frac{\langle\mathcal{H}^2\rangle_{\mathrm{T}} - \langle\mathcal{H}\rangle_{\mathrm{T}}^2}{k_b T^2}. \tag{3.12}$$

The last term follows directly from the calculation of the derivative of the energy $\langle\mathcal{H}\rangle_{\mathrm{T}}$. Moreover, the heat capacity can be expressed in terms of a derivative of the free energy F. Because a phase transition is characterised by a singular behaviour of the free energy, the fluctuation quantity C also exhibits a singular behaviour. This is the reason we shall investigate the method of computing the desired average properties, taking advantage of the running Monte Carlo process. In this manner, we shall be able to determine experimentally the temperatures of interest of a given system.

3.3.2 Monte Carlo sampling method

The Monte Carlo sampling method is based on the idea of approximating the integration (3.10) with the Gibbs configuration weight over the entire configuration space by performing an integration on a characteristic subset of states that constitute a statistical sample. By and large, the limit distribution of a Markov chain $\{\mathbf{x}_l\}_{l\geq 1}$ generated by the Monte Carlo method is the Gibbs distribution itself. Thus, a subset of this Markov chain appears to be an appropriate choice for the statistical sample. This simple arithmetical average, a sort of coarse-grained average, is defined for an observable A quantity as

$$\overline{A} \equiv \frac{1}{(M-M_0)} \sum_{m=M_0+1}^{M} A(\mathbf{x}_m) \cong \langle A \rangle_{\mathrm{T}} \equiv \frac{1}{Z} \int \mathrm{d}\mathbf{x}\, A(\mathbf{x}) \exp\bigl(-\beta\mathcal{H}(\mathbf{x})\bigr). \tag{3.13}$$

In principle, the limit distribution of the Markov chain is independent of the initial state, but in practice, only a finite Markov chain can be generated. Hence, the correlations between the configurations generated sequentially by the Monte Carlo process will reflect the choice of the initial configuration.

We shall now attempt to understand how these correlations may affect the accuracy obtained by the sampling method. This task requires interpreting the Monte Carlo process dynamically. Therefore, we shall associate a time t with the scale of the subsequent configuration set of the Markov chain, and consider the probability $P(\mathbf{x}_l, t) = P(\mathbf{x}_l)$ that a configuration $\mathbf{x}_l$ occurs at time t. The normalisation of the time scale can be such that $N\tau_s^{-1}$ single component transitions are performed in a time unit for a system of N degrees of freedom. Thus, the time t and the Monte Carlo step l are related by $t = \tau_s l/N$. In this case, the probability $P(\mathbf{x}, t)$ satisfies [Reic91, p. 157–163] the *master equation*

$$\frac{\mathrm{d}P(\mathbf{x},t)}{\mathrm{d}t} = -\sum_{\mathbf{x}'} W(\mathbf{x} \to \mathbf{x}')P(\mathbf{x},t) + \sum_{\mathbf{x}'} W(\mathbf{x}' \to \mathbf{x})P(\mathbf{x}',t).$$

In thermodynamical equilibrium, the principle of detailed balance ensures that the two terms of the sum cancel each other, and hence for $P(\mathbf{x}, t) = P_{eq}(\mathbf{x})$, we have $\mathrm{d}P(\mathbf{x}, t)/\mathrm{d}t = 0$. The stationary distribution of the master equation is seen to be $P_{eq}(\mathbf{x})$, as expected.

In this context, the sampling average previously introduced can simply be interpreted as a time average along a stochastic trajectory in phase space controlled by the master equation. Thus, we obtain

$$\overline{A} = \frac{1}{t_M - t_{M_0}} \int_{t_{M_0}}^{t_M} \mathrm{d}t\, A(t), \tag{3.14}$$

where t_M and t_{M_0} are the times elapsed after M and M_0 configurations, respectively, have been generated.

3.3.3 Statistical errors

The dynamic interpretation [BiHe88] proves to be powerful in understanding the statistical errors of the Monte Carlo sampling method.

Let us consider $n \gg 1$ successive observations of a quantity A during a Monte Carlo process. The statistical error on the values $\{A_\mu\}_{\mu=1,\ldots,n}$ is given as

$$\delta A \equiv \frac{1}{n}\sum_{\mu=1}^{n}(A_\mu - \langle A\rangle_{\mathrm{T}}),$$

and the expectation value of its square becomes

$$\big\langle(\delta A)^2\big\rangle_{\mathrm{T}} = \frac{1}{n^2}\sum_{\mu=1}^{n}\big\langle(A_\mu - \langle A\rangle_{\mathrm{T}})^2\big\rangle_{\mathrm{T}} + \frac{2}{n^2}\sum_{\mu_1=1}^{n}\sum_{\mu_2=\mu_1+1}^{n}\big(\langle A_{\mu_1}A_{\mu_2}\rangle_{\mathrm{T}} - \langle A\rangle_{\mathrm{T}}^2\big).$$

The second term explicitly contains the correlations between the sequential observations of the quantity A.

We shall now denote by δt the time interval between two successive observations A_μ and $A_{\mu+1}$. Hence the time associated with the Monte Carlo process is $t_\mu = \mu\,\delta t$. By changing the summation index and interpreting the Monte Carlo process dynamically, we obtain

$$\big\langle(\delta A)^2\big\rangle_{\mathrm{T}} = \frac{1}{n}\big(\langle A^2\rangle_{\mathrm{T}} - \langle A\rangle_{\mathrm{T}}^2\big)\left(1 + \frac{2}{\delta t}\int_0^{t_n} \mathrm{d}t\,\left(1 - \frac{t}{t_n}\right)\Phi_A(t)\right), \tag{3.15}$$

where $\Phi_A(t)$ denotes the normalised relaxation function of the quantity A, given by

$$\Phi_A(t) \equiv \frac{\big\langle A(0)A(t)\big\rangle_{\mathrm{T}} - \langle A\rangle_{\mathrm{T}}^2}{\langle A^2\rangle_{\mathrm{T}} - \langle A\rangle_{\mathrm{T}}^2}.$$

We note that $\Phi_A(0) = 1$, and since $A(\infty)$ is independent of the initial observation $A(0)$, $\lim_{t\to\infty}\Phi_A(t) = 0$. We shall now assume that $\Phi_A(t)$ has essentially decayed to zero on a time scale τ_A with $\tau_A \ll t_n$, which means that the observation time $\tau_{obs} = t_n$ has been chosen to be long enough to avoid significant correlation with the initial state of the system. We can give a precise meaning to τ_A in terms of the integral

$$\tau_A \equiv \int_0^{\infty}\mathrm{d}t\;\Phi_A(t).$$

Since we have assumed that $\Phi_A(t)$ differs appreciably from zero only for times $t \ll t_n$, the term t/t_n in equation (3.15) can be neglected, and the integration can be extended to infinity. Finally, equation (3.15) yields

$$\big\langle(\delta A)^2\big\rangle_{\mathrm{T}} = \frac{1}{n}\big(\langle A^2\rangle_{\mathrm{T}} - \langle A\rangle_{\mathrm{T}}^2\big)\big(1 + 2\frac{\tau_A}{\delta t}\big). \tag{3.16}$$

There are two cases to be distinguished. First, the case $\delta t \gg \tau_A$, signifying, by comparison to the decay time of the relaxation function, a large time interval between two observations, leads to

$$\left\langle (\delta A)^2 \right\rangle_{\mathrm{T}} = \frac{1}{n}\left(\langle A^2 \rangle_{\mathrm{T}} - \langle A \rangle_{\mathrm{T}}^2 \right).$$

Second, the choice of a small observation step $\delta t \ll \tau_A$ implies

$$\langle (\delta A)^2 \rangle_{\mathrm{T}} = 2\frac{\tau_A}{\tau_{obs}}\left(\langle A^2 \rangle_{\mathrm{T}} - \langle A \rangle_{\mathrm{T}}^2 \right),$$

where $\tau_{obs} = n\ \delta t$ is the time over which the averaging is extended. These formulas make explicit that the statistical error is completely independent of the time interval δt between two observations. Consequently, despite a choice of a smaller value δt for a given averaging time τ_{obs}, the statistical error does not decrease at all.

3.3.4 Finite-size behaviour of the statistical error

We shall now extend our considerations to the study of systems of finite size. Consider [MiBH86] a system of linear size L, and n statistically independent observations of a quantity A. The question of the dependency of

$$\Delta(n, L) \equiv \sqrt{\frac{\langle A^2 \rangle_L - \langle A \rangle_L^2}{n}} \tag{3.17}$$

on the size of the system, where $\langle\ \cdot\ \rangle_L$ denotes the ensemble average on a system of linear size L, arises naturally. If $\Delta(n, L)$, in the limit $L \to \infty$, does not go to zero and reaches a value independent of L, then A is said to exhibit a lack of *self-averaging*. On the other hand, if $\Delta(n, L)$ goes to zero, it is worth investigating whether one obtains the same statistical error $\Delta(n, L)$ by studying larger systems but reducing the number n of statistically independent observations, so that the total cost in computing time remains the same. This question can be formulated better by properly introducing the notion of self-averaging.

A system is said to be self-averaging if the following fluctuation satisfies

$$\left\langle (A - \langle A \rangle_L)^2 \right\rangle_L = \langle A^2 \rangle_L - \langle A \rangle_L^2 \propto L^{-m},$$

where $0 < m \leq d$, with d denoting the dimensionality of the system. If $m = d$, the system is considered to be *strongly self-averaging*.

The next step consists of introducing a scale factor $b > 1$ for a rescaled system of linear size $L' = bL$, for which $n' = b^{-d}n$ observations have been recorded. According to equation (3.17), the behaviour of the statistical error of a self-averaging quantity A is $\lim\limits_{L\to\infty} \Delta(n, L) = 0$. Furthermore, a strongly self-averaging quantity A satisfies the stronger condition

$$\Delta(n, L) = \Delta(n', L') = \Delta(b^{-d}n, bL).$$

Indeed, this means that the computational cost remains independent of the scale factor b.

If we now suppose that the quantity A observed in a simulation is the density of an extensive quantity, e.g., the per-site energy or the per-site magnetisation, then, away from any criticality, the Einstein fluctuation theory [Reic91] predicts the probability distribution for the fluctuations of A about its average $\langle A\rangle_{\mathrm{T}}$. In the limit $L \to \infty$, the distribution of $\delta A \equiv A - \langle A\rangle$ is a Gaussian and can be written as

$$P_L(\delta A) = (2\pi C_A)^{-\frac{1}{2}} L^{\frac{d}{2}} \exp\left(\frac{L^d}{2C_A}(\delta A)^2\right). \tag{3.18}$$

Therefore, δA is related to the parameter C_A which determines the width of the fluctuation, and we obtain

$$\left\langle(\delta A)^2\right\rangle_L = \left\langle A^2\right\rangle_L - \left\langle A\right\rangle_L^2 = L^{-d} C_A. \tag{3.19}$$

For instance, in the case of the per-site energy $E \equiv \langle\mathcal{H}\rangle_{\mathrm{T}}/N$, we have $C_E = k_\beta TC/N$, where C is the heat capacity introduced in equation (3.12). The limits of such intensive quantities exist, i.e., $\lim\limits_{L\to\infty} C_A$ exists, and it follows that A is strongly self-averaging. This important property holds for quantities such as E and the per-site magnetisation M.

However, the situation becomes drastically different if the sample is taken on quantities that are themselves fluctuations of densities, like the per-site specific heat C/N or the *susceptibility* $\chi \equiv \partial M/\partial h$. The error $\Delta(n, L)$ of the quantity C_A can be written, using equation (3.19), as

$$\Delta(n, L) = \sqrt{\frac{\left\langle C_A^2\right\rangle_L - \left\langle C_A\right\rangle_L^2}{n}} = \frac{L^d}{\sqrt{n}}\sqrt{\left\langle(\delta A)^4\right\rangle_L - \left\langle(\delta A)^2\right\rangle_L^2}.$$

The fourth-order cumulant $C_4(A)$ of the cumulant expansion of the Gaussian distribution (3.18) yields

$$C_4(A) = \langle A^4\rangle_L - 3\langle A^2\rangle_L^2 = 0.$$

We finally obtain

$$\Delta(n, L) = L^d\sqrt{2}\frac{\left\langle(\delta A)^2\right\rangle_L}{\sqrt{n}} = \sqrt{\frac{2}{n}}C_A.$$

Consequently, the relative error $\Delta(n, L)/C_A$ remains independent of L, and only depends on the number of observations n.

Therefore, we can conclude that increasing the linear size L of a system at a fixed n strongly improves the accuracy of intensive quantities, but, on the other hand, nothing is gained with respect to the accuracy of the fluctuation quantities. In the case of fluctuation quantities, it is more economical in terms of computing time to increase the number of observations, and to choose the smallest size L for which the correlation length between the parts of the system remains much smaller than L, ensuring in this way that the system is not close to a critical point.

Chapter 4

From harmony networks to spin glasses

Simulated annealing in *harmonium Boltzmann machines* (HBM) turns out not to be free of the emergence of various critical phenomena. At certain noticeable temperatures, the cooling process gives rise to rapid and significant changes in the harmony function, conceivably connected with a collective phenomenon among the nodes. Such critical phenomena offer a good ground for a concrete application of the physical interpretation of simulated annealing in association with the analogy between the neural networks and the Ising spin models.

Therefore, we shall exploit the parallel between the binary nodes and the Ising spins, and in this manner obtain a spin model described by a *Hamiltonian.* Whereas the characteristics of the harmony networks are well-defined, it is crucial to determine the distribution of the interactions between the spins composing the spin glass model. Furthermore, we shall see that in this physical analogy local fields come into existence. We shall establish their prevailing influence on the emergence of critical phenomena. It is also relevant to determine to what extent the simulated annealing schedule has to be pursued. Hence, we shall investigate the encountered phase transition in low temperature regions, and we shall conclude with its irrelevance for the correctness of the optimisation task.

4.1 Mapping to an Ising spin model

The harmony network introduced in Chapter 2 is composed of two sets of different nodes, namely, a set of P pattern nodes $\mathbf{a} = (a-1, \ldots, a_P)$, where $a_i \in \{0, 1\}$, and a set of F feature nodes $\mathbf{u} = (u-1, \ldots, u_F)$, where $u_i \in \{-1, +1\}$. These two sets of nodes are connected to each other by arcs whose signs are determined by the components $\omega_{ij} \in \{-1, 0, +1\}$ of the P pattern vectors $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iF})$, and whose strengths are given by

$$\sigma_i^R \equiv \frac{\lambda_i^R}{1-\kappa_i}, \tag{4.1}$$

where

$$\kappa_i \equiv 1 - \frac{2}{1+|\boldsymbol{\omega}_i|} \qquad \text{and} \qquad \lambda_i^R \equiv \lambda_i + \lambda^* \max|\lambda_i| \geq 0. \tag{4.2}$$

We note that $\lambda^* \geq 1$, and since $|\boldsymbol{\omega}_i| \equiv \sum_{j=1}^F |\omega_{ij}|$, we have $2 \leq |\boldsymbol{\omega}_i| \leq F$. The harmony function that maps the configurations of the nodes to real numbers is given by

$$H^R(\mathbf{a}, \mathbf{u}) \equiv \sum_{i=1}^P \sigma_i^R a_i h(\mathbf{u}, \boldsymbol{\omega}_i), \tag{4.3}$$

where

$$h(\mathbf{u}, \boldsymbol{\omega}_i) \equiv \frac{\mathbf{u} \cdot \boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|} - \kappa_i.$$

Despite the apparent lack of relationship of the harmonium Boltzmann machine with physical systems, both the existence of the binary nodes and an inherent dynamics associated with a harmony function to be optimised and the binary nodes suggest the possibility of resorting to an Ising spin model for the study of the main characteristics of this network. Consequently, we shall introduce the spin variables of values ± 1, for the feature nodes $S_i^f \equiv u_i$ and for the pattern nodes $S_i^p \equiv 2a_i - 1$. This way, the harmony function reformulated in terms of a Hamiltonian for the spin model becomes

$$\begin{aligned} \mathcal{H}(\{S_i^f\}, \{S_i^p\}) &\equiv -H^R\Big(\mathbf{a}(\{S_i^p\}), \mathbf{u}(\{S_i^f\})\Big) & (4.4) \\ &= -\sum_{i=1}^P \sum_{j=1}^F J_{ij} S_i^p S_j^f - \sum_{i=1}^P \tilde{h}_i^p S_i^p - \sum_{i=1}^F h_i^f S_i^f + \text{constant}, & (4.5) \end{aligned}$$

where the set of interactions is given by

$$J_{ij} = \frac{\sigma_i^R \omega_{ij}}{2|\boldsymbol{\omega}_i|}. \tag{4.6}$$

Because pattern nodes only connect to feature nodes, and vice versa, the interactions can be chosen to be symmetric, i.e., $J_{ji} = J_{ij}$. As a result, the HBM can be compared to an Ising spin model, where the interactions are physically associated with quenched random variables, whose values are not supposed to change during the simulated annealing process.

Equation (4.5) gives evidence for the appearance of local fields, which can be physically associated with external magnetic constraints. The field at the location of the feature spin S_i^f is given by

$$h_i^f = \sum_{j=1}^P \frac{\sigma_j^R \omega_{ji}}{2|\boldsymbol{\omega}_j|} = \sum_{j=1}^P J_{ji}, \tag{4.7}$$

and the field at the location of the pattern spin S_i^p is

$$\tilde{h}_i^p = -\frac{\sigma_i^R \kappa_i}{2} \leq 0. \tag{4.8}$$

Moreover, the constant term appearing in the expression of $\mathcal{H}$ does not influence the distribution of the configurations if it is the Gibbs distribution, as expected during the simulation. The transformation of a harmony network to an Ising spin model is now complete.

The primary purpose of the HBM is the Maximum Likelihood Estimation task as discussed in Chapter 2. In this context, some feature nodes are supposed to be clamped to fixed values during simulated annealing. The clamping of nodes corresponds to a query or to an incomplete assignment in the original Bayesian paradigm. Physically, the local freezing of a spin can be achieved by locally applying extremely strong fields along whose direction the spin will align. The artificial freezing of some of the feature spins at the values $\hat{S}_i^f$ reduces the number of stochastic variables. The interactions of the clamped spins with the pattern spins have to be considered. From equation (4.5), one easily obtains for the remaining stochastic spins the effective Hamiltonian

$$\mathcal{H}_{\text{eff}}(\{S_i^f\},\{S_i^p\}) = -\sum_{i=1}^{P} \sum_{\substack{j=1 \\ \text{unclamped}}}^{F} J_{ij} S_i^p S_j^f - \sum_{i=1}^{P} h_i^p S_i^p - \sum_{i=1}^{F} h_i^f S_i^f, \tag{4.9}$$

where the effective local field is

$$h_i^p = \tilde{h}_i^p + \sum_{\substack{k \\ \text{clamped}}} \hat{S}_k^f \sum_{i=1}^{P} J_{ik}. \tag{4.10}$$

Thus, a freezing of some stochastic variables has the major effect of modifying the local fields of the pattern nodes. If it occurs that the induced local fields become fairly large compared to the scale of the interactions J_{ij}, their effects can overwhelm the contribution of the interactions with the free spins, aligning in this way the concerned spins along the direction of the local fields.

4.2 The interaction distribution

Before discussing the implications of the existence of the local fields, we shall investigate more precisely the statistical characteristics of the aforementioned quenched interactions.

Each pattern node is associated with the parameter

$$\lambda_i = \ln \mathcal{P}\{U_i = u_i \mid \bigwedge_{U_j \in \mathbf{F}_{U_j}} U_j = u_j\}. \tag{4.11}$$

For zero probability $\mathcal{P}$, the values of the occurring probabilities are approximated by a small constant $\varepsilon \in [10^{-15}, 10^{-1}]$ as mentioned in Section 2.6. Thus, ε provides us with the value of $\lambda_0 \equiv \max_i|\lambda_i| = |\ln \varepsilon|$. On the other hand, we can conveniently express the interactions J_{ij} as a function of the independent variables $\boldsymbol{\omega}_i$ and λ_i, which yields

$$|J_{ij}| = \frac{\lambda_i + \lambda^* \lambda_0}{4}(1 + |\boldsymbol{\omega}_i|^{-1}).$$

If we now assume that a sufficiently large NEULA description code provides a large number of conditional probabilities $\mathcal{P}$, it is likely that we will find them uniformly distributed in the range $[\varepsilon, 1]$. Hence, equation (4.11) implies the following distribution for the λ_i:

$$P(|\lambda_i|) \propto \exp(-|\lambda_i|).$$

Moreover, we have

$$P\big(1 + |\boldsymbol{\omega}_i|^{-1}\big) \propto |\boldsymbol{\omega}_i|^2 \, P(|\boldsymbol{\omega}_i|).$$

Consequently, the distribution of the interactions depending on the independent variables $|\boldsymbol{\omega}_i|$ and λ_i is

$$P(|J|, |\boldsymbol{\omega}_i|) \propto |\boldsymbol{\omega}_i|^2 P(|\lambda_i|) P(|\boldsymbol{\omega}_i|) \propto |\boldsymbol{\omega}_i|^2 P(|\boldsymbol{\omega}_i|) \exp\Big(4|J|\big(1 + |\boldsymbol{\omega}_i|^{-1}\big)^{-1}\Big). \quad (4.12)$$

Finally, the total distribution of the interaction magnitude, disregarding the $P(|\boldsymbol{\omega}_i|)$ distribution, can be expressed as

$$P(|J|) \propto \sum_{|\boldsymbol{\omega}_i|=2}^{F} P(|J|, |\boldsymbol{\omega}_i|). \quad (4.13)$$

The exponential dependence in equation (4.12) clearly predicts a strong concentration of the interaction magnitudes close to their maximum. The original crude assumption is effectively corroborated by the experimental results (see Figure 4.1). Nonetheless, the experimental results exhibit an asymmetry of the proportion of negative and positive interactions that is not apparent from equation (4.12). This effective asymmetry is simply caused by the difference of positive and negative components ω_{ij} determining the sign of the interactions. We also note that the probabilities $\mathcal{P} = 0$ occur relatively frequently. In this case, we have $|J_{ij}| = \lambda_0(\lambda^* - 1)(1 + |\boldsymbol{\omega}_i|^{-1})/4$ that results in extremely weak interactions for the usual choice $\lambda^* \approx 1$.

The strongly peaked asymmetrical bond distribution constitutes a nonnegligible statistical feature of the harmonium Boltzmann machine. For later convenience, we shall introduce the asymmetry factor a that measures the proportion of the negative and the positive interactions. In the ideal case, the theoretical analysis justifies the simplification of only considering the dominant role played by the large interactions in $\mathcal{H}_{\text{eff}}$, formally written as

$$P(J_{ij}) = a\,\delta(J_{ij} + J) + (1 - a)\,\delta(J_{ij} - J),$$

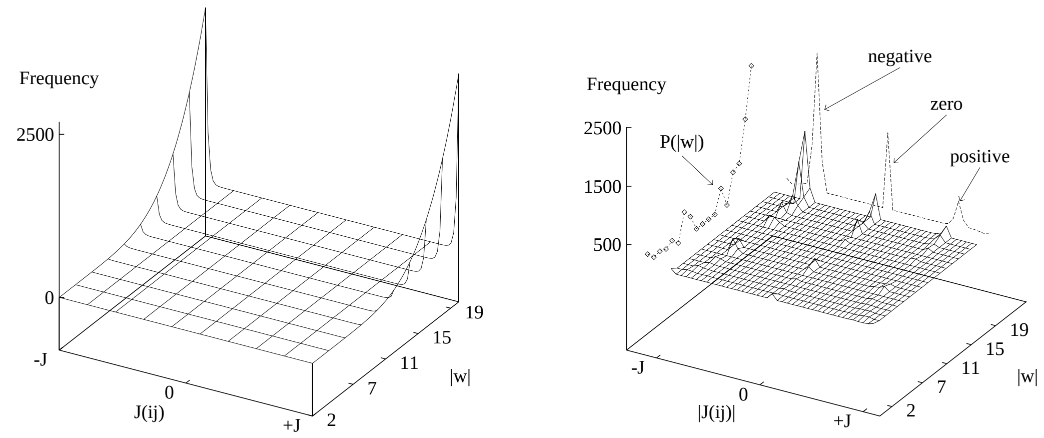


Figure 4.1: Distribution landscape of the interactions J_{ij}. In the left-hand graph, the theoretical distribution $P(|J|, |\boldsymbol{\omega}_i|)$ is depicted assuming that $P(|\boldsymbol{\omega}_i|) \propto 1$ and that there is no asymmetry between the negative and positive interactions. In the right-hand graph, a typical interaction landscape for a large harmony network generated by the NEULA compiler is shown with the corresponding projections of the frequencies. The zero-interactions arise from the conditional probabilities $\mathcal{P} = 0$ after rescaling.

where $J > 0$ represents the maximum of the absolute values of the interactions.

A relevant aspect of the specific topology of the harmonium Boltzmann machine now has to be brought to the fore. Equation (4.11) describes the relationship of the pattern nodes with the feature nodes, and hence, it follows from the generation process of the HBM that the number of pattern nodes depends exponentially on the number of feature nodes. One expects $P \sim \exp(F)$. Moreover, due to the bipartitional structure of the nodes, for large networks the order of the number of interactions per nodes is $P \cdot F/(P + F) \sim F \ll P$. By comparison, a fully connected network provides a ratio of order P. We can conclude that the harmonium Boltzmann machine is in fact very sparsely connected.

4.3 Local fields and critical phenomena

As we established previously, there exist local fields acting upon the spins of the network. We shall see later in this section that their effects cannot be omitted, and hence, the study of the field distribution will receive much attention. In view of the magnitude of these fields, their influence on the relaxation of the spins through Glauber dynamics during the cooling process finds a clear explanation widely verified by the numerous simulation runs. However, the lack of homogeneity of these fields constitutes an obstacle to any statistical treatment of the dynamics of the simulated annealing process. Especially, the attempt of predicting the temperatures at which critical phenomena occur boils down to giving their order of magnitude.

The case of the local fields of the feature spins and of the pattern spins must be handled separately.

The feature fields happen to be strongly dependent on the previously introduced asymmetry factor a of the HBM. In fact, because of the relative frequency of the two kinds of nodes, the number of incoming arcs from the P pattern nodes for each of the F feature nodes is large, of order P. Therefore, we can estimate the order of magnitude of the feature fields as

$$h_i^f = \sum_{i=1}^{P} J_{ij} \sim (1-2a)PJ, \tag{4.14}$$

where $J > 0$ and a represents the proportion of negative to positive interactions. Since $P \gg F$, the feature fields are extremely sensitive to the asymmetry of the network, as attested in Figure 4.2.

On the other hand, equation (4.8) yields

$$\frac{|\tilde{h}_i^p|}{|J_{ij}|} = |\boldsymbol{\omega}_i| \frac{|\boldsymbol{\omega}_i|-1}{|\boldsymbol{\omega}_i|+1}, \tag{4.15}$$

where $|\boldsymbol{\omega}_i| \in \{2, \ldots, F\}$. This ratio takes its maximal value at $|\boldsymbol{\omega}_i| = F$, where the minimum value $\tilde{h}_i^p \sim -FJ < 0$ is attained. Contrary to the case of the feature fields, the ratio in equation (4.15) is independent of a. The pattern fields become more important for large F — which means large networks — and their magnitude remains significantly larger than the interaction magnitude J (see Figure 4.2).

4.3.1 Field effects on Glauber dynamics

Glauber dynamics constitutes the principal key for the understanding of the dominant role played by the local fields. As a matter of fact, although the spins interact, a very strong local field renders a spin almost independent of these interactions by tending to align the spin along its direction. We can see this by considering Glauber dynamics given by

$$P\{S_i = \pm 1\} = \frac{1}{1+\exp(\mp 2\, I/T)}, \tag{4.16}$$

where the input field, or molecular field, for the spin S_i is

$$I = \sum_j J_{ij} S_j + h_i. \tag{4.17}$$

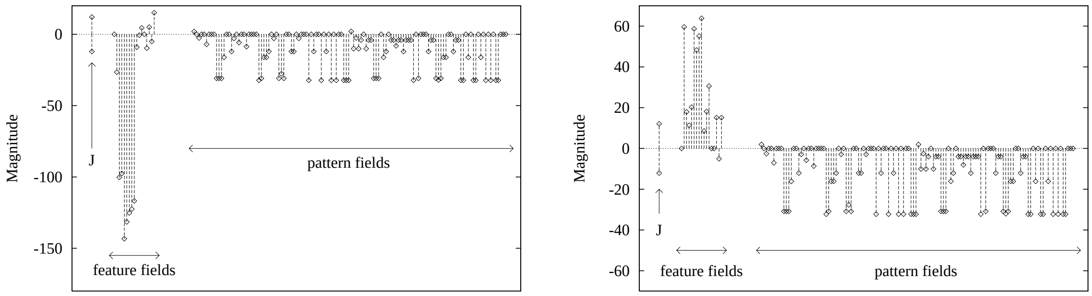


Figure 4.2: Distribution of the local fields of a harmony network with two different asymmetries and the magnitude of the maximal encountered interaction. In the left-hand graph, the asymmetry $a = 0.72$ induces strong negative local fields for the feature nodes. In the right-hand graph, the same network with an asymmetry $a = 0.2$ clearly exhibits positive feature fields. We note that the pattern fields are independent of the asymmetry of the network.

If the values of the spins interacting through I are random and the magnitude of the field h_i is larger than the typical interaction J_{ij}, we can expect h_i to be the dominant term in equation (4.17). Consequently, as soon as the temperature achieves the order of magnitude of h_i, the probability $P\{S_i = \pm 1\}$ will be entirely determined by the sign of I, which is itself determined by the sign of the external constraint h_i.

It is needless to say that this argument slightly resembles handwaving. Nevertheless, the dominant effect of the pattern fields, as well as the correlation between the temperature and the values of the fields, have clearly been established by simulation, as suggested in Figure 4.3. In the left-hand graph, the abrupt changes in the values of the Hamiltonian $\mathcal{H}_{\text{eff}}$ appear at much higher temperatures when the original local fields prevail, while the behaviour of the same network with zero fields gives evidence of the absence of these fields, as expected. Furthermore, the encountered evolution of the heat capacity clearly attests the correlation between the network asymmetry and the onset of the critical phenomenon.

The evidence of the predominance of the local field is particularly significant in the case of a strongly asymmetrical network. Indeed, the effect of the presence of large feature fields can be proved by considering the terms of the Hamiltonian separately. The effective Hamiltonian can be conveniently rewritten as

$$\mathcal{H}_{\text{eff}} = -\big(I + H_p + H_f\big), \tag{4.18}$$

where

$$H_p \equiv \sum_i h_i^p S_i^p \qquad \text{and} \qquad H_f \equiv \sum_i h_i^f S_i^f$$

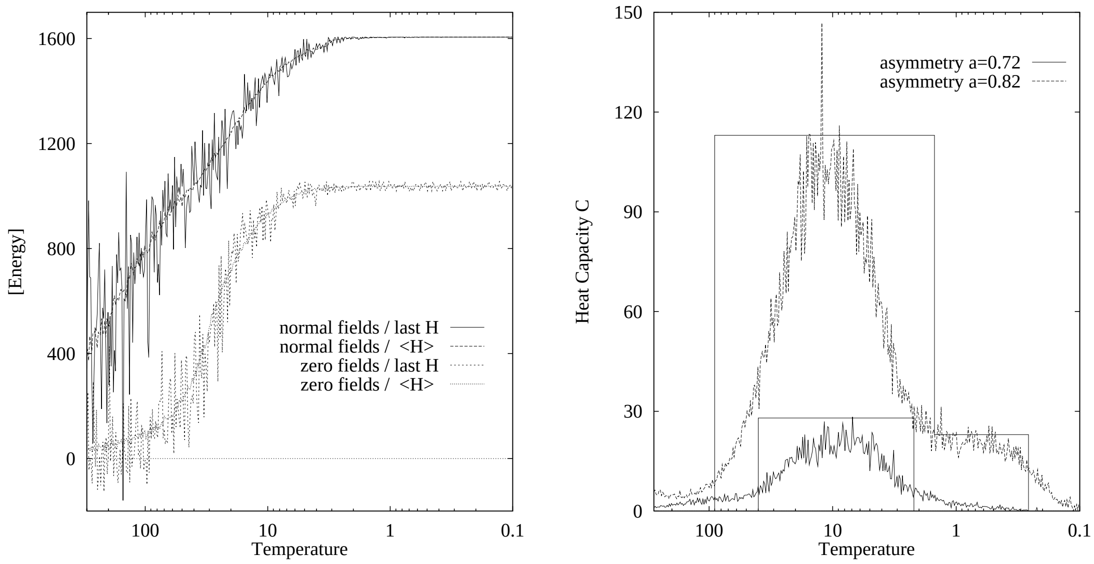


Figure 4.3: Critical phenomena during simulated annealing. In the left-hand graph, the evolution of the negative of the Hamiltonian of an HBM is depicted for the cooling process comparing the effect of the original local fields with the effect of zero fields, giving evidence for the influence of the fields. Moreover, the last sampling value of $\mathcal{H}_{\text{eff}}$ can be compared with $\langle\mathcal{H}_{\text{eff}}\rangle$. In the right-hand graph, two networks with different asymmetries clearly exhibit a critical phenomenon at different temperatures due to the magnitude of the feature fields.

and

$$I \equiv \sum_{ij} J_{ij} S_i^p S_j^f = \sum_i K_i\big(\{S_j^f\}\big) S_i^p .$$

The new dynamic variables $K_i\big(\{S_j^f\}\big)$ are defined as

$$K_i\big(\{S_j^f\}\big) \equiv \sum_j J_{ij} S_j^f .$$

We shall now consider a strongly asymmetrical HBM. The aim of decreasing the temperature continuously and updating the configuration of the nodes according to Glauber dynamics is minimising this Hamiltonian. According to the aforementioned argument, below a threshold temperature, the strong external constraints h_i^f tend to align the feature spins into a given fixed orientation. The larger h_i^f is, the more probable is that S_i^f points at the same direction. If we define $\Gamma \equiv \{S_i^f \in \Omega \mid |h_i^f| \gg J > 0\}$, we can write

$$H_f = \sum_{S_i^f \in \Gamma} h_i^f S_i^f + \sum_{S_i^f \in \Omega \setminus \Gamma} h_i^f S_i^f \approx \sum_{S_i^f \in \Gamma} h_i^f S_i^f \approx \text{constant},$$

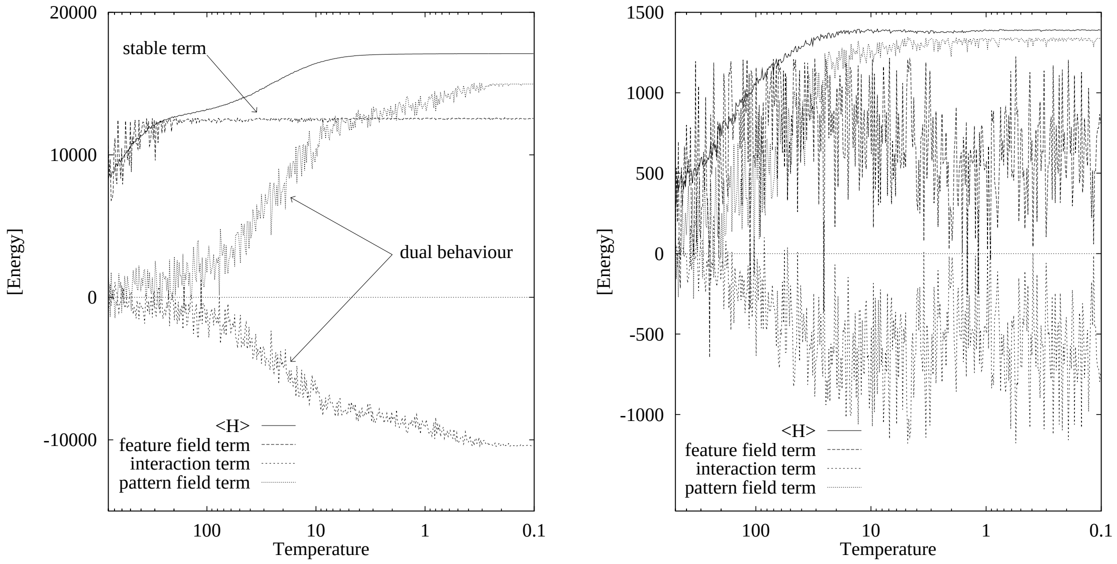


Figure 4.4: Behaviour of the different components of the negative of the Hamiltonian $\mathcal{H}_{\text{eff}}$ for two networks with strong asymmetries. In the left-hand graph, for an asymmetry $a = 0.82$, the feature field term stabilises at high temperatures, causing the dual fluctuations of the interaction term and the pattern field term. In the right-hand graph, the same behaviour is visible for a strong asymmetry $a = 0.2$.

because most of the feature spins in Γ are frozen by their local fields. This approximately constant term is positive because of the alignment. Consequently, we obtain

$$\mathcal{H}_{\text{eff}} \approx -(I + H_f) \; + \; c = -\sum_i \big(K_i + h_i^p\big) \; S_i^p \; + \; c,$$

where c is a constant. Because of the assumption of large asymmetry, we can expect the majority of the feature spins to be constrained by their fields, and henceforth we can consider $K_i\big(\{S_i^f\}\big) = K_i$ to be approximately constant. Next, we shall express the input field I_i^p of the pattern spin S_i^p as

$$I_i^p = \sum_j J_{ij} S_i^f + h_i^p = K_i + h_i^p.$$

Once more, Glauber dynamics determines the evolution of the values of the pattern spins. As the temperature decreases during simulated annealing, the pattern spins with a ratio I_i^p/T of order unity can be considered to have a well-defined orientation according to the sign of I_i^p, while the probability of the flipping of the remaining pattern spins is close to 1/2. Thus, the changes ΔI and ΔH_p after a Monte Carlo step are essentially due to these oscillating pattern spins. Because $K_i + h_i^p \approx$ constant, we have $\Delta I \approx -\Delta H_p$, which states that these two terms exhibit a simultaneously dual behaviour during simulated annealing. As the temperature decreases, the number of oscillating pattern spins also decreases, in this manner reducing the dual fluctuations of I and H_p and simultaneously minimising $\mathcal{H}_{\text{eff}}$.

This phenomenon, appearing in asymmetrical HBM and proving the predominance of the feature fields, is depicted in Figure 4.4 for two different networks with asymmetries $a = 0.2$ and $a = 0.8$. The phenomenon is reproducible and its emergence depends only on the value of the asymmetry factor.

4.3.2 Low temperature anomaly of the specific heat

The study of the HBM in low temperature region is instructive. Indeed, we can expect the final configuration of the nodes to have the same number of pattern nodes in the state $a_i = 1$ as the number of the cliques in the original Bayesian network, as discussed in Section 2.6. The coherence of the final value of the harmony function seems to depend on the fulfilment of such a requirement.

From the physical point of view, the third law of thermodynamics [Reic91] states that in the limit $T \to 0$, we have

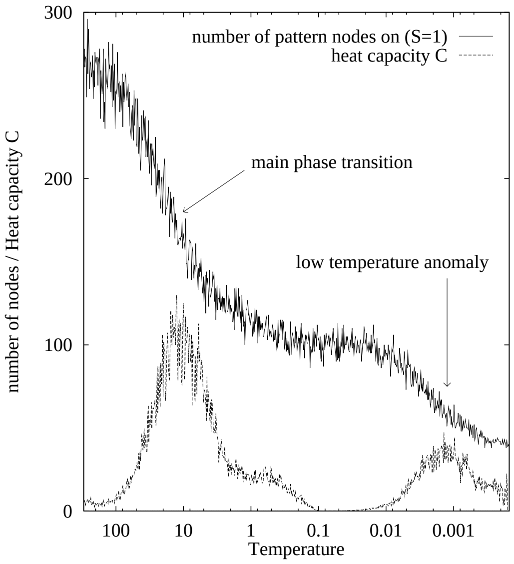
$$C = \left(\frac{\partial \langle \mathcal{H} \rangle_{\mathrm{T}}}{\partial T}\right) = T \left(\frac{\partial S}{\partial T}\right) \to 0.$$

Consequently, the occurring of critical phenomena emerging from a final ordering process of the spins at low temperatures is expected to end before the temperature reaches the zero limit. Surprisingly, this ordering is responsible for an anomalous increasing of the heat capacity. Even though the cooling down of the HBM to a very low temperature is a prerequisite to the achievement of the expected configuration, as can be seen in Figure 4.5, this does not constitute a matter of concern. In fact, the critical phenomenon in question involves only configurations of energies very close to each other and hence has no influence on the optimisation task of simulated annealing.

In order to justify this claim, we shall consider a set of configurations energetically close to the energy minimum. For simplicity, we shall approximate these states as a continuum of nondegenerated configurations of energy, E_{min} to $E_{min} + \Delta E$, where $\Delta E \ll 1$ is a small positive energy difference.

Simulated annealing causes the system to reach a stationary configuration distribution which, for an appropriate cooling schedule, is the Gibbs distribution P_{equ}. According to its definition $P_{equ} = Z^{-1} \exp(-\beta \mathcal{H}_{\text{eff}})$, at low temperatures, solely the states of low energy can be macroscopically occupied. Thus, we obtain the following expectation value of the energy:

$$\langle \mathcal{H} \rangle_{\mathrm{T}} = \int\limits_{E_{min}}^{\infty} \mathrm{d}E \; E \exp(-\beta E) = -\frac{\partial}{\partial \beta} \int\limits_{E_{min}}^{E_{min}+\Delta E} \mathrm{d}E \; \exp(-\beta E).$$



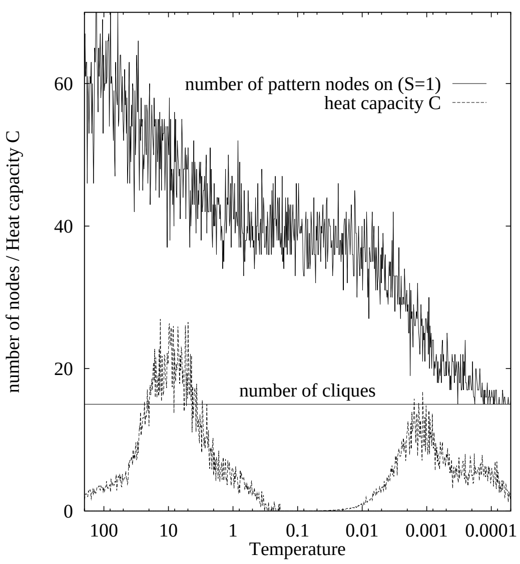


Figure 4.5: Behaviour of the heat capacity in relationship with the relaxation of the network in its final configuration. Both simulations show that the number of pattern nodes on, meaning $S_i^p = 1$, converge towards the number of cliques. Because of the weak energy difference between the involved configurations, a final alignment occurs at low temperatures and causes the heat capacity to peak.

Hence, the heat capacity C can be written as

$$C = -\beta^2 \frac{\partial \langle \mathcal{H} \rangle_{\mathrm{T}}}{\partial \beta} = \beta \exp(-\beta E_{min}) \big(E_{min}^2 - \exp(-\beta \Delta E)(E_{min} + \Delta E)^2 + O(T) \big).$$

It follows that the heat capacity C exhibits a peak with a maximum at the temperature

$$T \approx \frac{\Delta E}{3 \ln \left(1 + \frac{\Delta E}{E_{min}} \right)} = \frac{\Delta E}{3} + O\big(\Delta E^2\big).$$

At a low temperature, such a peak is a typical anomaly of the heat capacity. This phenomenon is widely observed in solid state physics — for instance, in malleable solids — and is due to the many degrees of freedom that are almost energetically equivalent.

We can conclude that despite the appearance of a critical phenomenon at low temperature, simulated annealing does not have to be extended to such low temperatures. Indeed, since the process is only concerned with configurations whose energy is very close to the global energy minimum, the accuracy of the optimisation task is not endangered.

4.4 Discussion

Despite the difficulties of any analytical treatment of the exact dynamics of simulated annealing, it remains possible to extract the main features of the critical phenomena. Whereas the local fields are site dependent, the estimation of their order of magnitude gives significant information about the temperature of interest.

It is particularly interesting to estimate the critical temperatures in order to achieve, by an adapted cooling schedule, a higher reliability of the result of simulated annealing in the frame of an optimisation task. The lowering of the temperature has to be slowed down at the sensitive points of high heat capacity when the energy of the system is susceptible to large changes. In this context, the estimation of the local fields can give important markers telling where the cooling schedule has to be slowed down, namely, coarsely at the temperatures corresponding to the magnitudes of the local fields. The strongly asymmetrical HBM appears to be sensitive to the existence of the feature fields, but in view of the pattern fields' independence of the parameter a, we can also expect the symmetrical HBM to display constrained behaviour.

Moreover, in spite of the anomalous behaviour of the heat capacity at low temperatures, it clearly appears that there is no need to extend the cooling process too far if we aim at optimising the energy, or synonymously, optimising the harmony function.

Chapter 5

Strongly diluted Ising spin glasses

In order to study the major features of spin glass phenomena, we must employ methods that capture the essential physics of the systems. Unfortunately, we are faced with the fundamental problem that we do not know all the parameters of such statistical ensembles; the randomness of the interactions introduces special features into this theoretical problem of statistical mechanics. Consequently, we could not normally calculate the properties of a single sample nor those of a particular set of values of the random interactions J_{ij}, even if we knew them.

Fortunately, as statistical physics teaches us [Reic91], the relative fluctuations of the energy $\Delta E/E$ around its thermal average mean value are of the order $N^{-1/2}$. Hence, we expect that its sample-to-sample fluctuations vanish in the *thermodynamical limit*, namely, the limit of large systems. Therefore, one hopes that the relevant physics occurring in spin glasses demonstrates such a behaviour, which is called *self-averaging*.

Under certain assumptions, the attempt to carry out formally the average over the distribution of the random parameters J_{ij} reveals itself as practicable. This so-called quenched averaging leads us to a problem in which the disorder no longer appears explicitly; the problem is reduced to a nonrandom, translationally invariant, single-site statistical mechanical problem. The spins are no longer formally correlated, but independent, and therefore this method can be qualified as a *Mean Field Theory* (MFT). MFT is especially appropriate for models with long-range interactions, because it provides an exact solution in these cases.

Harmony networks, particularly those generated by the NEULA compiler, appear to fulfil the requirements that lead to sensitive, but nontrivial thermodynamics. This property of the harmony network motivated our choice of the MFT. MFT can be applied to the class of networks described in Section 5.1. Harmony networks are a subset of this large class.

5.1 The model

Let us consider N Ising spins, whose random interactions are strong (of order unity), but only a fraction of them (decreasing as N^{-1}) are present. This condition on the bond dilution is necessary to make the model exactly soluble [ViBr85].

The *Hamiltonian* $\mathcal{H}$ describing the spin glass is defined as

$$\mathcal{H}(\{J_{ij}\},\{S_i\}) \equiv -\sum_{(i,j)} J_{ij} S_i S_j - h \sum_i S_i, \qquad \text{where} \qquad S_i = \pm 1. \tag{5.1}$$

The random interactions are symmetrical, i.e., $J_{ij} = J_{ji}$, and the field h is homogeneous.

We shall now consider the following distribution of interactions:

$$P(J_{ij}) \equiv \frac{p}{N}\, P_p(J_{ij}) + (1 - \frac{p}{N})\, \delta(J_{ij}), \tag{5.2}$$

where P_p denotes the distribution of the nonzero diluted bonds. MFT can only be applied if the interactions are sufficiently diluted; namely, if we want to maintain in the thermodynamical limit a suitable distribution P_p whose mean and variance are nontrivial, we must impose $p < \infty$. Nevertheless, one recovers in the limit $p \to \infty$ the results of the *Sherrington-Kirkpatrick model* [KiSe78], where the mean of the distribution P_p is of the order N^{-1} and its variance of the order $N^{-1/2}$. In other words, the strength of the interactions has to be balanced by their sparseness.

The parameter p introduced above represents, disregarding a factor of $(1 - 1/N)$, the number of nonzero interactions per site. Clearly, if one alters a fully connected system with N spins but preserves a fraction p/N of its links after the pruning of the other links, one obtains

$$L = \frac{p}{N} \frac{N(N-1)}{2} = \frac{p}{2}(N-1), \tag{5.3}$$

where L denotes the number of nonzero interactions among the N spins. The bipartitional structure of the nonconnected nodes of the harmony network particularly reduces the effective maximal possible number of arcs among the nodes. Especially, the mapping from a Bayesian network provides two strongly asymmetrical clusters of nodes characterised by the exponential dependence of the number P of pattern nodes on the number F of feature nodes. Therefore, we have to consider a number of arcs per node proportional to the number of feature nodes, which implies a ratio p/N of the order $\ln N/N$ for a large network mainly consisting of pattern nodes, where $N \approx P \approx \exp(F)$. Moreover, the NEULA compiler pruning of the unnecessary arcs increases the sparseness of the nonzero arcs. Consequently, it is reasonable to suppose that the requirement $p < \infty$ of strong dilution is fulfilled for the harmony networks under consideration.

5.2 The replica method

5.2.1 The quenched average

A proper treatment of a closed and isolated spin glass, thermally coupled with its surroundings and with quenched disordered interactions, involves the calculation of the free energy per site. This thermodynamic potential has the property of reaching its minimum in equilibrium, de facto maximising the entropy of the system [Reic91].

In the case of random interactions, the thermodynamic limit of the free energy per site is defined as

$$\beta \mathrm{f}(\{J_{ij}\}) \equiv \lim_{N\to\infty} -\frac{1}{N}\ln Z(\{J_{ij}\}),$$

where Z denotes the partition function

$$Z\{J_{ij}\} \equiv \sum_{\{S_i\}\in\{-1,+1\}^N} \exp\bigl(-\beta\mathcal{H}(\{J_{ij}\},\{S_i\})\bigr) = \mathrm{Tr}_S \exp\bigl(-\beta\mathcal{H}(\{J_{ij}\},\{S_i\})\bigr). \quad (5.4)$$

We have simultaneously introduced the trace convention $\mathrm{Tr}_S = \sum_{S_i}$, representing the sum over the configuration.

Unfortunately, we have to work on systems having different sets of interactions. Hence, predicting the behaviour of the various representatives for a given bond distribution separately is a complex problem. Despite this difficulty, the methods of statistical mechanics can give meaningful results under the *self-averaging* assumption. The concept of self-averaging can be illustrated by the following example.

Consider a single very large system of spins and divide it into large number of subsystems. Each subsystem is macroscopic and contains a different set of random interactions. If we assume that the coupling between the subsystems is negligible, then we can expect that the result of averaging over a large number of subsystems will differ by only a small amount from the result of averaging over all possible choices of the distribution of the interactions. In other words, if we look at, for instance, the per-spin magnetization

$$M = \frac{1}{N}\sum_i \langle S_i\rangle_{\mathrm{T}} = \frac{1}{N}\sum_i \left(\frac{1}{Z}\mathrm{Tr}_S\, S_i \exp\bigl(\beta\mathcal{H}(\{J_{ij}\},\{S_j\})\bigr)\right)$$

for a system with a given fixed interaction probability distribution, we can expect with reasonably large probability that

$$\lim_{N\to\infty}\bigl(M - \langle M\rangle_{\mathrm{J}}\bigr) = 0.$$

The same result holds for any other intensive quantity. This property of a system is called self-averaging. It can be proved for certain models [HePa82], but clearly it remains an assumption that is totally in accord with the experiments [BiYo86]. This assumption of self-averaging in the limit of a large system justifies the performing of the averaging over the bond distribution as a method for the statistical study of spin glasses with disordered interactions.

The average with respect to the fixed interaction distribution $P(J_{ij})$ is called the *quenched average*. Physically, the random variables may themselves fluctuate with time; however, in our case, the typical time of fluctuation is considered to be much larger than the time needed for the statistical variables to reach equilibrium; thus, in the time scale of the system, the random interactions can be treated as quenched. The quenched average $\langle \cdot \rangle_{\mathrm{J}}$ is defined as

$$\big\langle \ln Z\{J_{ij}\}\big\rangle_{\mathrm{J}} = \int \prod_{(i,j)} (\mathrm{d}J_{ij} P(J_{ij})) \ln Z\{J_{ij}\}. \tag{5.5}$$

We note that performing this average establishes a translational invariance. Consequently, the previously introduced magnetisation per site can be rewritten as

$$\langle M\rangle_{\mathrm{J}} = \big\langle \langle S_i\rangle_{\mathrm{T}} \big\rangle_{\mathrm{J}} = \frac{1}{N}\sum_i \langle S_i\rangle_{\mathrm{T}}$$

for any site i in the thermodynamical limit. The second equality follows from the self-averaging assumption and states that the configuration average $\big\langle \langle S_i\rangle_{\mathrm{T}} \big\rangle_{\mathrm{J}}$ is equal to the average over the sites for a single sample.

Assuming that the self-averaging property holds for the intensive free energy per site, we may now calculate its quenched average. Obviously, this task is very difficult, because both the calculation of $\ln Z$, which depends on the set of the interactions J_{ij} describing the quenched disorder, and averaging over the disorder in the thermodynamical limit remains hopeless due to the random variables occurring inside the logarithm. Furthermore, it is clear that attempting to compute $\ln\langle Z\rangle_{\mathrm{J}}$ leads to erroneous conclusions. Indeed, since the Einstein fluctuation theory [Reic91] states that

$$P(f) \propto \exp\left(-N\frac{(f-\langle f\rangle_{\mathrm{J}})^2}{2(\Delta f)^2}\right),$$

we obtain from the partition function $Z = \exp(-N\beta f)$ the result

$$f = -\frac{1}{N}\beta \ln \langle Z\rangle_{\mathrm{J}} = \langle f\rangle_{\mathrm{J}} + \beta(\Delta f)^2 \neq \langle f\rangle_{\mathrm{J}},$$

which would be the result of an average over nonquenched disordered interactions. However, there is a way to bypass these difficulties. The idea of calculating the average (5.5) rests on the use of the following relation for $n > 0$:

$$\langle \ln Z\rangle_{\mathrm{J}} = \lim_{n\to 0} \frac{1}{n}\big(\langle Z^n\rangle_{\mathrm{J}} - 1\big) = \lim_{n\to 0}\frac{1}{n}\ln\langle Z^n\rangle_{\mathrm{J}}.$$

The idea behind this so-called *replica method* is to duplicate the partition function and to obtain n different replicas of the spin glass, apparently all independent of each other. The partition function of the α^{th} replica is denoted Z_α. As we will see later, the quenched average $\langle Z^n\rangle_{\mathrm{J}}$ can be performed for any integer n, except $n = 0$. Consequently, one must interpret that the limit is reached by analytic continuation and that this provides the desired quantity. Nevertheless, the formal aspects of this analytic procedure leaves much to be desired, but this does not render the technique less useful.

5.2.2 Order parameters

The partition function Z can exhibit singularities. Such singularities may be physically associated with phase transitions. They are classified according to the behaviour of the free energy [Mour84]. A phase transition is characterised by a relative ordering of the spins below a freezing temperature. The ordering means that there is a nontrivial thermal correlation of many spins in the spin glass. Therefore, one needs to capture the appearance of correlated fluctuations of these spins in so-called *order parameters*. Fortunately, within the replica method, it is possible, in addition to computing the free energy, to get around performing the difficult quenched averages of these spins correlations. For that, we shall consider the expression

$$q \equiv \left\langle \langle S_i \rangle_{\mathrm{T}}^2 \right\rangle_{\mathrm{J}} = \left\langle \frac{Z^{n-2} \big(\mathrm{Tr}_S\, S_i \exp(-\beta \mathcal{H})\big) \big(\mathrm{Tr}_S\, S_i \exp(-\beta \mathcal{H})\big)}{Z^n} \right\rangle_{\mathrm{J}}. \tag{5.6}$$

The numerator involves a trace over n replicas, where two of them, say μ and ν (obviously $\mu \neq \nu$), are the replicas appearing explicitly in the equation (5.6). By virtue of the identity $\lim_{n\to 0} Z^n = 1$, the denominator can be ignored in the limit $n \to 0$. Performing the quenched average leads to

$$\lim_{n\to 0} q = \lim_{n\to 0} \mathrm{Tr}_S\, S_i^\mu S_i^\nu \exp(-\beta \mathcal{H}_{\mathrm{eff}}) = \langle S_i^\mu S_i^\nu \rangle_{\mathcal{H}_{\mathrm{eff}}} \equiv \widetilde{q}_{\mu\nu} \quad (\mu \neq \nu),$$

where we have to consider a new effective Hamiltonian $\mathcal{H}_{\mathrm{eff}}$ resulting from this averaging. The order parameters $\widetilde{q}_{\mu\nu}$ are very important in spin glass theory and emerge as a natural choice. We see that the left-hand side of equation (5.6) is independent of the arbitrary choice of the replicas. Consequently, one can expect $\mathcal{H}_{\mathrm{eff}}$ and the $\widetilde{q}_{\mu\nu}$ to have the same property of symmetry of permutation between the replica indices. We shall call this assumption the *replica symmetric ansatz*.

The generalisation to arbitrary products is straightforward and yields

$$\left\langle \langle S_i \rangle_{\mathrm{T}}^m \right\rangle_{\mathrm{J}} = \langle S_i^{\alpha_1} \cdots S_i^{\alpha_m} \rangle_{\mathcal{H}_{\mathrm{eff}}} \equiv \widetilde{q}_{\alpha_1 \cdots \alpha_m},$$

where all replica indices $\alpha_1, \ldots, \alpha_m$ are distinct. In other words, within the replica method, the m-spin correlation functions are independent of the arbitrary choice of the replicas for every integer m.

5.3 The free energy f

The computation of the free energy requires the averaging of the n^{th} power of the partition function over all configurations. Formally, the quenched average $\langle \cdot \rangle_{\mathrm{J}}$ with respect to the distribution $P(J_{ij})$ is

$$\langle Z^n \rangle_{\mathrm{J}} = \mathrm{Tr}_S \Big\langle \exp\Big(\beta h \sum_{i,\alpha} S_i^\alpha + \beta \sum_{(i,j)} J_{ij} \sum_\alpha S_i^\alpha S_j^\alpha\Big)\Big\rangle_{\mathrm{J}}, \tag{5.7}$$

where the trace includes all the possible configurations of the $N \times n$ spins of the n replicas.

We can perform the average in (5.7) over the bond distribution P. The zero interactions J_{ij} contribute by a factor $\exp\Big(\sum_{ij} \ln\big(1-\frac{p}{N}\big)\Big)$. If we denote by $\langle\cdot\rangle$ the average with respect to the nonzero bond distribution $P_p(J_{ij})$, we obtain

$$\begin{aligned}\langle Z^n\rangle_{\mathrm{J}} &= \mathrm{Tr}_S \exp\Big(\beta h \sum_{i,\alpha} S_i^\alpha + \sum_{(i,j)} \ln\Big(1 + \frac{p}{N}\big(\langle \exp(\beta J_{ij} \sum_\alpha S_i^\alpha S_j^\alpha)\rangle - 1\big)\Big)\Big) \\ &= \mathrm{Tr}_S \exp\Big(\beta h \sum_{i,\alpha} S_i^\alpha + \frac{p}{N}\sum_{(i,j)}\big(\langle \exp(\beta J_{ij} \sum_\alpha S_i^\alpha S_j^\alpha)\rangle - 1\big) + O(N^{-2})\Big). \qquad (5.8)\end{aligned}$$

On the second line, we have dropped the higher-order terms of the logarithm. Because $p < \infty$, they do not contribute as extensive terms to the free energy. This is a direct consequence of the strong dilution of the interactions.

On the other hand, because $S_i \in \{-1, +1\}$, we can write

$$\exp(\beta J_{ij} \sum_\alpha S_i^\alpha S_j^\alpha) = \prod_\alpha \cosh(\beta J_{ij})\big(1 + S_i^\alpha S_j^\alpha \tanh(\beta J_{ij})\big).$$

This product over the n replicas can be expanded and inserted into equation (5.8). For convenience, we shall now define the set of functions T_m occurring in the product:

$$\mathsf{T}_m = \langle \cosh^n(\beta J_{ij}) \tanh^m(\beta J_{ij})\rangle. \qquad (5.9)$$

The T_m depend on the temperature T and on the distribution P_p. After carrying out these transformations, we obtain the result

$$\begin{aligned}\langle Z^n\rangle_{\mathrm{J}} = \mathrm{Tr}_S \exp\Big(&\beta h \sum_{i,\alpha} S_i^\alpha + \frac{p}{N}\sum_{(i,j)}(\mathsf{T}_0 - 1) + \frac{p}{N}\mathsf{T}_1 \sum_{(i,j),a} S_i^\alpha \\ &+ \frac{p}{N}\mathsf{T}_2 \sum_{\substack{(i,j)\\ \alpha<\beta}} S_i^\alpha S_i^\beta S_j^\alpha S_j^\beta + \frac{p}{N}\mathsf{T}_3 \sum_{\substack{(i,j)\\ \alpha<\beta<\gamma}} S_i^\alpha S_i^\beta S_i^\gamma S_j^\alpha S_j^\beta S_j^\gamma + \cdots\Big). \qquad (5.10)\end{aligned}$$

Note that the replicas are now coupled, i.e., their behaviour is no longer independent of each other. The N spins still remain coupled, but we can get around this problem. First, the use of the identity

$$\sum_{\substack{(i,j)\\ \alpha_1<\cdots<\alpha_m}} S_i^{\alpha_1}\cdots S_i^{\alpha_m} S_j^{\alpha_1}\cdots S_j^{\alpha_m} = \frac{N^2}{2}\sum_{\alpha_1<\cdots<\alpha_m}\Big(\big(\frac{1}{N}\sum_i S_i^{\alpha_1}\cdots S_i^{\alpha_m}\big)^2 - \frac{1}{N}\Big), \quad (5.11)$$

reduces the coupling of the spins to quadratic terms. The second step consists of removing these quadratic dependences. This can only be achieved by introducing new degrees of freedom in the form of new integration variables. We shall realise this by means of the Gaussian integral, giving

$$\exp(\lambda x^2) = \left(\frac{\lambda}{\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} \mathrm{d}q \exp(-\lambda q^2 + 2\lambda q x).$$

Finally, by removing the nonextensive terms that appear in the argument of the exponential, equation (5.10) becomes

$$\langle Z^n \rangle_{\mathrm{J}} = \int \mathrm{d}\mathbf{q} \exp\bigl(-N f(\mathbf{q})\bigr), \tag{5.12}$$

where

$$f(\mathbf{q}) = \frac{p\mathsf{T}_1}{2} \sum_{\alpha} q_\alpha^2 + \frac{p\mathsf{T}_2}{2} \sum_{\alpha<\beta} q_{\alpha\beta}^2 + \frac{p\mathsf{T}_3}{2} \sum_{\alpha<\beta<\gamma} q_{\alpha\beta\gamma}^2 + \cdots - \ln \mathrm{Tr}_S \exp\bigl(-\beta \mathcal{H}_{\mathrm{eff}}\bigr), \tag{5.13}$$

and

$$\begin{aligned} -\beta\mathcal{H}_{\mathrm{eff}} = \beta h \sum_{\alpha} S^\alpha + p\mathsf{T}_1 \sum_{\alpha} q_\alpha S_a \\ + p\mathsf{T}_2 \sum_{\alpha<\beta} q_{\alpha\beta} S^\alpha S^\beta + p\mathsf{T}_3 \sum_{\alpha<\beta<\gamma} q_{\alpha\beta\gamma} S^\alpha S^\beta S^\gamma + \cdots . \end{aligned} \tag{5.14}$$

This way, we obtain a single-site problem. This and the additional translational invariance resulting from the averaging over the configuration, obviously render the site index i unnecessary. Thus, we can advantageously omit the site dependence of the $N \times n$ spins $\{S_i^\alpha\}$, and restrict to the thermal average Tr_S over the 2^n configurations of the $\{S^\alpha = \pm 1\}$.

However, we are still aiming at obtaining the free energy per site given by

$$f = -\frac{1}{\beta} \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{1}{n} \ln \langle Z^n \rangle_{\mathrm{J}}. \tag{5.15}$$

Indeed, the free energy can be derived from equation (5.12) by using the so-called *saddle point approximation*. Nevertheless, in order to justify this step, we shall assume that the thermodynamic limit $N \to \infty$ can be treated before the limit $n \to 0$. This interchange of the limits has received much attention, and it is believed that this step does not create trouble [BiYo86].

As an example of the saddle point approximation, we shall consider here its one-dimensional version [HeKP91], which asserts that

$$\lim_{N\to\infty} -\frac{1}{N} \ln \int_{-\infty}^{\infty} \mathrm{d}x \exp\bigl(-N f(x)\bigr) = \min_x f(x) = f(x_{saddle}),$$

where the minimum x_{saddle} of f is called *the saddle point*. We can prove this by assuming $f''(x_{saddle}) > 0$. We can now write

$$\int_{-\infty}^{\infty} \mathrm{d}x \exp\big(-Nf(x)\big)$$
$$= \exp\big(-Nf(x_{saddle})\big) \int_{-\infty}^{\infty} \mathrm{d}x \exp\big(-\frac{1}{2}Nf''(x_{saddle})(x - x_{saddle})^2 + \cdots\big).$$

The Gaussian integral contributes a nonextensive term to the logarithm, and thus the exponential becomes dominant.

Applying the generalisation of the saddle point approximation to the free energy per site in equation (5.15) gives

$$f = \frac{1}{\beta} \lim_{n\to 0} \frac{1}{n} \min_{\mathbf{q}} f(\mathbf{q}) = \frac{1}{\beta} \lim_{n\to 0} \frac{1}{n} f_n, \tag{5.16}$$

where we define $f_n = \min_{\mathbf{q}} f(\mathbf{q})$.

At this point, the meaning of the saddle point becomes evident and can be related to the order parameters $\widetilde{q}_\alpha, \widetilde{q}_{\alpha\beta}, \widetilde{q}_{\alpha\beta\gamma}, \ldots$ introduced previously. Indeed, the saddle point fulfils the condition $\partial f(\mathbf{q})/\partial \mathbf{q} = 0$, which applied to the definition of $f(\mathbf{q})$ in equation (5.12) gives the following values for the first components of the saddle point:

$$q_\alpha = \lim_{n\to 0} \frac{\mathrm{Tr}\, S^\alpha \exp(-\beta \mathcal{H}_{\text{eff}})}{\mathrm{Tr} \exp(-\beta \mathcal{H}_{\text{eff}})} = \lim_{n\to 0} \mathrm{Tr}\, S^\alpha \exp(-\beta \mathcal{H}_{\text{eff}}) = \lim_{n\to 0} \langle S^\alpha \rangle_{\mathcal{H}_{\text{eff}}} = \widetilde{q}_\alpha, \tag{5.17}$$

where $\langle \cdot \rangle_{\mathcal{H}_{\text{eff}}}$ denotes the normalised average with respect to the Gibbs distribution with the effective Hamiltonian $\mathcal{H}_{\text{eff}}$. The denominator in equation (5.17) has been dropped, because it converges towards 1 in the limit $n \to 0$ (because the existence of $\lim\limits_{n\to 0} f_n$ implies the existence of $\lim\limits_{n\to 0} \frac{1}{n} \ln \mathrm{Tr} \exp(-\beta \mathcal{H}_{\text{eff}})$). It is crucial to understand that equation (5.17) states that the saddle point components are exactly equal to the order parameters. Thus, we can identify the order parameters with the components of the saddle point, and omit the tilde in $\widetilde{q}$ in further calculations. However, this fact is not surprising, because this equality is characteristic of MFT. More generally, the order parameters can be obtained from the saddle point equation:

$$\frac{\partial f}{\partial q_{\alpha\beta\cdots\delta\gamma}} = 0 \qquad \Longrightarrow \qquad q_{\alpha\beta\cdots\delta\gamma} = \lim_{n\to 0} \langle S^\alpha S^\beta \cdots S^\delta S^\gamma \rangle_{\mathcal{H}_{\text{eff}}}. \tag{5.18}$$

From this equation we can also conclude that high-order spin correlations occur in the spin glasses. Furthermore, we obtain an infinite set $\{q_\alpha, q_{\alpha\beta}, q_{\alpha\beta\gamma}, q_{\alpha\beta\gamma\delta}, \ldots\}$ of order parameters describing the possible phases of the spin glass.

We have seen that for positive integers n, it is relatively simple to carry out the quenched average $\langle \cdot \rangle_{\mathrm{J}}$, and that one can express the result formally in terms of an effective Hamiltonian $\mathcal{H}_{\text{eff}}(n)$ which no longer contains any disorder and which is translationally invariant. Moreover, since $\mathcal{H}_{\text{eff}}(n)$ is obviously symmetrical under the permutation of the indices of the replicas, the replica symmetry ansatz seems to be justified. Unfortunately, this no longer remains true when n takes noninteger values to reach the limit $n \to 0$; we can thus expect the order parameters to take values dependent on this set of the indices. This fact leads to the natural idea of breaking the replica symmetry. Nevertheless, the replica symmetry approximation is expected to be a good approximation in the vicinity of the phase boundary. Thus, we will make use of this assumption in our further developments.

5.4 Phase boundaries

The different phases of a spin glass are determined by the correlations between the spins, which come into existence for nonvanishing order parameters. Thus, in order to find the phase boundaries, one needs to determine their explicit values. This can be achieved by making a cumulant expansion of the exponential factor in equation (5.17) for small $\mathcal{H}_{\text{eff}}$. This limits our calculation to a region above or close to the critical temperature, where by definition the $q_\alpha, q_{\alpha\beta}, q_{\alpha\beta\gamma}, \ldots$ are small. The magnitude of the homogeneous field h also has to remain negligible.

For the case $h = 0$, expanding equation (5.18) to order ${\mathcal{H}_{\text{eff}}}^2$ provides an infinite set of equations for the values of the order parameters. The condition $q_\alpha \neq 0$ for a nonvanishing magnetisation yields

$$\mathsf{T}_1(T = T_F) = \frac{1}{p}, \tag{5.19}$$

where the temperature T_F defines the upper boundary of the *ferromagnetic phase*. Moreover, one obtains

$$\mathsf{T}_2(T = T_{SG}) = \frac{1}{p}, \tag{5.20}$$

as well as similar equations for higher order phases. The temperature T_{SG} denotes the critical temperature for the *spin glass phase*. The highest of the resulting transition temperatures will give the corresponding boundary for the *paramagnetic phase* that refers to the disordered phase of noncorrelated spins. These equations obviously have no solutions for $p < 1$ and the MFT predicts no order at all in this case. Below this threshold concentration $p_c = 1$, the spins interact too weakly and can be considered isolated clusters.

Furthermore, the only phase transitions that can occur first are the ferromagnetic or the spin glass transition. This follows directly from the inequality

$$\mathsf{T}_{2m+l} - \mathsf{T}_{2m} = \int_{-\infty}^{\infty} \mathrm{d}J_{ij} P_p(J_{ij}) \tanh^{2m}(\beta J_{ij})\big(\tanh^{l}(\beta J_{ij}) - 1\big) < 0$$

for m, $l \geq 1$ (and $n \to 0$). Therefore, the lines $p\mathsf{T}_1 = 1$ and $p\mathsf{T}_2 = 1$ give us the boundaries with the paramagnetic phase, that is, the *paramagnetic/ferromagnetic* (P/F) boundary and the *paramagnetic/spin glass* (P/SG) boundary, respectively.

5.5 Stability of the phases

The correctness of the phase boundaries has to be corroborated by the stability of the phase close to its boundary. We must also prove that solutions provided by the saddle point equation (5.18) minimise the free energy. Consequently, one has to investigate the behaviour of the free energy around the saddle point.

5.5.1 High-order corrections

The study of the stability of the solutions close to the lines $p\mathsf{T}_1 = 1$ and $p\mathsf{T}_2 = 1$ must include high-order order parameters. Fortunately, the saddle point equation allows us to express the corrections to the order parameters in terms of the order parameters q_α and $q_{\alpha\beta}$ of the lowest order. This way, one can take into account the contribution to the free energy of the first four order parameters. We will limit ourselves to the reasonable approximation

$$f(\mathbf{q}) \approx f(q_\alpha, q_{\alpha\beta}, q_{\alpha\beta\gamma}, q_{\alpha\beta\gamma\delta}).$$

Before continuing, we shall introduce for later convenience the following variables:

$$t_\alpha = \beta h + p\mathsf{T}_1 q_\alpha, \qquad Q_{\alpha\beta} = p\mathsf{T}_2 q_{\alpha\beta},$$
$$Q_{\alpha\beta\gamma} = p\mathsf{T}_3 q_{\alpha\beta\gamma}, \qquad Q_{\alpha\beta\gamma\delta} = p\mathsf{T}_4 q_{\alpha\beta\gamma\delta}.$$

These changes of variable induce new equations, for instance,

$$\frac{\partial f}{\partial q_\alpha} = 0 \qquad \Longleftrightarrow \qquad p\mathsf{T}_1 \left(\frac{\partial f}{\partial t_\alpha}\right) = 0.$$

Finally, we shall often make use of the set of variables defined as

$$r_m = \frac{1}{p\mathsf{T}_m} - 1.$$

In order to find the corrections of the third order, we may now apply the saddle point equation (5.18). Expanding the exponential factor of $\ln \mathrm{Tr}_S \exp(-\beta\mathcal{H}_{\text{eff}})$ to order ${\mathcal{H}_{\text{eff}}}^3$ and evaluating the trace over the n spins S^α gives

$$Q_{\alpha\beta\gamma} = \frac{1}{2r_3}\big(t_\alpha Q_{\beta\gamma} + t_\beta Q_{\alpha\gamma} + t_\gamma Q_{\alpha\beta}\big) \; + \; \text{higher order terms}, \qquad (5.21)$$

where the replicas' indices α, β, γ are pairwise different.

The same procedure for the order parameter $Q_{\alpha\beta\gamma\delta}$ also requires the expansion to order ${\mathcal{H}_{\text{eff}}}^3$, and for pairwise different indices gives the equation

$$Q_{\alpha\beta\gamma\delta} = \frac{1}{8r_4} \sum_{\sigma \in S_4} Q_{\sigma(\alpha)\sigma(\beta)} Q_{\sigma(\delta)\sigma(\gamma)} \; + \text{ higher order terms,} \tag{5.22}$$

where S_4 denotes the group of permutations of four elements.

In order to determine the behaviour of the free energy in the vicinity of the critical temperature, we may explicitly compute f_n from its definition

$$f = \lim_{n\to 0} \frac{1}{n} f_n,$$

from equation (5.13). After quantifying the contribution of the order parameters $Q_{\alpha\beta\gamma}$ and $Q_{\alpha\beta\gamma\delta}$ and after some manipulations, we obtain at order n the equation

$$\begin{aligned} f_n &= (2p\mathsf{T}_1)^{-1} \sum_{\alpha} (t_\alpha - \beta h)^2 - \frac{1}{2} \sum_{\alpha} t_\alpha^2 + \frac{r_2}{4} \sum_{\alpha\beta} Q_{\alpha\beta}^2 \\ &- \frac{1}{6} \sum_{\alpha\beta\gamma} Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\alpha} - \frac{1}{2} \sum_{\alpha\beta} Q_{\alpha\beta} t_\alpha t_\beta + \frac{1}{2}(1 + \frac{2}{r_3}) \sum_{\alpha\beta} Q_{\alpha\beta}^2 t_\alpha^2 + \frac{1}{12} \sum_{\alpha} t_\alpha^4 \\ &- \frac{1}{12}(1 + \frac{3}{r_4}) \sum_{\alpha\beta} Q_{\alpha\beta}^2 + \frac{1}{4}(1 + \frac{2}{r_4}) \sum_{\alpha\beta\gamma} Q_{\alpha\beta}^2 Q_{\alpha\gamma}^2 - \frac{1}{8}(1 + \frac{1}{r_4}) \sum_{\alpha\beta\gamma\delta} Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\delta} Q_{\delta\alpha} \\ &- \frac{1}{2}(1 + \frac{1}{r_3}) \sum_{\alpha\beta\gamma} Q_{\alpha\beta} Q_{\beta\gamma} t_\alpha t_\beta \; + \text{ higher order terms.} \end{aligned} \tag{5.23}$$

We define the $Q_{\alpha\beta}$ to be symmetric, i.e., $Q_{\alpha\beta} = Q_{\beta\alpha}$, with the restriction $Q_{\alpha\alpha} = 0$, so that the sums over $\alpha < \beta$ can be replaced by the free sums as occurring in equation (5.23). We have omitted the terms that vanish in the limit $n \to 0$. This can only be achieved by considering that, for instance,

$$\lim_{n\to 0} \frac{1}{n} \sum_{\alpha\neq\beta} t_\alpha t_\beta = \lim_{n\to 0} n(\frac{1}{n} \sum_{\alpha} t_\alpha)^2 - \lim_{n\to 0} \frac{1}{n} \sum_{\alpha} t_a^2 = - \lim_{n\to 0} \frac{1}{n} \sum_{\alpha} t_\alpha^2,$$

where the sum runs over the n replicas. We can also emphasise that some of the couplings of the order parameters arise from the equation

$$\sum_{\alpha\neq\beta\neq\delta\neq\gamma} Q_{\alpha\beta}^2 Q_{\delta\gamma}^2 = \sum_{\alpha\beta\delta\gamma} Q_{\alpha\beta}^2 Q_{\delta\gamma}^2 + \sum_{\alpha\beta} Q_{\alpha\beta}^4 - 2 \sum_{\alpha\beta\gamma} Q_{\alpha\beta}^2 Q_{\beta\gamma}^2.$$

Unfortunately, the solutions that minimise f_n are not easy to find. We need an appropriate assumption to determine the values of the order parameters. This assumption is the previously mentioned replica symmetric ansatz.

5.6 Replica symmetric solutions

We may now assume the order parameters to be invariant under permutation of their replica indices. Formally, this is written

$$t_\alpha = t + R_\alpha \qquad \text{and} \qquad Q_{\alpha\beta} = Q + R_{\alpha\beta},$$

where t and Q are the mean field solutions and R_α and $R_{\alpha\beta}$ their fluctuations. The investigation of the behaviour of the free energy requires us to expand equation (5.23) around the minima t and Q of f_n to the second order. If we keep only the nonvanishing terms in the limit $n \to 0$, this procedure yields the following equation:

$$\begin{aligned} f_n &= L\sum_\alpha R_\alpha + M\sum_{(\alpha,\beta)} R_{\alpha\beta} \\ &+ A\sum_\alpha R_\alpha^2 + B\sum_{(\alpha,\beta)} R_\alpha R_\beta + C\sum_{(\alpha,\beta)} R_\alpha R_{\alpha\beta} + D\sum_{\alpha\neq\beta\neq\gamma} R_\gamma R_{\alpha\beta} \\ &+ P\frac{1}{2}\sum_{(\alpha,\beta)} R_{\alpha\beta}^2 + Q\sum_{\substack{(\beta,\gamma)\\ \alpha}} R_{\alpha\beta}R_{\alpha\gamma} + R\sum_{\alpha\neq\beta\neq\gamma\neq\delta} R_{\alpha\beta}R_{\gamma\delta}, \end{aligned} \tag{5.24}$$

where (α,β) denotes distinct pairs with $\alpha \neq \beta$.

The conditions $L = 0$ and $M = 0$ for the mean field order parameters t and Q are fulfilled by demanding[1] that

$$L = p\mathsf{T}_1\Big(r_1 t - \frac{\beta h}{p\mathsf{T}_1} + Qt + \frac{1}{3}t^3 - (2+\frac{3}{r_3})Q^2 t\Big) + O(Q^3 t, Qt^3) = 0 \tag{5.25}$$

and

$$M = p\mathsf{T}_2\Big(r_2 Q + 2Q^2 - t^2 + (4+\frac{6}{r_3})Qt^2 - (\frac{17}{3}+\frac{9}{r_4})Q^3\Big) + O(t^4, Q^2t^2) = 0. \tag{5.26}$$

The factors of the quadratic terms are

$$A = (p\mathsf{T}_1)^2\Big(\frac{r_1}{2} + \frac{1}{2}Q + \frac{1}{2}t^2 - (1+\frac{3}{2r_3})Q^2\Big) + \text{higher order terms,} \tag{5.27}$$

$$C = (p\mathsf{T}_1)(p\mathsf{T}_2)t\big(-1 + (4+\frac{6}{r_3})Q\big) + \text{higher order terms,} \tag{5.28}$$

$$P = (p\mathsf{T}_2)^2\Big(\frac{r_2}{2} + \frac{1}{r_3}t^2 + +\frac{1}{2}(3+\frac{1}{r_4})Q^2\Big) + \text{higher order terms,} \tag{5.29}$$

$$Q = (p\mathsf{T}_2)^2\Big(-\frac{1}{2}Q - \frac{1}{2}(1+\frac{1}{r_3})t^2 + \frac{1}{2}(5+\frac{7}{r_4})Q^2\Big) + \text{higher order terms,} \tag{5.30}$$

and trivially

$$B = D = R = 0. \tag{5.31}$$

[1] correcting an obvious mistake of sign in [ViBr85].

The saddle point can only be considered as an MFT stable solution if the free energy is minimised at this point. Since the Hessian matrix whose components are defined in equations (5.27–31) determines the fluctuations of the free energy around the saddle point, the investigation of its eigenvalues in the limit $n \to 0$ is necessary.

5.7 Stability of the MFT solutions

We now aim at finding the eigenvalues of the previously introduced Hessian matrix. According to [AlTh78], one can take advantage of its particular form. Furthermore, since the fluctuations $R_{\alpha\beta}$ are symmetrical under the permutations of their indices, we shall consider vectors of the form

$$\boldsymbol{\mu} = \begin{pmatrix} \{R_\alpha\} \\ \{R_{\alpha\beta}\} \end{pmatrix} \quad \text{where } \alpha, \beta \in \{1, 2, \dots, n\} \text{ and } \alpha < \beta.$$

The column-vectors $\{R_\alpha\}$ and $\{R_{\alpha\beta}\}$ have n and $\frac{1}{2}n(n-1)$ elements, respectively. The real Hessian matrix is symmetric, and hence the number of linearly independent eigenvectors is $\frac{1}{2}n(n+1)$. The eigenvectors are distributed among three classes which happen to contain the complete set of eigenvectors.

(a) The components of the first eigenvectors are defined as

$$R_\alpha = a \quad \text{and} \quad R_{\alpha\beta} = b \quad \text{for all } \alpha, \beta.$$

The eigenvalue equation gives the two nondegenerate eigenvalues

$$\begin{aligned} \lambda_\pm &= \frac{1}{2}\big(A + (n-1)B + P + 2(n-2)Q + \frac{1}{2}(n-2)(n-3)R\big) \\ &\pm \frac{1}{2}\Big(\big(A + (n-1)B - P - 2(n-2)Q - \frac{1}{2}(n-2)(n-3)R\big)^2 \\ &+ 2(n-1)\big(2C + (n-2)D\big)^2\Big)^{\frac{1}{2}}. \end{aligned}$$

Thus, in the limit $n \to 0$, we obtain the final eigenvalues

$$\begin{aligned} \lambda_\pm = \frac{1}{2}\Big(&(A - B + P - 4Q + 3R) \\ &\pm\big((A - B - P + 4Q - 3R)^2 - 8(C - D)^2\big)^{\frac{1}{2}}\Big). \end{aligned} \tag{5.32}$$

(b) Next, we shall consider the vectors of components

$$R_\theta = a \quad \text{and} \quad R_\alpha = b \quad \text{for } \alpha \neq \theta,$$

and

$$R_{\alpha\beta} = c \quad \text{for } \alpha = \theta \text{ or } \beta = \theta \quad \text{and} \quad R_{\alpha\beta} = d \quad \text{for } \alpha, \beta \neq \theta\ ,$$

where $\theta \in \{1, 2, \ldots, n\}$. The orthogonality to the previously mentioned eigenvectors can be ensured by demanding that $a = (1-n)b$ and $c = (1-\frac{1}{2}n)d$. The substitution in the eigenvalue equation provides two further eigenvalues

$$\begin{aligned}\widetilde{\lambda}_{\pm} = &\frac{1}{2}\Big(A - B + P + (n-4)Q - (n-3)R\Big)\\ &\pm\frac{1}{2}\Big(\big(A - B - P - (n-4)Q - (n-3)R\big)^2 + 4(n-2)(C-D)^2\Big)^{\frac{1}{2}}.\end{aligned}$$

Nevertheless, note that these eigenvalues with degeneracy $n-1$ no longer remain distinct from the eigenvalues $\lambda_{\pm}$ in the limit $n \to 0$.

(c) Finally, the remaining eigenvectors can be expressed as

$$R_\alpha = a \quad \text{for } \alpha = \theta, \phi \quad \text{and} \quad R_\alpha = b \quad \text{for } \alpha \neq \theta, \phi\ ,$$

and

$$R_{\theta\phi} = c,\ R_{\alpha\phi} = R_{\alpha\theta} = d \quad \text{for } \alpha \neq \theta, \phi \quad \text{and} \quad R_{\alpha\beta} = e \quad \text{for } \alpha, \beta \neq \theta, \phi\ ,$$

where $\theta, \phi \in \{1, 2, \ldots, n\}$. The orthogonality to the previous eigenvectors imposes the conditions $a = b = 0$, $c = (2-n)d$ and $d = \frac{1}{2}(3-n)e$. This set of eigenvectors spans an invariant subspace of dimension $\frac{1}{2}(3-n)$ and gives a unique eigenvalue

$$\lambda = P - 2Q + R. \tag{5.33}$$

In the limit $n \to 0$, one finally obtains three distinct eigenvalues λ and $\lambda_{\pm}$. Their values determine whether the Hessian matrix is positive definite, and consequently the stability of the solution of the saddle point equation. Beyond the analysis of the phase boundaries lines $p\mathsf{T}_1 = 1$ and $p\mathsf{T}_2 = 1$, the stability of the boundaries also has to be investigated.

5.8 The harmony network

The harmony network can be compared with the previously developed spin glass model. In order to use MFT as a statistical method applied to a network, we have exploited both the clustered structure of the network and the strong interactions among its nodes. Furthermore, as MFT handles the macroscopical behaviour of physical models in the limit of large systems, the analysis of the critical phenomena shall be restricted to relatively large networks. Only in this case can the networks generated by the NEULA compiler be expected to exhibit a self-averaging behaviour.

In Section 5.4, we have seen that the theoretical analysis of the phase boundaries can be carried out in the case of a system with a weak and homogeneous field h. However, the local fields of the harmony networks, resulting from the compilation of a

NEULA description code, cannot be assumed to be weak nor independent of the nodes, as shown in Chapter 4. Henceforth, we must restrict ourselves to the investigation of harmony networks for which the local fields have to be set to zero. In our model, this simply means taking $h = 0$. To put it concretely, this intrinsic limitation of the theory renders the analytical results significant solely for harmony networks that are nonrealistic in terms of Bayesian reasoning. Unfortunately, this is the price we have to pay to use statistical mechanics.

At this point, let us discuss the distribution of the nonzero bonds of those harmony networks we studied extensively in Chapter 4. It has been shown that the frequency distribution of the magnitudes of the various bonds peaks strongly at two symmetrical values, namely, at the minimum $-J$ and at the maximum J of these magnitudes. Formally, this distribution can be approximated reasonably well by

$$P_p(J_{ij}) = a\delta(J_{ij} + J) + (1 - a)\delta(J_{ij} - J), \tag{5.34}$$

where the constant $J > 0$ represents the maximal value as introduced above. The parameter $a \in [0, 1]$ represents the asymmetry between the frequency of the negative and the positive interactions. As a matter of fact, the harmony networks generated by means of the NEULA compiler are particularly asymmetric, and according to our experience, in most of the cases, the negative bonds seems to prevail. Despite this observation, we shall not limit the networks to be studied to those originally generated from a given NEULA description code. We will randomly redistribute the sign of the originally existing interactions according to a desired value asymmetry parameter a. In this manner, we hope to be able to explore all the possible transitions from the paramagnetic phase to a less disordered phase.

The behaviour of the lines $p\mathsf{T}_1 = 1$ and $p\mathsf{T}_2 = 1$ depends on the specific bond distribution $P_p(J_{ij})$ previously introduced. Before determining the resulting phase boundaries, we note that

$$\lim_{n\to 0} \mathsf{T}_m = \langle \tanh^m(\beta J_{ij}) \rangle = \begin{cases} \tanh^m(\beta J), & \text{for } m \text{ even;} \\ (1 - 2a)\tanh^m(\beta J), & \text{for } m \text{ odd.} \end{cases} \tag{5.35}$$

This implies that the existence of the SG phase boundary is tied with the condition $a < 1/2$. Provided this condition is fulfilled for a fixed parameter $p > 1$, the system can present a crossover of the two lines defining the P/SG and the P/F boundaries. The multicritical point (a^*, T^*) is defined as the point at which the paramagnetic, the ferromagnetic and the spin glass phases meet. Consequently, the value of the asymmetry parameter a is relevant in determining the kind of phase transition to be expected: either a ferromagnetic or a spin glass phase transition.

By cooling it down, we expect a harmony network exhibiting the asymmetry $a > a^*$ first to enter a spin glass phase from the paramagnetic phase (a P/SG transition). The transition temperature T_{SG} can be computed from equation (5.35). Indeed, as expected, T_{SG} happens to be higher than the corresponding transition temperature

Table 5.1: Characteristics of the two large harmony networks used for the simulated annealing. The number of feature nodes and pattern nodes are denoted by F and P, respectively. L refers to the number of nonzero arcs of the network.

Network	F	P	L	J	p	a^*	T_{SG}/J
Zoo 1	17	126	423	15.54	5.96	0.295	2.3
Zoo 2	47	546	4682	15.54	15.82	0.37	3.89

T_F. The equation $p\mathsf{T}_2 = 1$ yields

$$\frac{T_{SG}}{J} = \frac{1}{\operatorname{arctanh}\left(p^{-\frac{1}{2}}\right)}, \quad \text{for } a \geq a^*, \tag{5.36}$$

where

$$a^* \equiv \frac{1}{2}(1 - p^{-\frac{1}{2}}). \tag{5.37}$$

The P/SG boundary temperature in equation (5.36) appears to be independent of the asymmetry parameter a, and hence, regardless of the value of a^*, we can state that $T^* = T_{SG}$.

On the other hand, the P/F boundary temperature T_F, following from the equation $p\mathsf{T}_1 = 1$, depends on the asymmetry parameter of the network and is given by

$$\frac{T_F}{J} = \frac{1}{\operatorname{arctanh}\left((p - 2ap)^{-1}\right)}, \quad \text{for } a < a^*. \tag{5.38}$$

For later convenience, we shall denote T_{crit} to be the transition temperature. According to the first phase to be reached by the system during its cooling, T_{crit} will be either T_{SG} or T_F. In order to analyse these results, we shall consider two harmony networks of different sizes. These networks, described in Table 5.1, have been generated from two different NEULA description codes, and they appear to be sufficiently large for our purpose.

But before we conclude the existence of the described phase boundaries, we have to investigate carefully the region of stability of the phases in question. For this purpose, as we intend to carry out the analysis of these regions in the vicinity of the $p\mathsf{T}_m = 1$ lines, the expansion

$$p\mathsf{T}_m = 1 + \Delta_m$$

is an appropriate choice.

5.8.1 The paramagnetic phase

The paramagnetic phase is characterised by $t = 0$ and $Q = 0$. Trivially, the saddle point conditions (5.25) and (5.26) are fulfilled. The question of the stability of this phase comes down to the study of the signs of the three eigenvalues $\lambda_\pm$ and λ_1. The only nonzero matrix elements in equation (5.24) are A and P. This provides the following conditions for the stability of the phase:

$$\lambda_+ = A = \frac{r_1}{2}\,(p\mathsf{T}_1)^2 > 0 \quad \Longleftrightarrow \quad p\mathsf{T}_1 < 1$$

and

$$\lambda_- = \lambda_1 = P = \frac{r_2}{2}\,(p\mathsf{T}_2)^2 > 0 \quad \Longleftrightarrow \quad p\mathsf{T}_2 < 1.$$

It is easy to see from equation (5.35) that the entire region where $T > T_{crit}$ corresponds to the domain that satisfies both of the previous conditions, and that this domain is delimited by the phase boundaries lines $p\mathsf{T}_1 = 1$ and $p\mathsf{T}_2 = 1$. Therefore, the entire paramagnetic phase can be considered stable with regard to the saddle point approximation.

5.8.2 The spin glass phase

The existence of other solutions to the equations giving the extrema of the free energy corresponds to the existence of a spin glass phase. This phase is defined as having a vanishing per-spin magnetisation, which means $t = 0$, but having a nontrivial order parameter $Q \neq 0$. By solving the saddle point equations close to the line $p\mathsf{T}_2 = 1$, we obtain

$$Q = \frac{\Delta_2}{2} + \left(\frac{5}{24} + \frac{9}{8r_4}\right)\Delta_2^2 + O(\Delta_2^3),$$

which corresponds to the solution that goes to zero in the limit $\Delta_2 \to 0$. Furthermore, for the following eigenvalues at the saddle point, we obtain

$$\lambda_- = \frac{\Delta_2}{2} + O(\Delta_2^2) \qquad \text{and} \qquad \lambda_+ = (p\mathsf{T}_1)^2\frac{1}{2}\big(r_1 + \frac{\Delta_2}{2} + O(\Delta_2^2)\big).$$

It follows that $\lambda_- > 0$ for $p\mathsf{T}_2 > 1$, and $\lambda_+ > 0$ for $p\mathsf{T}_1 < 1$. So far, this suggests that the SG phase is stable in the vicinity of the phase boundary P/SG.

Finally, the remaining eigenvalue can be expressed as

$$\lambda_1 = -\frac{1}{8}(7 + \frac{13}{r_4})\Delta_2^2 + O(\Delta_2^3),$$

where we can insert

$$r_4 = \frac{p}{(p\mathsf{T}_2)^2} - 1 = (p-1) - 2\Delta_2 + 3p\Delta_2^2 + O(\Delta_2^3),$$

Table 5.2: Critical temperatures of the network Zoo 1 with different asymmetry factors. The multicritical point $a^* = 0.29$ determines the upper limit for a phase transition from the paramagnetic phase to the ferromagnetic phase (P/F boundary).

Asymmetry a	P/F boundary			P/SG boundary					
	0.0	0.1	0.2	0.3	0.4	0.6	0.8	0.9	1.0
Theoretical T_{crit}	91.7	73.0	54.1	35.7	35.7	35.7	35.7	35.7	35.7
Experimental T_{crit}	≈ 40	≈ 30	≈ 27	≈ 14	< 10	< 10	≈ 20	≈ 29	≈ 32

directly derived from the distribution P_p of the bonds. This shows, since $\lambda_1 \leq 0$ holds everywhere in the parameter space, that the SG phase is unstable. At this point, the replica symmetric ansatz breaks down. As previously mentioned, the correct solution should involve the breaking of the replica symmetry. Fortunately, the solution $\lambda_1 = 0$ coincides with the boundary line $p\mathsf{T}_2 = 1$; hence the solutions along the boundary line are marginally stable. This implies the correctness of the boundary P/SG, although the theory fails to predict the stability of the SG phase.

5.8.3 The ferromagnetic phase

The ferromagnetic phase is characterised by nonzero order parameters $t \neq 0$ and $Q \neq 0$. The solutions provided by the saddle point equation can be written as

$$Q = \frac{3}{3 + r_2}\Delta_1 + O(\Delta_1^2) \qquad \text{and} \qquad t^2 = Qr_2 + O(\Delta_1^2),$$

in the region far from the multicritical point, where $|\Delta_2| \gg |\Delta_1|$. On the other hand, close to the multicritical point, we obtain

$$Q = \Delta_1 + O(\Delta^2) \qquad \text{and} \qquad t^2 = 2\Delta_1^2 - \Delta_1\Delta_2 + O(\Delta^3).$$

This yields the eigenvalues

$$\lambda_- = (p\mathsf{T}_2)^2 r_2 + O(\Delta_1) \qquad \text{and} \qquad \lambda_+ = \frac{3 + 5r_2}{3 + r_2}\Delta_1 + O(\Delta_1^2).$$

Similarly to the spin glass phase, both eigenvalues are positive in the vicinity of the boundary line. Moreover, if we set $\Delta_1 = 0$, then we obtain $\lambda_+ = 0$, and this indicates the correctness of the P/F boundary.

Table 5.3: Critical temperatures of the network Zoo 2 with different asymmetry factors. The predicted temperatures for the P/SG boundaries are clearly below the experimental temperatures. The critical temperatures exhibits a strong dependence on the asymmetry factor $a > a^*$.

Asymmetry a	P/F	P/SG boundary					
	0.0	0.5	0.6	0.7	0.8	0.82	0.9
Theoretical T_{crit}	245.5	60.5	60.5	60.5	60.5	60.5	60.5
Experimental T_{crit}	≈ 109	< 10	≈ 60	≈ 90	≈ 110	≈ 115	≈ 130

5.8.4 Results and discussion

Below the temperature T_{crit}, the totally disordered paramagnetic phase does not prevail any more. Although the replica symmetric theory fails to predict the phase of the system, this theory ensures the existence of an ordering characterised by the onset of nonzero high-order correlation between two or more spins. We shall now attempt to corroborate, by means of the computational simulation, the theoretical bound for the critical temperature T_{crit}. The simulated annealing has been carried out with extremely long delay between the decreasing steps of the temperature. This way, the quasi-equilibrium is likely to be maintained throughout the cooling process.

The results obtained by the annealing of the network Zoo 1 are summarised in Table 5.2. As can be seen, the theoretical results clearly overestimate the experimentally found temperatures. Without any doubts, the simplifying assumption made about the bond distribution P_p may affect the predictions of the theory. If one supposes the magnitudes of the interactions are between the two values $-J$ and J, it can easily be shown that

$$\langle \tanh^{2k}(\beta J_{ij}) \rangle \leq \tanh^{2k}(\beta J), \quad \text{implying} \quad T_{SG}^{\text{eff}} \leq T_{SG}^{\text{th}}.$$

In reality, the interactions are strongly located around their minimal or maximal values. Obviously, this causes the temperature to be only slightly overestimated. However, the results for the network Zoo 2 in Table 5.3 show that this argument is not sufficient in general, because they clearly establish that the critical temperatures are higher than those theoretically predicted. Surprisingly, this phenomenon appears to be particularly significant for the P/SG boundary.

We also notice that in both cases, the effective transition temperature T_{crit} strongly varies with the values of the asymmetry parameter a, in spite of the theoretical prediction that states that the transition temperature on the P/SG phase boundary is constant. A more precise investigation of the phase occurring below T_{SG} establishes

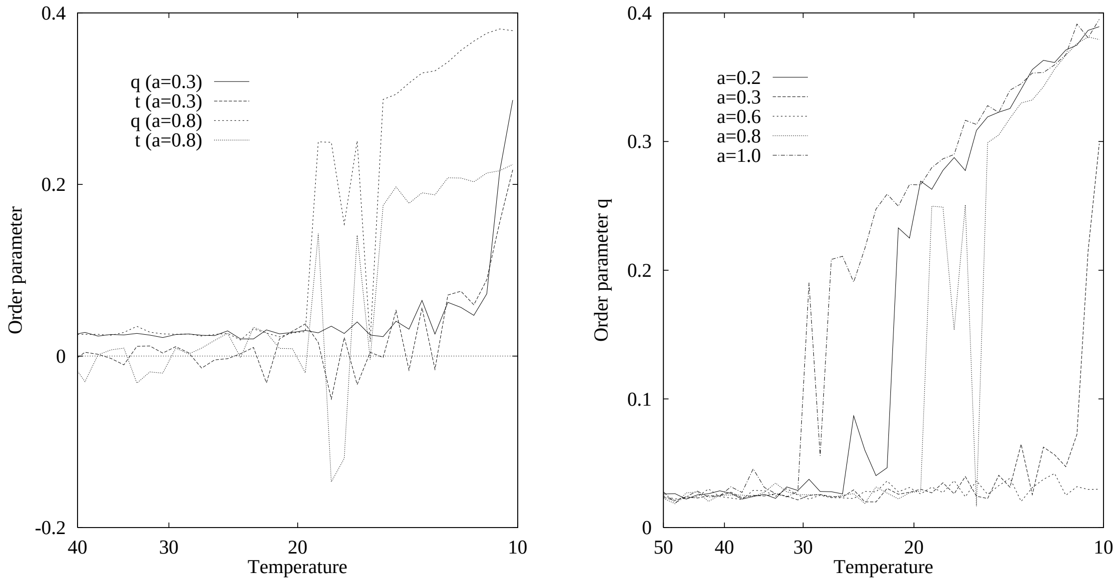


Figure 5.1: Typical behaviours of the network Zoo 1 resulting from simulated annealing processes. In the left-hand graph, the simultaneous behaviour of the two order parameters t and q, as a function of the temperature, are depicted for two different values of the asymmetry: $a = 0.3$ and $a = 0.8$. In the right-hand graph, the evolution of the order parameter q is represented for various values of the asymmetry parameter a.

that the spin glass phase, characterised by $t = 0$, does not appear at all (see Figure 5.1). Conceivably, the breakdown of the replica symmetric theory prevents us from predicting the dominant phase that should arise below T_{crit}. Consequently, the nonexistence of a SG phase does not contradict the theory at all.

One of the basic statements justifying the validity of MFT is the extensively discussed self-averaging assumption. We have attempted to bring to the fore the stated sample independence of the behaviour of the network. As an example, the right-hand graph of Figure 5.2 gives evidence that the onset of the boundary remains independent of the random distribution of the interaction signs for a given fixed parameter a. Apart from this experimental justification of the self-averaging assumption, this graph also highlights the fact that the sign distribution is not generated randomly by the NEULA compiling process. Indeed, the transition temperature of the originally compiled network differs significantly from that of a similar network, differing only in its sign distribution, but keeping the original asymmetry parameter a.

As a matter of fact, the major anomalies in the behaviour of the networks can be explained by the specific topology of the class of harmony networks under consideration. The strongly asymmetric bipartitional structure of the harmony networks fails to be statistically equivalent to the considered spin glass model, especially for large networks.

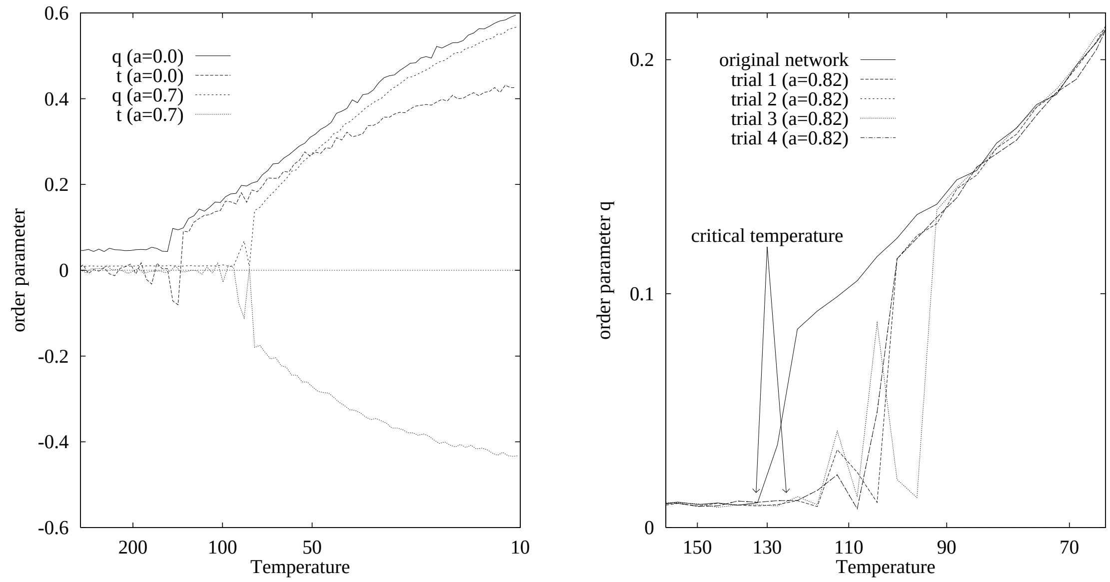


Figure 5.2: Examples of some annealing processes simulated on the network Zoo 2. In the left-hand graph, the behaviour of the two order parameters t and q during the cooling process shows that the prevailing phase below the transition temperature T_{crit} typically exhibits a nonzero per-spin magnetisation. In the right-hand graph, the asymmetry parameter of the originally compiled network Zoo 2 has been kept constant, but the interaction signs of the network have been randomised according to this constant value of $a = 0.82$. The only discordant result arises from the original network, where the sign distribution has not yet been randomised.

First, because of the relatively small number of feature nodes compared to the large number of pattern nodes sharing the same number of links, every feature node can be expected to be much more highly connected than every pattern node. Consequently, the resulting specific distribution of the number of interactions per node is not uniform, unlike the case for the spin glass model. Second, another piece of evidence for this disagreement is that the spin glass model, even when strongly diluted, can exhibit so-called closed loops [Toul77] of interacting spins, regardless of the parity of the number of connections composing these loops. But, in harmony network, only loops consisting of an even number of connected nodes can appear. In conclusion, the major part of the observed discrepancies can apparently be ascribed to these slight deviations from the most general distribution P_p of the interactions.

By and large, we have seen that the studied spin glass model differs slightly from the harmony network model. They are distinguished both by the underlying Bayesian reasoning and the bipartitional topology of the most general harmony networks. However, we would like to emphasise that MFT seems to offer an interesting theoretical framework for the study of more general sparsely connected Boltzmann machines. Without any doubts, the replica method can bring more accurate criteria than the prevailing empirical methods into the task of trapping the critical phenomena.

Bibliography

[AcHS85] D.H. Ackley, G.E. Hinton and T.J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science* **9** (1985), 147–169.

[Amit89] D.J. Amit, *Modeling Brain Function*, Cambridge University Press, Cambridge, 1989.

[AaKo89] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*, John Wiley & Sons, England, 1989.

[AlTh78] J.R.L. de Almeida and D.J. Thouless, "Stability of the Sherrington-Kirkpatrick solution of a spin glass model," *J. Phys. A: Math. Gen.* **11** (1978), 983–990.

[Bind80] K.Binder , "Statistical mechanics of Ising spin glasses," *Fundamental problems in Statistical Mechanics* V, E.G.D. Cohen (ed.), North-Holland, 1980.

[BiYo86] K. Binder and A. Young, "Spin glasses: experimental facts, theoretical concepts, and open questions," *Rev. Mod. Phys.* **58** (1986), 801–976.

[Bind87] K. Binder, *Applications of the Monte Carlo Method in Statistical Physics*, Topics in Current Physics, Springer, Berlin, 1987.

[BiHe88] K. Binder and D.W. Heerman, *Monte Carlo Simulation in Statistical Physics*, Springer Series in Solid-State Sciences, Berlin, 1988.

[BrMo78] A.J. Bray and M.A. Moore, "Replica-symmetry breaking in spin-glass theories," *Phys. Rev. Let.* **41** (1978), 1068–1079.

[ChHo93] M. Christoph and K.H. Hoffman, "Scaling behaviour of optimal simulated annealing schedules," *J. Phys. A* **26** (1993), 3267–3277.

[Fell68] W. Feller, *An Introduction to Probability Theory and Its Applications* Vol. I, John Wiley & Sons, New-York, 1968.

[FiHe91] K.H. Fischer and J.A. Hertz, *Spin Glasses*, Cambridge University Press, Cambridge, 1991.

[GeGe84] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (1984), 721–741.

[Glau63] R. Glauber, Time-dependent statistics of the Ising model, *J. Math. Phys.* **4** (1963), 294–307.

[HeKP91] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, California, 1991.

[HePa82] J.L. van Hemmen and R.G. Palmer, *J. Phys. A* **15** (1982), 3881.

[HiSe83] G.E. Hinton and T.J. Sejnowski, "optimal perceptual inference," *Proc. Conference on Computer Vision and Pattern Recognition* , IEEE, Washington DC (1983), 448–453.

[Jain92] S. Jain, *Monte Carlo Simulations of Disordered Systems*, World Scientific, Singapore, 1992.

[KaHa90] Y. Kamp, M. Hasler, *Réseaux de Neurones Récursifs pour Mémoires Associatives*, Presses polytechniques et universitaires romandes, Lausanne, 1990.

[KaSo87] I. Kanter, H. Sompolinsky, "Mean-Field theory of spin-glasses with finite coordination number," *Phys. Rev. Lett.* **58** (1987), 164–167.

[KiSe78] S. Kirkpatrick and D. Sherrington, "Infinite-ranged models of spin-glasses," *Phys. Rev. B* **17** (1978), 4384–4403.

[Krog89] A. Krogh, *The Boltzmann Machine*, Lecture notes for the Second Nordic Summer School on Neural Modeling and Computation, 1989.

[LaAa87] P.J.M. van Laarhoven and E.H.L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel Publishing Company, Dordrecht, The Netherland, 1987.

[MiBH86] A. Milchev, K. Binder and D.W. Heermann, "Fluctuations and Lack of Self-Averaging in the Kinetics of Domain Growth," *Z. Phys. B* **63** (1986), 521–535.

[MyOr91] P. Myllymäki and P.Orponen, "Programming the Harmonium," *Proc. International Conference on Neural Networks* , IEEE, New York, Vol **1** (1991), 671–677.

[Mour84] O.G. Mouritsen, *Computer Studies of Phase Transitions and Critical Phenomena*, Springer Series in Computational Physics, Berlin, 1984.

[MePV87] M. Mezard, G. Parisi and M.A. Virasoro, *Spin Glass Theory and beyond*, Lecture Notes in Physics Vol.9, World Scientific, Singapore, 1987.

[Myll93] P. Myllymäki, *Bayesian Reasoning by Stochastic Neural Networks*, Ph. Lic. Thesis, Report C-1993-67, University of Helsinki, Dept. of Computer Science, 1993.

[Myll94] P. Myllymäki, "Using Bayesian networks for incorporating probabilistic a priori knowledge into Boltzmann machines," To appear in *Proc. of SOUTHCON'94*, Orlando, March 1994.

[OFMT90] P. Orponen, P. Floréen, P. Myllymäki and H. Tirri, "A Neural Implementation of Conceptual Hierarchies with Bayesian Reasoning," Proc. International Conference on Neural Networks , IEEE, New York, Vol **1** (1990), 297–303.

[OFMT92] P. Orponen, P. Floréen, P. Myllymäki and H. Tirri, "NEULA: A hybrid neural-symbolic expert system shell," *Journal of the Finnish Computer Science Society*, 1992.

[Pear88] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, California, 1988.

[Pere92] P. Peretto, *An Introduction to the Modeling of Neural Networks*, Cambridge University Press, Cambridge, 1992.

[RuMc86] D. Rumelhart, J. McClelland et al. , *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, Massachusetts, Vol. **1**: foundations, 282–317, 1986.

[Reic91] L.E. Reichl, *A Modern Course in Statistical Physics*, Arnold, London, 1991.

[Toul77] G. Toulouse, "Theory of the frustration effect in spin glasses: I," *Comm. Phys.* **2** (1977), 115–119.

[ViBr85] L. Viana and A.J. Bray, "Phase diagrams for dilute spin glasses," *J. Phys. C.* **18** (1985), 3037–3051.