Université de Montréal

Statistical physics of constraint satisfaction problems

par

Elyes Lamouchi

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)

en Informatique

Orientation intelligence artificielle

Octobre 2020

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Statistical physics of constraint satisfaction problems

présenté par

Elyes Lamouchi

A été évalué par un jury composé des personnes suivantes:

Pierre L'Écuyer
(président-rapporteur)

Michel Gendreau
(directeur de recherche)

Alexander Fribergh
(codirecteur de recherche)

Margarida Carvalho
(membre du jury)

Sommaire

La technique des répliques est une technique formidable prenant ses origines de la physique statistique, comme un moyen de calculer des quantités du type:

$$\mathbf{E}\Big[\log\big(\sum_{i\in\{1\ldots q\}^N} e^{-\beta E_i}\big)\Big], \quad E_i \overset{iid}{\sim} \mathcal{N}(0,\Delta).$$

Dans le jargon de physique, cette quantité est connue sous le nom de *l'énergie libre*, et toutes sortes de quantités utiles, telle que l'entropie, peuvent être obtenue de là par des dérivées. Plus généralement, n'importe quel système qui peut être décrit par la distribution de probabilité : $\mu_\beta(i) \propto e^{-\beta E_i}$, avec $[q]^N \equiv \{1,\ldots,q\}^N$ comme support, requiert la computation d'une constante de normalisation en sommant un nombre exponentiel de termes : $\mathcal{Z} \equiv \sum_{i\in\Sigma} e^{-\beta E_i}$. Cependant, ceci est un problème NP difficile, qu'une bonne partie de statistique computationelle essaye de résoudre, et qui apparaît partout; de la théorie des codes, à la statistique en hautes dimensions, en passant par les problèmes de satisfaction de contraintes. Dans chaque cas, la méthode des répliques, et son extension par (Parisi et al., 1987), se sont prouvées fortes utiles pour illuminer quelques aspects concernant la corrélation des variables dans μ_β et la nature fortement nonconvexe de $-\log(\mu_\beta())$. Algorithmiquement, il existe deux principales méthodologies adressant la difficulté de calcul que pose $\mathcal{Z}$:

a). *Le point de vue statique*: dans cette approche, on reformule le problème en tant que graphe dont les nœuds correspondent aux variables individuelles de μ_β, et dont les arêtes reflètent les dépendances entre celles-ci. Quand le graphe en question est localement un arbre, les procédures de message-passing sont garanties d'approximer arbitrairement bien les probabilités marginales de μ_β et de manière équivalente $\mathcal{Z}$. Les prédictions de la physique concernant la disparition des corrélations à longues portées se traduise donc, par le fait que le graphe soit localement un arbre, ainsi permettant l'utilisation des algorithmes locaux de passage de messages. Ceci va être le sujet du chapitre 4.

b). *Le point de vue dynamique*: dans une direction orthogonale, on peut contourner le problème que pose le calcul de $\mathcal{Z}$, en définissant une chaîne de Markov le long de laquelle, l'échantillonnage converge à celui selon μ_β, tel qu'après un certain nombre d'itérations (sous le nom de temps de relaxation), les échantillons sont garanties d'être approximativement générés selon μ_β.

Afin de discuter des conditions dans lesquelles chacune de ces approches échoue, il est très utile d'être familier avec la méthode de *replica symmetry breaking* de Parisi. Cependant, les calculs nécessaires sont assez compliqués, et requièrent des notions qui sont typiquemment étrangères à ceux sans un entrainement en physique statistique. Ce

mémoire a principalement deux objectifs : i) de fournir une introduction a la théorie des répliques, ses prédictions, et ses conséquences algorithmiques pour les problèmes de satisfaction de constraintes, et ii) de donner un survol des méthodes les plus récentes adressant la transition de phase, prédite par la méthode des répliques, dans le cas du problème $k-$SAT, à partir du point de vu statique et dynamique, et finir en proposant un nouvel algorithme qui prend en considération la transition de phase en question.

Mots-clés: problémes de satisfaction de contraintes, k-SAT, transition de phase, méthode des replicas, replica-symmetry-breaking, chaînes de Markov Monte Carlo, marche aléatoire

Summary

The replica trick is a powerful analytic technique originating from statistical physics as an attempt to compute otherwise intractable quantities of the type:

$$\mathbf{E}\Big[\log\big(\sum_{i\in\{1\dots q\}^N} e^{-\beta E_i}\big)\Big], \quad E_i \overset{iid}{\sim} \mathcal{N}(0,\Delta).$$

In physics jargon this quantity is known as the *free energy*, and all kinds of useful quantities, such as the entropy, can be obtained from it using simple derivatives. More generally, any system whose state can be described by a probability distribution of the form: $\mu_\beta(i) \propto e^{-\beta E_i}$, with $[q]^N \equiv \{1,\dots,q\}^N$ as a support, requires computing a normalizing constant by summing over an exponentially large state space: $\mathcal{Z} \equiv \sum_{i\in[q^N]} e^{-\beta E_i}$. This is however an NP-hard problem that a large part of computational statistics attempts to deal with, and which shows up everywhere from coding theory, to high dimensional statistics, compressed sensing, protein folding analysis and constraint satisfaction problems. In each of these cases, the replica trick, and its extension by (Parisi et al., 1987), have proven incredibly successful at shedding light on keys aspects relating to the correlation structure of μ_β and the highly non-convex nature of $-\log(\mu_\beta())$. Algorithmic speaking, there exists two main methodologies addressing the intractability of $\mathcal{Z}$:

a) *Statics*: in this approach, one casts the system as a graphical model whose vertices represent individual variables, and whose edges reflect the dependencies between them. When the underlying graph is locally tree-like, local message-passing procedures are guaranteed to yield near-exact marginal probabilities or equivalently compute $\mathcal{Z}$. The physics predictions of vanishing long range correlation in μ_β, then translate into the associated graph being locally tree-like, hence permitting the use message passing procedures. This will be the focus of chapter 4.

b) *Dynamics*: in an orthogonal direction, we can altogether bypass the issue of computing $\mathcal{Z}$, by defining a Markov chain along which sampling converges to μ_β, such that after a number of iterations known as the relaxation-time, samples are guaranteed to be approximately sampled according to μ_β.

To get into the conditions in which each of the two approaches is likely to fail (strong long range correlation, high energy barriers, etc..), it is very helpful to be familiar with the so-called *replica symmetry breaking* picture of Parisi. The computations involved are however quite involved, and come with a number of prescriptions and prerequisite notions (s.a. large deviation principles, saddle-point approximations) that are typically foreign to those without a statistical physics background. The purpose of this thesis is then twofold: *i*) to provide a self-contained introduction to replica theory, its predictions, and its algorithmic implications for constraint satisfaction problems, and *ii*) to give an account of state of the art methods in addressing the predicted phase transitions in the case of $k-$SAT, from both the statics and dynamics points of view, and propose a new algorithm takes takes these into consideration.

Keywords: constraint satisfaction problems, k-SAT, phase transitions, combinatorial optimization, replica trick, replica-symmetry-breaking, Markov chain Monte Carlo, self-avoiding-walk

Contents

List of Figures

List of Tables

Remerciements

Tout d'abord, j'aimerai remercier mes deux supervieurs: Michel Gendreau et Alexander Fribergh, pour leur encouragement, leurs commentaires judicieux, et le grand degré de liberté qu'ils m'ont accordé dans la direction de ce travail.

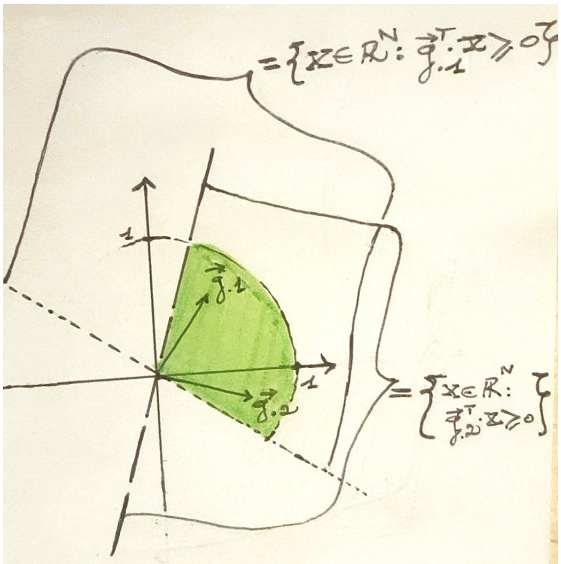
Ensuite, j'aimerai remercier mes parents Mohamed Hedi et Halima, pour le support moral et fiancier qu'il m'ont accordé au cours de mes études. Finalement, j'aimerai remercier mon professeur au études secondaires, Chokri Saida qui a alimenté mon intérêt pour toute chose mathématique, et sans lequel je n'aurais probablement pas entrepris ce chemin.

Introduction

0.1 A toy problem

We start with a geometric problem that is easy to visualize. Consider the N-dimensional ball of radius $\sqrt{N}$, centered at the origin: $\mathcal{S}_N \equiv \{x \in \mathbf{R}^N : ||x||_2 \leq \sqrt{N}\}$, and suppose we would like to determine its intersection with some random half-space of $\mathbf{R}^N$, whose supporting hyper-plane also passes though the origin. A straightforward way to do this, is by generating a random vector $g_1 \in \mathbf{R}^N$, and defining the corresponding half space as the set of points whose coordinate vector's angle with g_1 has a positive cosine: $U_1 \equiv \{x \in \mathbf{R}^N : \; g_1^T.x \geq 0\}$.

Now, suppose we reiterate this procedure by generating another vector g_2, and looking at the intersection of its corresponding half-space with the previously generated half-space inside the unit ball $\mathcal{S}_N \bigcap_{k\in\{1,2\}} U_k$, (see the figure below). If we keep on reiterating this procedure, we will ultimately come up a vector g_m satisfying $g_m^T.x < 0, \forall x \in \mathcal{S}_N \bigcap_{k\leq m-1} U_k$, such that the intersection of its corresponding half-space with all of the previous ones is the empty set.



$=\{x \in \mathbb{R}^N : \vec{g}_{.1}^T . x \geqslant 0\}$
1
1
$\vec{g}_{.1}$
$\vec{g}_{.2}$
$=\{x \in \mathbb{R}^N : \vec{g}_{.2}^T . x \geqslant 0\}$

Figure 1: The intersection of two random half spaces inside $\mathcal{S}_N$

The problem of characterizing $\mathcal{S}_N \bigcap_{1\leq k\leq M} U_k$, is known as *the perceptron problem*, and is one of many constraint satisfaction problems that have been fruitfully studied by statistical physicists. To make matters more precise:

Definition 1. *Consider a random vector $g_k \in \mathbf{R}^N$, where $(g_k)_i \overset{iid}{\sim} \Psi$ for all $1 \leq i \leq N$, for some distribution Ψ. We define its corresponding half-space U_k as the set of points whose angles with g_k have a positive cosine (or dot product):*

$$U_k \equiv \{x \in \mathbf{R}^N : \; g_k^T.x \geq 0\}.$$

Given a realization of the random variables $\{(g_k)_i : 1 \leq i \leq N\}_{1 \leq k \leq M}$, we are interested in the intersection of the M corresponding half spaces $\{U_k\}_{1 \leq k \leq N}$ inside the N dimensional ball of radius $\sqrt{N}$ centered at the origin:

$$\mathcal{S}_N \bigcap_{1 \leq k \leq M} U_k,$$

where

$$\mathcal{S}_N \equiv \{x \in \mathbf{R}^N : ||x||_2 \leq \sqrt{N}\}, \quad U_k \equiv \{x \in \mathbf{R}^N : \; g_k^T.x \geq 0\} \quad \text{for all } 1 \leq k \leq M.$$

A fascinating result of (Derrida and Gardner, 1988) states that:

- For $\alpha \equiv M/N > 2$, $\mathcal{S}_N \bigcap_{1 \leq k \leq M} U_k = \emptyset$ with high probability (w.h.p.).
- For $\alpha \leq 2$, $\log\left(\mu_N(\mathcal{S}_N \bigcap_{1 \leq k \leq M} U_k)\right)/N \approx \zeta_{rs}(\alpha)$ w.h.p. for some explicitly known function ζ_{rs}, where μ_N is the uniform measure on $\mathcal{S}_N$.

In other words, for N sufficiently large, a *phase transition* occurs at a threshold around $\alpha \equiv M/N = 2$.

0.2 A classical constraint satisfaction problem

A more elementary constraint satisfaction problem is the *random assignment problem.* Suppose you are throwing a dinner party on short notice, and there are only two tables at your disposition. You are asked to assign a seat to each guest such that no pair of guests that dislike each other are seated at the same table.

Let $\sigma_i \in \{-1, 1\}$ denote the position of the i^{th} guest, with minus one signifying a seat at the first table, and plus one a seat at the second one. Moreover, let $\{J_{ij}\}_{1 \leq i < j \leq N}$ be a set of independent zero-mean Gaussians with unit variance, representing the compatibility between guests, with $J_{ij} \geq 0$ signifying that the guests i and j get along well, and $J_{ij} < 0$ meaning that they do not. It is easy then to see that the assignment problem is equivalent to determining:

$$\text{argmax}_{(\sigma_1, \ldots \sigma_N) \in \{-1,1\}^N} \sum_{1 \leq i < j \leq N} J_{ij} \sigma_i \sigma_j, \tag{1}$$

where N is the number of guests.

Since we are interested in the least conflicting assignment, it can be useful to introduce a random probability measure over $\Sigma \equiv \{-1, 1\}^N$ assigning more weight to those assignments with the least conflicting pairs:

$$\mu_\beta(\sigma) \equiv \frac{e^{-\beta\mathcal{H}(\sigma)}}{\sum_{\sigma\in\Sigma} e^{-\beta\mathcal{H}(\sigma)}}, \tag{2}$$

where $\mathcal{H}(\sigma) \equiv -(1/\sqrt{N})\sum_{1\le i<j\le N} J_{ij}\sigma_i\sigma_j$.

In (Talagrand, 2011), the author proposes the following high level argument. Since $\mathcal{H}(\sigma)$ is about $\sqrt{N}$, if the random variables $\{\mathcal{H}(\sigma)\}_{\sigma\in\Sigma}$ are "not too correlated" (in a sense that will be thoroughly explained in chapter 5), we get $\max_{\sigma\in\Sigma}\mathcal{H}(\sigma) = \mathcal{O}(\sqrt{N}\sqrt{\log(2^N)}) = \mathcal{O}(N)$, such that the normalizing constant at the denominator of the random measure: $\mathcal{Z} \equiv \sum_{\sigma\in\Sigma} e^{-\beta\mathcal{H}(\sigma)}$ is dominated by a few summands that are of the same order (in logarithmic scale) to the entire sum. This wide disparity between the contributions of the summands in $\mathcal{Z}$ makes it considerably harder to approximate.

A central theme in this thesis, and in the study of constraint satisfaction problems or spin systems in general, is the relationship between the correlation structure of μ_β and the geometry of the minima of $\mathcal{H} : \Sigma \mapsto \mathbf{R}$ in the metric space (Σ, d_h) where $d_h(\sigma,\tau) \equiv |\{k : \sigma_k = \tau_k\}|$ is the *Hamming distance* between the assignments $\sigma, \tau \in \Sigma$.

To bear this fact to light, consider the correlation of the random function $\mathcal{H}(\sigma)$ whose randomness comes from the distribution of $\{J_{ij}\}$.

Proposition 1 (Talagrand, 2011). *Consider the real valued function $\mathcal{H} : \Sigma \mapsto \mathbf{R}$ given by: $\mathcal{H}(\sigma) = -(1/\sqrt{N})\sum_{i<j} J_{ij}\sigma_i\sigma_j$, where $J_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ for all $1 \le i < j \le N$. We define the overlap between two assignments as $q(\sigma,\tau) \equiv \big(\sum_{i\le N}\sigma_i\tau_i\big)/N$. The overlap is then related to the correlation of $\mathcal{H}(\sigma)$ through the following identity:*

$$\mathbf{E}\Big(\mathcal{H}(\sigma)\mathcal{H}(\tau)\Big) = \frac{Nq^2(\sigma,\tau)}{2} - \frac{1}{2}. \tag{3}$$

Proof. Since the compatibility variables are independent standard Gaussian, if $(ij) \neq (k,l)$, we have $\mathbf{E}[J_{ij}J_{kl}] = \mathbf{E}[J_{ij}]\mathbf{E}[J_{kl}] = 0$, while in the other case we have $\mathbf{E}[J_{ij}J_{ij}] = \mathbf{E}[J_{ij}^2] = Var(J_{ij}) = 1$, such that:

$$\mathbf{E}\Big(\mathcal{H}(\sigma)\mathcal{H}(\tau)\Big) = \frac{1}{N}\mathbf{E}\Big(\big(\sum_{i<j} J_{ij}\sigma_i\sigma_j\big).\big(\sum_{k<l} J_{kl}\tau_k\tau_l\big)\Big) \tag{4}$$

$$= \frac{1}{N}\left(\sum_{i<j}\sum_{\substack{k<l \\ (k,l)\neq(i,j)}} \underbrace{\mathbf{E}(J_{ij})}_{=0}\underbrace{\mathbf{E}(J_{kl})}_{=0}\sigma_i\sigma_j\tau_k\tau_l + \sum_{i<j}\underbrace{\mathbf{E}(J_{ij}^2)}_{=1}\sigma_i\sigma_j\tau_i\tau_j\right) \tag{5}$$

$$= \frac{\sum_{i<j}\sigma_i\sigma_j\tau_i\tau_j}{N} = \frac{N}{2}\left(\frac{\sum_{i\le N}\sigma_i\tau_i}{N}\right)^2 - \frac{1}{2}. \quad \blacksquare \tag{6}$$

Moreover, it is easy to verify that the overlap satisfies:

$$q(\sigma,\tau) = 1 - \frac{2d_h(\sigma,\tau)}{N}. \tag{7}$$

Hence, studying the distribution of the overlap can tell us a great deal about the typical distance between least conflicting pairs of assignments, and is intimately related to the study of the correlation structure of $\mathcal{H}(\sigma)$ and by extension of the correlation between individual variables $\{\sigma_i\}_{i\leq N}$ under μ_β.

More importantly, another non-obvious way in which the overlap comes up, is when trying to approximate the normalization constant $\mathcal{Z} \equiv \sum_{\sigma\in\{-1,1\}^N} e^{-\beta\mathcal{H}(\sigma)}$. As we discussed, this task is very hard even for very simple spin systems such as the assignment problem above, and no known rigorous technique is able to approximate $\mathcal{Z}$ in most cases when $\beta \gg 1$. However, in the late eighties, a number of statistical physicists have succeeded in doing so though a very powerful, yet non-rigorous, analytic technique known as the *replica trick*.

In this approach, one considers the harder problem of approximating $\mathbf{E}[\log \mathcal{Z}]$ by reducing it, through a Taylor expansion, to the problem of computing $\lim_{n\to 0} \mathbf{E}[\mathcal{Z}^n]$. When doing so, it is very useful to make a change of variables inside the sum in order to express $\mathbf{E}[\mathcal{Z}^n]$ as a function of the overlap and approximate it by a few dominant terms that are called *saddle-points*. Moreover, the choice of the correct saddle-point depends crucially on a few assumption about the distribution of the overlap.

0.3 Thesis organization

In the first part of this thesis, we attempt to present a self-contained introduction to *replica theory* and its predictions for a large class of *constraint satisfaction problems*, mainly following (Mezard and Montanari, 2009) and (Mezard et al., 1987), while filling in the missing proofs to some of the presented results. Then, in the second part, we focus on a classical NP-hard constraint satisfaction problem known as the k-SAT problem, which consists of determining the satisfiability of a given boolean formula and generating the set of solutions in the case in which the formula is satisfiable.

After having presented the algorithmic consequences of the physics predictions in the case k-SAT, we survey the state of the art methods in generating satisfying assignments, and the pros and cons of each approach. Finally, we present a novel algorithm that we name *SAW-SAT*, that builds upon previous work, and addresses some of the problems of alternative approaches. The organization of the thesis is then as follows:

- **Chapter 1 & 2.** We start by introducing some basic notions from spin glass theory and the necessary physical jargon, building up to establishing the overlap parameter as *the* central parameter to study a large class of disorder systems, which encompasses $q-$coloring, $k-$SAT problem, etc.

- **Chapter 3.** We give a self contained account of the replica trick, where it fails, and its extension by Parisi, on a toy model called the *REM*, then on a more general model called the $p-$spin model, of which a large number of constraint satisfaction problems are a special case.

- **Chapter 4.** We start by introducing the $k-$SAT problem and casting it in the language of probabilistic graphical models, then go on to discuss the algorithmic implications of the replica predictions regarding the uniform measure on satisfying assignments, building up to a message passing algorithm under the name of *Survey propagation*, that takes these into account.

- **Chapter 5.** After presenting the statics point of view in chapter 4, we start by introducing Markov chain Monte Carlo methods (MCMC), their provable guarantees, and the physics prediction that result in exponential relaxation times for single-flip dynamics. Finally, we survey some state of the art methods, their uniformity-efficiency trade-off, and propose an algorithm under the name of *SAW-SAT* which takes inspiration from some interesting ideas from computational physics, and overcomes some of the issue present in previous algorithms.

Part I

Spin glass theory and the 1RSB universality class

Chapter 1

An informal introduction to spin glasses

Chapter organization

We start by an introductory discussion of spin glasses as a physical object and their main properties (**1.1**). Then, after introducing a physical model under the name of the *Gibbs measure* that is used to describe it, we move on to describe its dependence on a temperature parameter (**1.1.1**). Afterwards, we begin the discussion of phase transitions and introduce the natural objects used to characterize it: *thermodynamic quantities* (**1.2.1**), *order parameters* (**1.2.2**), and *correlation length* (**1.2.3**). Subsequently, we describe a phenomenon that is widely known in statistical physics as *universality*, that relates very different models by a set of constants called *critical exponents* (**1.2.4**), and move on to a brief historical note on experimental studies on spin glasses that motivates the previous sections (**1.3**). Finally, we add an appendix for some first-principles (non-rigorous) physical derivation of a central object of the thesis called *the free energy*, in order to provide some intuition for the interested reader. However, since there is no mention of this appendix later on in the thesis it can be skipped without loss of continuity.

Referencing theorems

Since one of the main goals of this thesis is to provide a self-contained introduction to the subject of replica theory and its connection with constraint satisfaction problems, we have filled in some gaps in the literature by formalizing a number of definitions and proving a number of theorems/lemmas/proposition left as exercises to the reader or simply mentioned in passing. To distinguish the results whose proof can be found in the mentioned reference from those we completed: If a theorem's proof can be found in the literature we will reference it *inside* the statement of the theorem, e.g. (**Theorem 1** (Mezard and Montanari, 2009) *Consider an arbitrary ...*) , and will only prove it if it provides further intuition. On the other hand, if a result was left as an exercise to the reader or whose proof is missing from the standard references, we will state the theorem without the reference, e.g. (**Theorem 1** *Consider an arbitrary...*), but we will mention the reference where we found the statement of the result in the preceding paragraph (or just after it). The same goes for lemmas/propositions and definitions.

1.1 Disorder and frustration

Consider a metal (such as Cu) with a small amount of a magnetic species (e.g. Mn) diluted in it. Such a mixture is called a magnetic alloy. An important feature of magnetic alloys is that, while they keep the properties of the original metal, the magnetic impurities are strong enough to create a peculiar magnetic behaviour.

Spin glasses (such as CuMn or AuFe) are a famous example of such alloys belonging to the class of *disordered systems with quenched disorder*. A system is said to be *disordered* if some parameters defining its behaviour are random.

In the case of a spin glass, the magnetic species form local moments or *spins* whose location in the metal is random and the spatial distribution of spins drives the system into a *frustrated state* which determines its magnetic behaviour.

As an example, consider a dilute solution of Mn in Cu. This can be modelled as a grid of Cu with Mn spins located randomly on this grid, each pointing in a certain direction. Now given the locations of the magnetic impurities in the metal, the conduction electrons at each spin scatter onto the neighboring sites, inducing a strongly oscillating interaction potential between spins called *the RKKY interaction* (Ruderman and Kittel, 1954; Kasuya, 1956; Yosida, 1957):

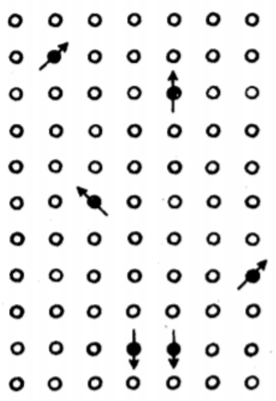


Figure 1.1: Magnetic moments in a metallic matrix (K.Binder, 1977a)

$$J_{ij}(r_{ij}) = J_0 \frac{cos(2k_F r_{ij} + \phi_0)}{(k_F r_{ij})^3}, \tag{1.1}$$

where r_{ij} is the the distance between the pair of spins (i, j) while J_0 and ϕ_0 are constant terms and k_F is the Fermi wave number of the host metal.

Thus, if the spins are not too far apart, we can expect an effective interaction or *coupling* between them. Moreover this interaction should oscillate between positive and negative values depending on how far apart they are, as evident in the figure above.

A positive coupling J_{ij} indicates that the system is favourable to the alignment of the pair (i, j) while a negative coupling indicates that they should be in opposite di-

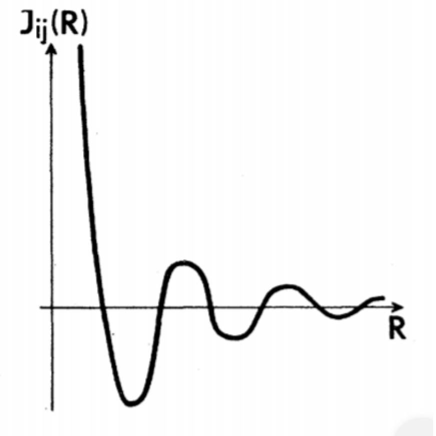


Figure 1.2: Oscillating potential with $R \equiv r_{ij}$ (K.Binder, 1977a)

rections.

The resulting optimization problem is then to find the most favorable spin directions $\{\sigma_i\}_{1\leq i\leq N}$ subject to the sign of the couplings:

$$\max_{(\sigma_1,\ldots,\sigma_N)\in\{-1,1\}^N} \sum_{1\leq i<j\leq N} sgn(J_{ij})\sigma_i\sigma_j, \quad \text{subject to} \quad \{J_{ij}\} = \mathcal{J}, \tag{1.2}$$

where $\mathcal{J}$ is a symmetric $\mathbf{R}^{N\times N}$ matrix with zero diagonal entries representing the set of couplings associated with the given spin system.

In a *ferromagnetic* system all coupling constants are positive such that the most favourable state has all spins aligned, whereas in an *antiferromagnetic* one, the couplings are all negative and so each pair of spins should point in opposite directions. A spin glass is essentially a mixture of both situations and the basic physics arises as a competition between ferro- and antiferromagnetic interactions.

Through the interaction potentials, the spins are encouraged by their neighboring local moments to point in contradictory directions simultaneously such that the interactions pertaining to a given spin cannot all be simultaneously satisfied. The resulting failure to satisfy all couplings drives the system into a strange kind of quasi-equilibrium referred to as *frustration.*

In a pure ferromagnetic system, all couplings can be satisfied by having all spins point in the same direction and hence there can be no frustration, which is not the case for an antiferromagnet. In particular, if we suppose that all spins interact with each other, then any antiferromagnetic system of size $N \geq 3$ spins cannot possibly satisfy all negative couplings and hence it will invariably be in a frustrated state.

If a physical system depends on a large number of variables with many degrees of freedom, an exact solution is often not possible and even if possible not very realistic. Thus, it can be very useful to add stochasticity into the description of the system at the miscroscopic level by specifying the physical picture for a miniature $N-$particle system and then take the scaling limit $N \to \infty$ to deduce the macroscopic quantities of interest.

Historically, this statistical point of view started with Boltzmann who managed to deduce the second law of thermodynamics by starting with a probabilistic description of the individual particles and their interactions.

Here and throughout the thesis, we use the following notation: $[N] \equiv \{1, 2, \ldots N\}$ and we write $i, j \in [N]^2$ to mean $\{(i, j) : i \in [N], j \in [N]\}$.

Consider a system of N random variables or *particles*: $\sigma = (\sigma_1, \ldots \sigma_N)$, each individual particle σ_k takes value in $\mathcal{X}$. The support of the system or the *state space* (as in state of all possible states) is then given by $\Sigma \equiv \mathcal{X}^N$.

Equilibrium statistical mechanics (Fischer and Hertz, 1991) postulates that the probability that a system with a fixed number of particles, at a fixed temperature T, is in a given state $\sigma \in \Sigma$, is given by the *Gibbs-Boltzmann* distribution:

$$\mu_\beta(\sigma) = \frac{e^{-\beta\mathcal{H}(\sigma)}}{\sum_{\sigma\in\Sigma} e^{-\beta\mathcal{H}(\sigma)}} \quad \text{in the discrete case or}$$

$$\mu_\beta(\sigma) = \frac{e^{-\beta\mathcal{H}(\sigma)}}{\int_\Sigma e^{-\beta\mathcal{H}(\sigma)} d\sigma} \quad \text{in the continuous case,}$$

where $\beta \equiv 1/T$ is the inverse temperature and $\mathcal{H} : \Sigma \mapsto \mathbf{R}$ is *the energy function* or *Hamiltonian.* The Gibbs-Boltzmann distribution is used to model a wide array of $N-$particle systems throughout statistical physics such as ideal gases, crystals, etc. The choice of the Hamiltonian then differs depending on the system we are trying to model.

For magnetic alloys, each system is identified with a single symmetric matrix with zero diagonal entries representing the coupling: $\mathcal{J} \in \mathcal{R}^{N\times N}$. A rather general form for the Hamiltonian of spin systems can be given by:

$$\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j - B\sum_{i\in[N]} \sigma_i, \tag{1.3}$$

where $\mathcal{J}$ is the coupling matrix whose entries $\{J_{ij}\}$ are supposed to model the RKKY interactions noted above, and B is a global magnetic field acting uniformly on all spins $\{\sigma_k\}_{k\in[N]}$.

Note that in a given $N-$particle system, the values of the couplings $\{J_{ij}\}$ are fixed (and assumed to be known) such that the system is associated with a unique Gibbs-Boltzmann distribution that is a function of just $\sigma = (\sigma_1, \ldots \sigma_N)$:

$$\mu_\beta(\sigma) = \frac{\exp\left(\beta\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j + \beta B\sum_{i\in[N]}\sigma_i\right)}{\sum_{\sigma\in\Sigma}\exp\left(\beta\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j + \beta B\sum_{i\in[N]}\sigma_i\right)}. \tag{1.4}$$

At $B = 0$, the proposed Hamiltonian closely resembles the optimization problem described above, and finding the most likely configuration: $\max_{\sigma\in\Sigma}\mu_\beta(\sigma)$, in the low temperature limit $T \to 0$ ($\beta \to \infty$), is in fact equivalent to the original hard constrained problem.

Oftentimes, μ_β is simply referred to as a Gibbs measure, and denoted $\mu_\beta(\sigma) \propto e^{-\beta\mathcal{H}(\sigma)}$ up to a normalization constant called *the partition function* $\mathcal{Z} \equiv \sum_{\sigma\in\Sigma} e^{-\beta\mathcal{H}(\sigma)}$.

Definition 2. *An $N-$particle spin system is a collection of dependent random variables: $\sigma = (\sigma_1, \dots \sigma_N)$, whose support is Σ and whose distribution is the Gibbs-Boltzmann distribution, that is parameterized by a symmetric zero-diagonal matrix of couplings $\mathcal{J}$ an inverse temperature parameter β, and a global magnetic field B:*

$$\sigma \sim \mu_\beta(\sigma) \quad \textit{where} \quad \mu_\beta(\sigma) \propto \exp\big(\beta \sum_{1\le i<j\le N} J_{ij}\sigma_i\sigma_j + \beta B \sum_{i\in[N]} \sigma_i\big). \tag{1.5}$$

Following statistical physics convention, we will write the Gibbs measure with the inverse temperature β subscript but we will keep the other two parameters B and $\mathcal{J}$ implicit, even though we do assume they are fixed (and known).

For a spin glass model with binary or *Ising* spins: $\mathcal{X} \equiv \{-1, 1\}$, the state space is given by $\Sigma = \{-1, 1\}^N$ where the value of an individual spin $\sigma_k = +1$ or -1 is interpreted as the k^{th} spin pointing up or down respectively.

Definition 3. *Given an $N-$particle spin system $\sigma \sim \mu_\beta$, the set of minimum energy configurations are called ground states, and correspond to those with the highest probability:*

$$\sigma_{gs} = \underset{\sigma\in\Sigma}{argmax}\{\mu_\beta(\sigma) \propto \exp(-\beta\mathcal{H}(\sigma))\} = \underset{\sigma\in\Sigma}{argmin}\{\mathcal{H}(\sigma)\}. \tag{1.6}$$

Note that, computing the partition function $\mathcal{Z}$ cannot be done naively as it requires summing over an exponential ($|\mathcal{X}|^N$) number of configurations. In fact, normalizing discrete probability distributions over exponentially growing supports is a central problem in statistical inference, information theory and computational complexity theory.

The second part of this thesis (*Part II*) deals with analytic and algorithmic ideas originating from the physics literature which have proven very fruitful in approximating $\mathcal{Z}$.

1.1.1 High vs low temperature regimes

An important observation regarding the temperature dependence of the Gibbs measure and the correlation between individual spins can already be made:

- At high temperature, $\beta \approx 0$ and therefore, $\mu_0(\sigma) = 1/|\Sigma|, \forall\sigma \in \Sigma$, independently of the energy function and hence of the signs of the couplings $\{J_{ij}\}$. Thus, in the high temperature regime, all states have nearly equal probability such that the energetically favourable states are as probable as those that do not satisfy the couplings (i.e. those with relatively many individual spin pairs in $\{(i,j):\ J_{ij}\sigma_i\sigma_j < 0\}$).

- On the other hand, at low temperature, β is large and therefore, a small change in the energy corresponds to a large change in the probability of a given state, such that nearly all of the probability mass is concentrated on the minima of the energy function. Hence, the spins are locked together with high probability in a few configuration and are thus strongly correlated.

In fact, as the temperature is lowered, the energy landscape undergoes a series of structural changes corresponding to multiple phase transitions, where each phase is characterized by subtle changes in the correlation structure. The duality between the correlation structure and the geometry of the minima of the energy function in the sense of the Hamming distance between low energy configurations: $d(\sigma,\tau) \equiv |\{i \in [N] : \sigma_i \neq \tau_i\}|$, will be a major theme in later chapters where we will study the different spin glass phases at length.

Furthermore, since the inverse temperature parameter β controls the sensitivity of μ_β w.r.t. changes in the energy, it allows one to consider smoothed versions of distributions with zero probability events. As pointed out above, this is very useful when considering models with hard constraints that have zero probability in a significant portion of the state space, as it helps avoiding numerical issues in dynamical algorithms, as we shall see in the last chapter, when discussing the use of Markov chains for sampling $\overset{approx}{\sim} \mu_\infty(.)$.

1.2 Phase transitions

In the study of phase transitions, each phase exhibits a different kind of order, and *order parameters* serve to delineate the boundaries of each phase.

For example, consider the gas-to-liquid phase transition. As the temperature is lowered at constant pressure, we reach a critical temperature where the liquid component instantly becomes much denser than the gas (Simons, 1997), the difference between the two densities can therefore serve as an order parameter that signals the gas-to-liquid phase transition. As a rule, systems are ordered at low temperature and transition into less ordered states upon heating. Boiling water is a well known example for this phenomenon.

To shed some light on the low temperature ordered phase of a spin glass, i.e. *the spin glass phase*, we will contrast it to the simpler case of a ferromagnet. In general, the order parameter is defined such that it is zero below the transition temperature and becomes nonzero at a critical temperature signaling the onset of a phase transition (Castellani and Cavagna, 2005).

In spin systems, *the magnetization* is a central order parameter which describes the orientation of spins. Physically speaking, each spin feels the combined effect of the global magnetic field produced by its neighbors through the couplings, which pushes the spin to point in a specific direction completely determined by the Hamiltonian. The resulting orientation is called the *local magnetization* and is denoted by m_i. More formally:

Definition 4 (Mezard and Montanari, 2009)**.** *Given an $N-$particle spin system, the local magnetization at the i^{th} spin is the expectation of σ_i with regards to μ_β and is denoted by m_i:*

$$m_i \equiv \sum_{\{\sigma_k\}_{k\neq i}\in\mathcal{X}^{N-1}} \mu_\beta(\sigma_{[N]\backslash i}, \sigma_i = 1) - \mu_\beta(\sigma_{[N]\backslash i}, \sigma_i = -1). \tag{1.7}$$

Here and throughout the thesis, we will denote to the expectation with regards to μ_β by brackets such that: $m_i = \langle\sigma_i\rangle$. Since we want to characterize the general orientation of the entire system, we define a second order parameter: the total magnetization (or simply the magnetization) as the arithmetic mean of local magnetizations:

$$m \equiv \frac{1}{N}\sum_{i=1}^{N}\langle\sigma_i\rangle. \tag{1.8}$$

As pointed out above, a variation in the temperature redistributes the probability mass by either flattening it over all states (at high temperature) or putting all the mass on the the minima of the energy function.

In a ferromagnet, all spin couplings are positive and hence, at low temperature, the system is ordered in the sense that all spins are aligned. As the temperature increases, less favourable spin configurations gain probability mass. In the high temperature regime, the system is in a disordered state and has spins pointing in apparently random directions.

Consider an $N-$particle spin system with the above Hamiltonian and let $\mu_\beta^{(i)}(\sigma_i = .)$ denote the marginal probability of the i^{th} spin. Considering that, when $\beta \approx 0$, the probabilities of all states are nearly equal independently of the couplings, the marginal probabilities that an individual spin be ±1 are equal, such that the magnetization is zero:

$$m = \frac{1}{N}\sum_i \mu_\beta^{(i)}(1) - \mu_\beta^{(i)}(-1) \approx 0, \quad \text{when} \quad \beta \approx 0. \tag{1.9}$$

This results holds independently of the signs of the couplings. However, at low temperature, ferromagnetic and glassy systems behave much differently. Consider the Hamiltonian introduced above

$$\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j - B\sum_{i\in[N]}\sigma_i, \tag{1.10}$$

the lowest possible value this function can take is $-\sum_{1\leq i<j\leq N}|J_{ij}| - BN$. In a ferromagnet it is possible to reach the lowest energy since $J_{ij} \geq 0,\ \forall 1 \leq i < j \leq N,$ therefore, depending on the sign of B, the ground state configuration is either $\sigma_{gs} = (1, \ldots 1)$ if $B > 0$ or $(-1, \cdots - 1)$ if $B < 0$.

In the low temperature regime, β is large, and the Gibbs measure is sensitive to the slightest perturbation in energy, in which case, the $\pm$ symmetry is broken, and all local magnetizations take a nonzero value with the same sign, this phenomenon is called *spontaneous symmetry breaking* (Fischer and Hertz, 1991), and will be expanded upon in the next chapter.

For a spin glass, the high temperature results still hold, since no matter the signs of the couplings the probability of any state $\sigma \in \Sigma$ will be $\mu_{\beta\approx 0}(\sigma) = 1/\mathcal{Z}$, such that $\mu_\beta^{(i)}(1) - \mu_\beta^{(i)}(-1) = 0$, but the fact that the interaction is allowed to take negative values complicates the analysis for the low temperature regime. The next chapter will introduce the appropriate framework needed to discuss a more complex variant of symmetry breaking characteristic of the spin glass phase.

1.2.1 Thermodynamic quantities

Since we're ultimately interested in a macroscopic piece of spin glass material, after specifying the behaviour at the microscopic scale (of an $N-$particle system) with an appropriate Hamiltonian, we take the large N limit to derive quantities of interest. This limit is often referred to as *the thermodynamic limit*, in reference to it's original use in Boltzmann's work.

Definition 5 (Mezard and Montanari, 2009)**.** *Given an $N-$particle spin system, with the Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j - B\sum_{i\in[N]}\sigma_i$, its three thermodynamic quantities (or potentials) are as follows:*

- *The free energy: $F(\beta) = \log(\mathcal{Z})/\beta$.*
- *The internal energy: $U(\beta) \equiv \langle \mathcal{H} \rangle = \sum_{\sigma\in\Sigma} \mathcal{H}(\sigma)\, \mu_\beta(\sigma)$.*
- *The entropy: $\mathcal{S}(\beta) \equiv -\sum_{\sigma\in\Sigma} \mu_\beta(\sigma) \log\big(\mu_\beta(\sigma)\big)$.*

The single most important thermodynamic quantity in a disordered system is the free energy and all other quantities can be derived directly from it (as shown in the appendix). Physically speaking, given a system in thermodynamic equilibrium, the evolution of its state can be best described by the free energy. In particular, changes in the value of the free energy can be used to pinpoint spontaneous changes in the system and determine their direction (Simons, 1997).

Although, it is not directly related to the rest of the discussion, we have found it illuminating to delve deeper into the thermodynamics of the free energy, and relate it to the other two main thermodynamic potentials. We have therefore included an appendix detailing the derivation of $F(\beta)$ from first principles and it's physical significance, following (Claudius, 2017) and (Simons, 1997), as well as the simple proof of the proposition below. Note that the rest of the thesis makes no further mention of this appendix, and can therefore be skipped without loss of understanding, for those not interested.

Proposition 2 (Mezard and Montanari, 2009)**.** *Given an N-particle spin system $\sigma \sim \mu_\beta$, and $F(\beta), U(\beta), S(\beta)$ as defined above, we then have:*

$$F(\beta) = U(\beta) - \frac{S(\beta)}{\beta}. \tag{1.11}$$

1.2.2 Order parameters

As previously discussed, at low enough temperatures β is large and the system becomes very sensitive to small perturbations. In particular, the slightest change in energy in the vicinity of the critical temperature can cause dramatic changes in the system. Hence, we can probe the existence of a phase transition by perturbing the energy and monitoring its response (Simons, 1997).

These perturbations can be either local or global, depending on the information we are seeking. In magnetic systems, we can either perturb the system globally by varying a global parameter such as the temperature and look at some temperature dependent observable such as *the specific heat* $C(\beta) \equiv \partial \mathcal{H}/\partial \beta$, or we can add a local magnetic field acting on a given spin and look at the effect of this local perturbation on faraway spins. If the system exhibits long range correlation, we expect local perturbations to have an effect on far off spins (Fischer and Hertz, 1991).

Definition 6 (Mezard and Montanari, 2009)**.** *Consider an $N-$particle spin system with the Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j$, and suppose we add a local magnetic field B_i at the i^{th} spin, such that the new (perturbed) energy function of the system becomes $\underline{\mathcal{H}}_{B_i}(\sigma) = \mathcal{H}(\sigma) + B_i\sigma_i$. The spin glass susceptibility is then defined as*

$$\chi^{SG} \equiv \frac{\beta^2}{N}\sum_{i=1}^{N}\chi_{ji}^2 \quad where \quad \chi_{ji} \equiv \frac{dm_j}{dB_i}_{|B_i=0}. \tag{1.12}$$

Physically speaking, if $\chi_{ji} > 0$, the perturbation induces a positive response in the system such that the material is attracted by the local magnetic field B_i and the j^{th} spin's orientation, characterized by m_j, shifts towards $sgn(B_i)$. This follows by definition of the susceptibility, if $\chi_{ji} = dm_j/dB_i > 0$ then $sgn(dm_j) = sgn(B_i)$. The following theorem is a classical result in statistical physics, called *the fluctuation-dissipation relation* (Mezard and Montanari, 2009), as it relates the correlation between spins within the unperturbed system with its response to an infinitesimal perturbation.

Theorem 1 (Mezard and Montanari, 2009)**.** *Consider an $N-$particle spin system with the locally perturbed Hamiltonian $\underline{\mathcal{H}}_{B_i}(\sigma) = \mathcal{H}(\sigma) + B_i\sigma_i$, the spin glass susceptibility then satisfies the following "fluctuation-dissipation" relation:*

$$\chi^{SG} = \frac{\beta^2}{N}\sum_{ij}[\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle]^2.$$

Proof.

$$\begin{aligned}
\chi_{ji} &= \frac{d}{dB_i}\langle\sigma_j\rangle = \frac{d}{dB_i}\left[\frac{\sum_{\underline{\sigma}} \underline{\sigma}_j e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}}\right] \\
&= \frac{\sum_{\underline{\sigma}} \underline{\sigma}_j \frac{d}{dB_i} e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}} - \frac{\sum_{\underline{\sigma}} \frac{d}{dB_i} e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}^2} \sum_{\underline{\sigma}} \underline{\sigma}_j e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i} \\
&= \beta\left[\sum_{\underline{\sigma}} \underline{\sigma}_j\underline{\sigma}_i \frac{e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}} - \left(\frac{\sum_{\underline{\sigma}} \underline{\sigma}_i \frac{d}{dB_i} e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}}\right)\left(\frac{\sum_{\underline{\sigma}} \underline{\sigma}_j \frac{d}{dB_i} e^{-\beta\mathcal{H}(\sigma)+\beta B_i\underline{\sigma}_i}}{\mathcal{Z}}\right)\right] \\
&= \beta\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle.
\end{aligned}$$

Hence the spin glass susceptibility satisfies

$$\chi^{SG} = \frac{\beta^2}{N}\sum_{ij}[\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle]^2. \quad \blacksquare \tag{1.13}$$

In order to detect the transition to glassy phases, a necessary condition to be checked is that χ^{SG} should diverge as $N \longrightarrow \infty$ (Mezard and Montanari, 2009).

1.2.3 Correlation length

Suppose we take a sample of spin glass, cut it in half and measure some observables for each piece. Assuming the interactions taking place are identical in each piece and that the temperature and magnetic field B are kept fixed, the two pieces will have the same properties as the whole. However, if we keep repeating this process long enough, at some point, we will reach a length scale where the magnetization, susceptibility and other observables of the subsystems start to differ from the ones measured at the previous iteration. Since the interactions arise from the scattering of electrons which only happen within a certain reach, this length scale defines a correlation length below which the spins are highly correlated within each subsystem (Simons, 1997).

An intuitive description of this phenomenon is to consider a spin system where the spins are not independent under the Boltzmann measure but the correlation decays above a certain length scale ξ, the idea proposed in (Mezard and Montanari 2009) is to consider blocks of length $\xi \in \mathbf{N}$, taking value in $|\mathcal{X}|^{\xi}$, that are nearly independent under μ_β. Since the system becomes more correlated at low temperatures, the correlation length should be a function of β.

Theorem 2 (Mezard and Montanari, 2009)**.** *Consider the one-dimensional Ising model where the state space is given by* $\Sigma \equiv \{-1, 1^N\}$, *and with the following interactions:* $\mathcal{J} = \{J_{i,i+1} : i \in [N-1]\}$, *with the usual Hamiltonian at zero magnetic field* $B = 0$*:*

$$\mathcal{H}(\sigma) = -\sum_{i\in[N-1]} J_{ij}\sigma_i\sigma_{i+1} \quad \textit{such that} \quad \mu_\beta(\sigma) \propto \exp\Big(\beta\sum_{i\in[N-1]} J_{ij}\sigma_i\sigma_{i+1}\Big). \tag{1.14}$$

Then, for all pairs of spins within a distance: $\delta N < i < j < (1-\delta)N$, *for some*

$\delta > 0$, *we have:*

$$\lim_{N\to\infty} \langle \sigma_i \sigma_j \rangle = \exp\Big\{ -\frac{|i-j|}{\xi(\beta)} \Big\} + \Theta(e^{-\alpha N}), \quad \xi(\beta) \equiv -\frac{1}{\log(\tanh\beta)}. \tag{1.15}$$

In other words, for certain models, the correlation between spins decreases exponentially fast above a temperature-dependent critical distance $\xi(\beta)$. Moreover, since the susceptibility can be expressed in terms of correlations, this critical distance can also be understood in terms of a fluctuation dissipation relation. In particular, in the simpler case where $B = 0$ such that by symmetry we get $\langle \sigma_i \rangle = 0$, we have the following result (Mezard and Montanari, 2009):

$$\chi_M(\beta) = \beta \sum_{ij} \langle \sigma_i \sigma_j \rangle = \beta \sum_{i=-\infty}^{+\infty} \exp(-|i|/\xi(\beta)) + \Theta(e^{-\alpha N}) \quad \text{when} \quad N \gg 1. \tag{1.16}$$

1.2.4 Universality

When undergoing a phase transition, the system is subject to dramatic changes in its order parameters such that its response functions typically become singular (Simons, 1997):

- the compressibility in the liquid-to-gas phase transition: $\kappa_T = -\frac{1}{V}\frac{\partial V}{\partial P}_{|T=T_c} = \infty$
- and the magnetic susceptibility in the paramagnetic-to-ferromagnetic phase transition in the case of a spin system with all positive couplings: $\chi_M = \frac{\partial m}{\partial B}_{|T=T_c} = \infty$.

Since most of these observables can be written as derivatives of the free energy, phase transitions typically correspond to singularities in the free energy.

Note that, since the partition function of finitely many particles is a sum of exponential functions, it is always analytic in β, such that the free energy $F(\beta) = -\log \mathcal{Z}/\beta$ is analytic as well. Therefore, singularities corresponding to phase transitions can only occur in the thermodynamic limit as $N \to \infty$.

Thus, the study of phase transitions is in large parts reducible to finding the origin of singularities in the free energy and characterizing them by a set of *critical exponents* (s.a. α, γ below).

For instance, consider a ferromagnetic $N-$particle spin system, i.e. one where all coupling J_{ij} are positive, and with the Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1\le i<j\le N} J_{ij}\sigma_i\sigma_j - B\sum_{i\in[N]}\sigma_i$, and let $t \equiv \frac{T-T_c}{T_c}$ be *the reduced temperature.* The ferromagnetic phase transition can then be characterized by a couple of critical exponents (α, γ) (Simons, 1997) relating to the magnetic susceptibility χ_M and the *specific heat*: $C \equiv \partial\mathcal{H}/\partial\beta$:

$$\chi_M = \frac{\partial m}{\partial B}_{|B=0^+} \propto |t|^{-\gamma} \quad , \quad C = \frac{\partial \mathcal{H}}{\partial \beta} \propto |t|^{\alpha}. \tag{1.17}$$

Surprisingly, some singularities in the free energy of very different systems (i.e. with very different Hamiltonians) can be characterized by the same set of critical exponent. For example the liquid-to-gas and para-to-ferromagnetic transitions are described by the same set of critical exponents and are therefore said to belong to the same *Universality class* (Simons, 1997).

1.3 Peculiar magnetic behaviour

Experimental studies of spin glasses started in the 70s, the basic question was whether such systems display a phase transition at low enough temperatures, the high temperature regime being trivially paramagnetic for all magnetic systems. The magnetic susceptibility of CuMn was measured as a function of temperature, as shown in the figure below.

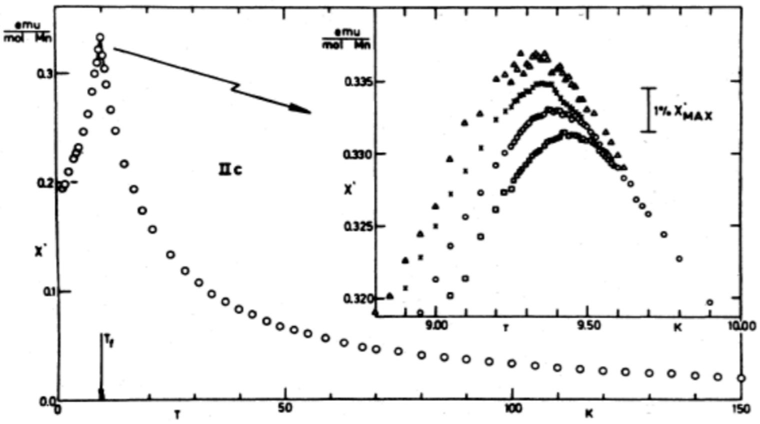


Figure 1.3: A cusp in the susceptibility (Binder and Young, 1986).

At high temperature, experiments suggested that the magnetic susceptibility decreases, as expected by the Curie law for paramagnetic materials $\chi_M \propto T^{-1}$. Furthermore, at some critical temperature we find a sharp peak in the susceptibility, which seem to indicate a phase transition (Binder and Young, 1986).

Since the critical system is highly susceptible to perturbations, the susceptibility was expected to diverge at some critical temperature, as in the ferromagnetic case. However, experiments demonstrated a cusp in the susceptibility at the critical temperature rather than a full blown singularity. Moreover, the specific heat: $C = \partial\mathcal{H}/\partial\beta$, which measures the change in energy induces by varying the temperature, was observed to be smooth around the critical temperature with a broad maximum only at higher temperatures, which is rarely the case in standard phase transitions.

Neutron-scattering experiments also revealed the absence of any kind of ferro- or antiferromagnetic (i.e. $J_{ij} < 0, \forall i, j$) spatial ordering below the critical temperature, but rather an irregular kind of equilibrium with strongly correlated local magnetic moments (i.e. spins) frozen in apparently random directions.

The main reason for this type of behaviour was conjectured to be the indirect nature of the RKKY interactions, whereby placing a magnetic impurity in a sea of conducting electrons has a damping effect on the susceptibility.

These observations, along with many other experimental studies uncovering strange glassy phenomena, motivated an extensive amount of theoretical work with the goal of understanding what would a phase transition look like in a solvable model of spin glass.

1.4 Appendix: Thermodynamics of the free energy

1.4.1 Potentials of an ideal gas

Consider a system of N particles moving randomly in a space of volume V with internal pressure P and colliding with each other in a perfectly elastic fashion, i.e. without losing any kinetic energy. Such a system is called an *ideal gas* and is governed by the *the ideal gas law*:

$$PV = nRT, \tag{1.18}$$

with n being the quantity of gas in moles, R is the gas constant and T the temperature.

The *internal energy* of a disordered system is the microscopic energy resulting from the random motion of molecules and their interactions, it is essentially the statistical mechanics counterpart of the energy of moving objects in classical mechanics.

The internal energy is defined as a differential quantity rather than an absolute one. More precisely, the first law of thermodynamics defines the change in internal energy of a system as:

$$dU = \underbrace{\delta Q}_{\text{energy flowing \textit{into} the system as heat}} + \underbrace{\delta W}_{\text{work done \textit{on} the system}}. \tag{1.19}$$

In the case of pressure-volume work or *PV-work*, $\delta W = -PdV$ and $\delta Q = TdS$, such that the internal energy U is related to the entropy S through $U = TdS - PdV$.

If the gas is kept at constant temperature and volume with fixed number of particles then its state can be described by the Boltzmann distribution, where the probability of an *eigenstate* of energy E_r is given by

$$\mu_\beta(r) = \frac{e^{-\beta E_r}}{\mathcal{Z}} \quad \text{with} \quad \beta \equiv \frac{1}{kT} \quad \text{for some constant } k. \tag{1.20}$$

Since the state of the system is random, its *internal energy* is the expected value of the energy with regards to the Boltzmann distribution:

$$U(\beta) \equiv \langle E_r \rangle = \sum_{r \in \Sigma} \mu_\beta(r) E_r. \tag{1.21}$$

The rationale for this is that given a suitable distribution for the energy function, the fluctuations of $U(\underline{\beta})$ at fixed temperature will vanish in the thermodynamic limit and the system will be associated with a unique internal energy potential with high probability.

1.4.2 The Legendre transform and the Helmholtz free energy

A neat way to shift variable dependencies in physics is through the *Legendre transform.* Formally:

Definition 7. *The Legendre transform of a function $f : \mathcal{A} \mapsto \mathbf{R}$ is given by*

$$f^*(t) = \underset{x \in \mathcal{A}}{sup}\{tx - f(x)\} \quad where \quad t \in \{s \in \mathbf{R} : f^*(s) < \infty\}. \tag{1.22}$$

However, in the physics literature (McKay et al., 2009), this technique is formulated slightly differently while preserving the same aim of shifting the dependency of a function from one variable to another. Consider a function of two variables $f(x, y)$, we define *the conjugate variables* u and v of x and y respectively, as follows:

$$df = \underbrace{\left(\frac{df}{dx}\right)_y}_{u} dx + \underbrace{\left(\frac{df}{dy}\right)_x}_{v} dy, \tag{1.23}$$

Definition 8 (McKay et al., 2009)**.** *The Legendre transform $g(.,.)$ of a function of two variables $f(x, y)$ on one of its variables, e.g. y, is defined by fixing said variable to some value $y = \underline{y}$, and defining the transform as:*

$$g(x, v) \equiv f(x, y = \underline{y}) - v\underline{y}. \tag{1.24}$$

The Legendre transform of $f(x, y)$ on one of its variables, e.g. y, then switches the dependency from said variable to its conjugate:

$$\begin{aligned} dg &= d(f - vy) = df - vdy - ydv = udx + vdy - vdy - ydv \\ &= udx - ydv \end{aligned}$$

resulting in a function $g(x, v)$ depending on v instead of y. The most classical example of this formulation in statistical physics can be seen in the derivation of the *Helmholtz free energy.*

The *Helmholtz free energy* of a system is defined as the Legendre transform of its internal energy $F \equiv U - TS$ (Claudius, 2017). Since $S = -(\partial F/\partial T)_{V,N}$, we shift the dependency of U on the entropy to its conjugate T :

$$dU = TdS - PdV, \quad d(TS) = SdT + TdS \quad \text{hence} \quad dU = d(TS) - SdT - PdV, \tag{1.25}$$

and therefore

$$\begin{aligned} dF \equiv d(U - TS) &= \underbrace{dU - TdS}_{=-PdV} - SdT \qquad (1.26) \\ &= -SdT - PdV. \qquad (1.27) \end{aligned}$$

In the last section below, we present the physics derivation of this result from first principles, expanding on the notes of (Claudius, 2017).

1.4.3 The free energy from first principles

The first law of thermodynamics stipulates that in a closed system, such as an ideal gas with fixed number of particles, the internal energy of the system satisfies

$$dU = \underbrace{\delta Q}_{\text{energy flowing } \textit{into} \text{ the system as heat}} + \underbrace{\delta W}_{\text{work done } \textit{on} \text{ the system}} , \tag{1.28}$$

while the second law states that if the system is undergoing a reversible process, then $\delta Q = TdS$ and $\delta W = -PdV$, hence $dU = TdS - PdV$ which yields a potential with the entropy and volume as independent variables $U(S, V)$.

In a more general sense, we have *the basic thermodynamic relation* (*BTR* for short)

$$dU = TdS + \sum_{i}^{k} F_i dq_i \tag{1.29}$$

where $\{F_i, q_i\}$ are the pairs of conjugate variables characterizing the system.

In a magnetic system such as a spin glass, we have $\{F, q\} \equiv \{B, m\}$ where B is the uniform magnetic field and m is the magnetization of the system, whereas in a gas $\{F, q\} \equiv \{-P, V\}$ (Claudius, 2017). Note that the pair of conjugates always involve an *intensive* variable, independent of the size of the system (i.e. of order $\mathcal{O}(1)$), such as the pressure, and its *extensive* conjugate, such as the volume, which does depend on the size and hence is generally of order $\mathcal{O}(N)$.

Back to the ideal gas scenario, as we have shown from first principles, $\{F, q\} = \{-P, V\}$. Now suppose that the system depends on an external variable x, in the sense that the energy function becomes $E : \mathcal{R} \times \mathcal{X} \mapsto \mathbf{R}$ where $\mathcal{R}$ is the state space and $\mathcal{X}$ is the domain of the external variable x, hence

$$\frac{\partial \log \mathcal{Z}}{\partial x} = \frac{1}{\mathcal{Z}} \sum_r -\beta \frac{\partial E_{r,x}}{\partial x} e^{-\beta E(r,x)} = \frac{1}{\beta} \langle \frac{\partial E_{r,x}}{\partial x} \rangle. \tag{1.30}$$

Let $X \equiv \frac{\partial E_{r,x}}{\partial x}$, then the log-partition function satisfies

$$dlog\mathcal{Z} = \underbrace{\frac{\partial \log \mathcal{Z}}{\partial \beta}}_{=-U} d\beta + \underbrace{\frac{\partial \log \mathcal{Z}}{\partial x}}_{=\beta X} dx. \tag{1.31}$$

Hence,

$$Ud\beta = -dlog\mathcal{Z} + \beta X dx \tag{1.32}$$

and since $d(\beta U) = \beta dU + Ud\beta$, we have

$$- dlog\mathcal{Z} + \beta X dx = d(\beta U) - Ud\beta \tag{1.33}$$

Therefore

$$dU = \frac{1}{\beta} d(log\mathcal{Z} + \beta U) - X dx. \tag{1.34}$$

We recall that in the thermodynamic limit (as $N \longrightarrow \infty$), the following BTR holds

$$dU = TdS - Xdx \quad \text{and since} \quad dU = \frac{1}{\beta}d(log\mathcal{Z} + \beta U) - Xdx, \tag{1.35}$$

the entropy is given by

$$dS = \frac{1}{T}\underbrace{\frac{1}{\beta}}_{\equiv kT} dlog\mathcal{Z} + \frac{U}{T} = klog\mathcal{Z} + \frac{U}{T}. \tag{1.36}$$

Now that we have all the ingredients we can straightforwardly compute the free energy. By integrating over the differential of the entropy we have

$$S = k\log\mathcal{Z} + \frac{U}{T} + \underbrace{c_0}_{\text{some constant}}, \tag{1.37}$$

culminating in the free energy relation

$$-kT\log\mathcal{Z} = U - TS \equiv F. \tag{1.38}$$

Chapter 2

Dynamics: the study of time-evolution of glassy systems

Chapter organization

We start with a brief note on the historical development of the theory of spin glasses building up to mean field models (**2.1**). Then, we discuss a crucial condition called *self-averaging*, that any satisfactory theory of spin glasses needs to satisfy (**2.2**). Afterwards, we introduce a Markov chain called *Glauber dynamics*, that is meant to model the time evolution of spin glasses (**2.3.1**), and briefly discuss a slowness (or *freezing*) phenomenon that is characteristic of Glauber dynamics in the spin glass phase (**2.3.2**). Subsequently, we give a preliminary discussion of a certain property of the Gibbs measure at low temperature, called *the pure state decomposition*, where the Gibbs measure can be approximated by a convex combination of some quantities (**2.3.3**), while leaving some details for the last two chapters. Then, building up on the discussion so far, we introduce a new order parameter that is better suited for spin glass models than those commonly used for other magnetic alloys, such as (anti)ferromagnets (**2.4**). Finally, we close the chapter with a brief discussion of different temperature-dependent timescales in which the system reaches equilibrium (**2.5**).

2.1 Introduction

In contrast to crystals that are characterized by atoms located at regular intervals, forming a lattice in $\mathbf{Z}^3$, the random locations of the magnetic moments in a real spin glass are typically very irregular. Therefore, a physically realistic model would have the spin locations distributed according to a Poisson point-process in $\mathbf{R}^3$ with *spherical spins* i.e. with spin orientations in $\{\sigma_i \in \mathbf{R}^3 : ||\sigma_i||_2 = 1\}$.

However, this model poses hard technical difficulties and is far from being approachable. Therefore, we resort to a series of simplifications, the first of which is to move from euclidean space to the lattice $\mathbf{Z}^d$ and start by considering the two dimensional case. Fortunately, convincing experimental evidence points to the fact that quite different glassy systems seem to exhibit the same qualitative critical behaviour: susceptibility cusp, long-range correlation, frozen magnetization, etc...

This apparent universality suggests that it is possible to capture the essential physics of spin glasses by starting with a very simple model provided it incorporates two main features present in all systems displaying glassy behaviour, namely disorder and competing interactions.

The earliest attempt at a working model of spin glasses started by a simple description where each spin sits at a random position and interacts with the others through the oscillating RKKY interactions (Klein and Brout, 1963). The disorder is thus described by a set of random variables $\{\epsilon_{ij}\}$ indicating the presence of a spin in site $(i, j) \in \mathbf{Z}^2$. Since these occupation variables allow us to determine the distances between spins and hence the values of the interactions, each spin glass sample is completely determined by the realization of $\{\epsilon_{ij}\}$.

And although this model incorporates both above features, it misses an important property of physical spin glasses, namely *isotropy.* In a nutshell, isotropy measures the dependence of the response of a system on the direction in which an external magnetic field is trying to steer it in.

More specifically, a system is *isotropic* if there is no preferential orientation for its magnetic moment, upon the application of an external magnetic field, whereas in a highly *anisotropic* system, there exists two equally favourable opposite orientations forming the *easy axis* along which an external magnetic field will have an easier time magnetizing the spins than along any other directions. Considering that most real spin glasses are highly anisotropic, we can further simplify our model by considering binary or *Ising* spins $\sigma_{ij} \in \{\pm 1\}$ for all $(i, j) \in \mathbf{Z}^d$ (Binder et al., 2008).

Glassy behaviour has been observed in a wide variety of systems with interactions very different from the RKKY functions. If we aim to provide a toy model general enough to describe an array of glassy systems, we should choose a more universal interaction potential, one which preferably facilitates theoretical treatments while still retaining the macroscopic features of the system. In this case, we can then consider the location of the spins non-random, i.e. with spins occupying every point of the lattice, and compensate for the loss of spatial disorder by the random oscillating interactions.

Since it is precisely the cooperative microscopic phenomena that give rise to the peculiar macroscopic behaviour, this is a natural substitution. In fact, as far as the magnetic behaviour is concerned, the spatial randomness of the spins comes into play only through the deterministic RKKY interaction function. In this sense, the spatial randomness of the spins is baked into the random interactions. Furthermore, given that far off spins have vanishing interactions, we can restrict $\mathcal{J}$ to nearest neighbors interactions, i.e. $\mathcal{J} = \{J_{\mathbf{a},\mathbf{b}} : \mathbf{a}_1 = \mathbf{b}_1 \pm 1, .., \mathbf{a}_d = \mathbf{b}_d \pm 1\}$, which leads us to the *Edwards-Anderson* model. Here and throughout the thesis we will denote the Normal distribution by $\mathcal{N}$, and its variance by Δ.

Definition 9. *Consider a sequence of consecutive integers of length* $N : \mathcal{B}_N \equiv \{a, \ldots, a+N-2, a+N-1\}$, *starting from an arbitrary integer* $a \in \mathbf{Z}$. *The* $k-$*dimensional Edwards-Anderson model is an* $N-$*particle spin system, where each individual spin* σ_{i} *is identified with a vector* $\mathrm{i} \in \mathcal{B}_N^k$ *locating its position inside the* $k-$*dimensional subset* *of the integer lattice $\mathcal{B}_N^k \subset \mathbf{Z}^k$, such that the set of couplings $\mathcal{J}$ consists of:*

$$J_{\mathrm{i,j}} \overset{iid}{\sim} \mathcal{N}(0,\Delta).\, \mathbf{1}\big\{(\mathrm{i,j}) \in \mathcal{B}_N^k \times \mathcal{B}_N^k :\ ||\mathrm{i,j}||_2 = 1\big\}, \tag{2.1}$$

meaning that all the couplings are zero except for nearest neighbors pairs on the lattice. The Hamiltonian of the system is then given by:

$$\mathcal{H}(\sigma) = -\sum_{\mathrm{i,j}\in\mathcal{B}_N^k} J_{\mathrm{i,j}}\sigma_{\mathrm{i}}\sigma_{\mathrm{j}} - B\sum_{\mathrm{i}\in\mathcal{B}_N^k}\sigma_{\mathrm{i}}. \tag{2.2}$$

In the large N limit, the system spans the entire $k-$dimensional integer lattice $\mathbf{Z}^k$.

If the aim is to provide a general model for glassy systems, the macroscopic observables should not depend strongly on the realization of the disorder $\underline{\mathcal{J}}$ in a given sample. This is what is referred to as *self-averaging.*

2.2 Self-averaging

To follow conventional statistical physics notation, we introduce the following terminology:

- *Observables* are physical quantities which can measured though experiments. Formally speaking, an observable is a real valued function on the state space $\mathcal{O} : \Sigma \mapsto \mathbf{R}$ and its macroscopic limit is referred to as a thermodynamic obervable and is defined on the large N limit of the state space, e.g. let $\Sigma_N \equiv \mathcal{X}^N, \mathcal{O}_N : \Sigma_N \mapsto \mathbf{R}$, we are ultimately interested in quantities like $\lim_{N\to\infty}\mathcal{O}_N(\sigma)$ for some $\sigma \in \Sigma_N$. If a quantity of interest cannot be observed through experiments, we will instead refer to it as a *potential.* For example, the specific heat $C \equiv \partial\mathcal{H}/\partial\beta$ can be observed through experiments and is thus an observable, while the entropy $S(\beta) \equiv -\sum_{\sigma\in\Sigma}\mu_\beta(\sigma)\log\big(\mu_\beta(\sigma)\big)$ cannot be inferred through purely experimental means and is therefore referred to as a potential.

- *Disorder* (as in *sample dependent disorder*) is always meant to refer to the randomness of the couplings $\{J_{ij}\}$. As previously noted, when referring to an $N-$particle spin system with a given Hamiltonian (that depends on $\{J_{ij}\}$) we always assume that the couplings are fixed such that $\mathcal{H}$ is a function of only σ. However, as we discussed in the previous section, the $N-$particle system is supposed to model a generic spin glass sample, whose characteristics (e.g. energy landscape, thermodynamic potentials, etc) are uniquely determined by a set of fixed couplings $\{J_{ij}\}$. Hence, we generate the couplings according to a distribution that mirrors said characteristics and some more general conditions, the most important of which is *self-averaging*, that we will expand upon below.

Given these 2 sources of randomness: *i*) the distribution of the couplings $\{J_{ij}\}$, and *ii*) the distribution of μ_β with fixed $\underline{\mathcal{J}}$, to distinguish between the two, here and throughout the thesis, we follow the standard statistical physics notation (Mezard et al., 1987):

- We underline a variable (e.g. $\underline{\sigma}$) to signal that it is fixed.

- The expectation w.r.t. μ_β with fixed couplings $\underline{\mathcal{J}}$ will be denoted by brackets $\langle\rangle$.
- Moreover, the expectation of some observable over the distribution of the couplings $\{J_{ij}\} \overset{iid}{\sim} \Psi$ will be denoted by an overline, and will be referred to as the *average over the disorder*. For example, the disorder-averaged total magnetization is denoted:

$$\overline{m_k} \equiv \overline{\langle\sigma_k\rangle} \equiv \mathbf{E}_{J_{ij} \overset{iid}{\sim} \mathcal{N}(0,\Delta)}\big[\langle\sigma_k\rangle\big] \equiv \mathbf{E}_{J_{ij} \overset{iid}{\sim} \mathcal{N}(0,\Delta)}\big[\sum_{\sigma_k \in \{-1,1\}} \sigma_k \mu_\beta^{(i)}(\sigma_k)\big], \tag{2.3}$$

  where $\mu_\beta^{(k)}(\sigma_k = .)$ is the marginal probability of the k^{th} spin, and $\mathcal{N}(0,\Delta)$ is the zero-mean Gaussian distribution with variance Δ.

Since the entropy is an extensive potential (see appendix of ch1), the free energy $F \equiv U - TS \propto S$ is expected to be of order $\mathcal{O}(N)$ (Claudius, 2017). Given a $N-$particle spin system with the usual Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1 \leq i < j \leq N} \underline{J}_{ij}\sigma_i\sigma_j - B\sum_{i \in [N]} \sigma_i$ and fixed (i.e. non random) interactions $\underline{\mathcal{J}}$, we define the free energy of the system to be:

$$F_N(\beta) = -\frac{1}{\beta} log\mathcal{Z} \equiv -\frac{1}{\beta} log\Big(\sum_{\sigma \in \Sigma} e^{-\beta\mathcal{H}(\sigma)}\Big). \tag{2.4}$$

The intensive (i.e. $\mathcal{O}(1)$) version of the free energy is the *free energy density*, where *"density"* is meant in the sense of the amount of free energy per spin, in the large N limit:

$$f(\beta) = \lim_{N \longrightarrow \infty} \frac{1}{N\beta} F_N(\beta). \tag{2.5}$$

In a satisfactory theory of spin glasses, the free energy density should be the same across different sample realizations $\underline{\mathcal{J}}$. Since the goal is to characterize the behaviour of a typical spin glass sample, the free energy should converge (in mean) to a unique limit $f(\beta)$, regardless of the realization of the sample-dependent disorder $\underline{\mathcal{J}}$. In that case, the potential is said to be *weakly dependent on the disorder* or *self-averaging* (Castellani and Cavagna, 2005).

The free energy is self-averaging if it is well concentrated around its mean. In other words, the distribution of $-log\mathcal{Z}/\beta$ whose only source of randomness is from the couplings $\{J_{ij}\}$, should be sharply peaked around its mean $\overline{F_N(\beta)} \equiv -\overline{log\mathcal{Z}}/\beta$, with vanishing fluctuations in the large N limit:

$$\lim_{N \longrightarrow \infty} \overline{F_N(\beta)^2} - \overline{F_N(\beta)}^2 = 0. \tag{2.6}$$

Hence, we need to define a disorder distribution $\mathcal{J} \sim \Psi$ such that $F_N(\beta)$ is well concentrated. Assuming that all interactions are independent and identically distributed, an analytically convenient choice for Ψ is the zero mean Gaussian with $1/N$ variance, or a binomial with $J_{ij} = \pm 1/N$ with equal probability for faster computations using the multi-spin technique, for more details see chapter 4 in (Binder et al., 2008).

Either way, as long as the interactions are of order $\mathcal{O}(1/N^{\frac{1}{2}})$, the existence of the limiting free energy density should only depend on the first two moments of J_{ij} (Mezard et al., 1987).

2.3 Time evolution of a spin system

2.3.1 Glauber dynamics

Definition 10. *A discrete-time Markov chain is a sequence of random variables: $x_1, x_2, \ldots x_T$, representing the state of the Markov chain at each time-step t, where for all $1 < t \leq T$, the probability that x_t is equal to some value $\underline{x}_t$ depends only on the previous state x_{t-1}:*

$$\mathbf{P}[x_t = \underline{x}_t |\, x_1 = \underline{x}_1, x_2 = \underline{x}_2, \ldots x_{t-1} = \underline{x}_{t-1}] = \mathbf{P}[x_t = \underline{x}_t |\, x_{t-1} = \underline{x}_{t-1}]. \tag{2.7}$$

We assume that the variables $\{x_t\}_{t\in[T]}$ share the same support Σ that we call the state space of the Markov chain.

The simplest picture of the time evolution of an N-particle spin system can be illustrated by a single-spin flip Markov chain called *Glauber dynamics* (also known as *the heat bath algorithm*), that we describe below. Since we are considering a sequence of states $\sigma^{(t)} \in \Sigma$ for $0 \leq t \leq T$, to distinguish the individual spin indices $k \in [N]$ from the time index t of the Markov chain, we will write $\sigma_k^{(t)}$ to denote the value of the k^{th} spin of the state $x^{(t)} \in \Sigma$, for the remainder of this section.

Definition 11 (Mezard and Montanari, 2009)**.** *Consider an $N-$particle spin system $\sigma \sim \mu_\beta$ with some Hamiltonian $\mathcal{H}(\sigma)$ and whose state space is Σ. Glauber dynamics is a discrete-time Markov chain that is conditioned on an initial state sampled from the target distribution: $\sigma^{(0)} \sim \mu_\beta$. Given the current state of the Markov chain $\sigma^{(t)} = \underline{\sigma}^{(t)}$, the next state is generated as follows:*

1. *Propose the next state uniformly at random from the set of immediate neighbors of the current state: $\sigma^{(t+1)} \sim \mathcal{U}(\mathcal{N}(\underline{\sigma}^{(t)}))$ where $\mathcal{N}(\underline{\sigma}^{(t)}) \equiv \{y \in \Sigma :\ d(y, \underline{\sigma}^{(t)}) = 1\}$, with $d(,)$ begin the Hamming distance.*

2. *Accept the proposed state according to the probability:*

$$\alpha(\underline{\sigma}^{(t+1)}, \underline{\sigma}^{(t)}) \equiv \min\left\{1, e^{-\beta(\mathcal{H}(\underline{\sigma}^{(t+1)}) - \mathcal{H}(\underline{\sigma}^{(t)}))}\right\}.$$

Algorithm 1: The heat bath algorithm

```
Result: $\sigma^{(T)} \overset{approx}{\sim} \mu_\beta(.)$
$\underline{\sigma}^{(0)} \leftarrow \tau$;
for $t = 1, ...T$ do
    draw $i \sim \{1 \dots N\}$ uniformly at random;
    $\gamma \leftarrow (\underline{\sigma}_1^{(t)} \dots \underline{\sigma}_{i-1}^{(t)}, -\underline{\sigma}_i^{(t)}, \underline{\sigma}_{i+1}^{(t)} \dots \underline{\sigma}_N^{(t)})$;
    $\Delta E_t \leftarrow \mathcal{H}(\gamma) - \mathcal{H}(\underline{\sigma}^{(t-1)})$;
    $\alpha(\gamma, \underline{\sigma}^{(t-1)}) \leftarrow e^{-\beta max\left\{\Delta E_t, 0\right\}}$;
    draw $u \sim [0,1]$ uniformly at random;
    if $u \leq \alpha(\gamma, \underline{\sigma}^{(t-1)})$ then
        $\underline{\sigma}^{(t)} \leftarrow \gamma$;
    else
        $\underline{\sigma}^{(t)} \leftarrow \underline{\sigma}^{(t-1)}$
    end
end
```

Definition 12. *Given an N-particle spin system $\sigma \sim \mu_\beta$, and an observable $\mathcal{O} : \Sigma \mapsto$ $\mathbf{R}$, we define the time average of $\mathcal{O}(t) \equiv \mathcal{O}(\sigma^{(t)})$ as its expectation w.r.t. Glauber dynamics starting from $\sigma^{(0)} \sim \mu_\beta$, that we denote by brackets:*

$$\langle \mathcal{O}(t) \rangle_{\underline{\sigma}^{(0)}} = \mathbf{E}_{Glauber}\left[\mathcal{O}(\sigma^{(t)}) | \sigma^{(0)} = \underline{\sigma}^{(0)}\right] = \sum_{\underline{\sigma}^{(t)} \in \Sigma} \mathcal{O}(\underline{\sigma}^{(t)})\, \mathbf{P}_{Glauber}\left[\sigma^{(t)} = \underline{\sigma}^{(t)} | \sigma^{(0)} = \underline{\sigma}^{(0)}\right]. \tag{2.8}$$

To make matters more explicit on a simple example, consider the time average of the value of the k^{th} spin after t Glauber steps starting from the realization of $\sigma^{(0)} \sim \mu_\beta$:

$$\begin{aligned}
\langle \sigma_k(t) \rangle_{\underline{\sigma}^{(0)}} &= \sum_{\underline{\sigma}^{(t)} \in \Sigma} \underline{\sigma}_k^{(t)}\, \mathbf{P}_{Glauber}\left[\sigma^{(t)} = \underline{\sigma}^{(t)} | \sigma^{(0)} = \underline{\sigma}^{(0)}\right] \\
&= \sum_{\underline{\sigma}^{(t)} \in \Sigma} \underline{\sigma}_k^{(t)}\, \frac{\mathbf{P}_{Glauber}\left[\sigma^{(t)} = \underline{\sigma}^{(t)}, \sigma^{(0)} = \underline{\sigma}^{(0)}\right]}{\mu_\beta(\sigma^{(0)} = \underline{\sigma}^{(0)})} \\
&= \frac{1}{\mu_\beta(\sigma^{(0)} = \underline{\sigma}^{(0)})} \sum_{\underline{\sigma}^{(1)} \dots \underline{\sigma}^{(t)} \in \Sigma} \underline{\sigma}_k^{(t)} \left[\prod_{s=2}^{t} p_{Gb}\left[\underline{\sigma}^{(s)} | \underline{\sigma}^{(s-1)}\right]\right] p_{Gb}\left[\underline{\sigma}^{(1)} | \underline{\sigma}^{(0)}\right],
\end{aligned}$$

where the transition probability of the Markov chain along Glauber dynamic is as previously discussed: $p_{Gb}(\underline{\sigma}^{(t)} | \underline{\sigma}^{(t-1)}) \equiv k(\sigma^{(t)} = \underline{\sigma}^{(t)} | \sigma^{(t-1)} = \underline{\sigma}^{(t-1)}) = \big(\mathbf{1}\big\{d(\underline{\sigma}^{(t)}, \underline{\sigma}^{(t-1)}) = 1\big\}/N\big).\ \min\{1, e^{\beta(\mathcal{H}(\underline{\sigma}^{(t)}) - \mathcal{H}(\underline{\sigma}^{(t-1)}))}\}$.

2.3.2 Breaking of ergodicity in the spin glass phase

Definition 13 (Fischer and Hertz, 1991)**.** *Given an N-particle spin system, we define the energy spectrum of a particular value e, as the number of states with the corresponding energy that we denote by $\mathcal{N}_\delta(e)$:*

$$\mathcal{N}_\delta(e) \equiv\ |\{\sigma \in \Sigma :\ e - \delta \leq \mathcal{H}(\sigma) \leq e + \delta\}|, \tag{2.9}$$

for some $\delta > 0$.

The ergodic hypothesis states that the amount of time spent in a given state (in terms of visiting frequency) is proportional to the energy spectrum of that state. In other words, given enough time, the system will explore the entirety of the state space Σ in such a way that the average over time is equal to the average over states:

$$\underset{T \longrightarrow \infty}{lim} \frac{\sum_{t=0}^{T} \mathcal{O}(\sigma^{(t)})}{T} = \sum_{\underline{\sigma} \in \Sigma} \mu_\beta(\underline{\sigma})\mathcal{O}(\underline{\sigma}) \quad \text{for all} \quad \mathcal{O} : \Sigma \mapsto \mathbf{R}. \tag{2.10}$$

When this equality holds with T growing polynomially in N, we say that the system is *ergodic*.

A *glassy system* is defined through its dynamical properties (i.e. changes over time). In the spin glass phase, the relaxational dynamic (i.e. the speed in which the system reaches equilibrium) is exponentially slow (Mezard et al., 1987), such that the system is essentially frozen and the positions of spins change extremely slowly as a function of time.

The exponentially slow convergence to equilibrium is caused by the fact that the energy landscape contains wells within wells of low energy, such that if we keep flipping spins in an attempt to reach one of the *global* minima of $\mathcal{H}(\sigma)$, according to changes in energy (as in Glauber dynamics), the system will keep running into energy barriers (i.e. large differences in energy), such that the acceptance probability $\alpha(,)$ stays prohibitively small, and the system will keep rejecting proposal states thus remaining stuck in local minima. These wells trap the system for large time scales, and are therefore called *metastable states*.

The more critical points the Hamiltonian has, the more rugged or *complex* the energy landscape is (see figure below). Moreover, since the number of local minima depends on the couplings, it is the frustration phenomenon that is responsible for the proliferation of metastable states.

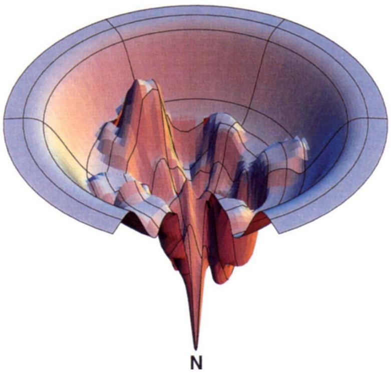


Figure 2.1: Rugged landscape (Dill and Chan, 1997).

2.3.3 The pure state decomposition

Throughout the thesis, there will be two major points of view: *statics* and *dynamics*. The statics perspective, that will be expanded upon further in chapter 4, deals with the equilibrium properties of μ_β, i.e. the study of the Gibbs measure in the zero temperature limit, while the dynamics point of view (that we expand further in ch.5) concerns the evolution of the system towards equilibrium. Although there are a number of models for the time evolution of spin systems (s.a. Langevin dynamics), for the remainder of the discussion, we will assume that it transitions according to Glauber dynamics. Following statistical physics terminology, we will refer to the Gibbs measure at low temperature $\mu_\beta, \beta \gg 1$ as the *equilibrium measure*.

From the statics point of view, in the low temperature limit, the exponentially high energy barriers correspond to the breaking of the support of the Gibbs distribution into well separated high probability density regions taking almost all of the probability mass, that we call *pure states*: $\Sigma = \left(\bigsqcup_{i\in[\eta]} \alpha_i\right) \bigsqcup \mathcal{S}$, where η is the number of pure states $\{\alpha_i\}$ and $\mathcal{S}$ are the configurations in between (Fischer and Hertz, 1991).

This phenomenon is called *the pure state decomposition* and is characteristic of the low temperature spin glass phase. The technical definition of the pure state decomposition lists a number of technical conditions that require prerequisite notions that we have not introduced yet. That being said, we will give a hand-wavy definition in terms of an approximation for the time being, and we will give the proper definition in chapter 5.

Definition 14. *Given an N-particle spin system $\sigma \sim \mu_\beta$, whose state space can be expressed as a disjoint union ($\sqcup$) of sets called pure states: α_i for $i \in [\eta]$ (where each set is a connected component of Σ in Hamming space) and the remaining states surrounding each pure state whose union is $\mathcal{S}$, such that $\Sigma = \left(\bigsqcup_{i\in[\eta]} \alpha_i\right) \bigsqcup \mathcal{S}$. The pure state decomposition refers to the exponential vanishing of $\mu_\beta(\mathcal{S})$ when $\beta \gg 1$, such that:*

$$\mu_\beta(\tau) \approx \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i} \mu_\beta^{\alpha_i}(\tau), \quad \textit{where} \quad \mu_\beta^{\alpha_i}(\tau) \equiv \frac{\mathbf{1}\{\tau \in \alpha_i\} e^{-\beta\mathcal{H}(\tau)}}{\mathcal{Z}_{\alpha i}},$$

$$w_{\alpha_i} = \frac{\mathcal{Z}_{\alpha_i}}{\sum_{\alpha_i : i\in[\eta]} \mathcal{Z}_{\alpha i}}, \quad \textit{and} \quad \mathcal{Z}_{\alpha_i} \equiv \sum_{\underline{\sigma}\in\alpha_i} e^{-\beta\mathcal{H}(\underline{\sigma})}.$$

Note that, the fact that the Gibbs measure can be approximated by a convex combination of pure states weights $\{w_{\alpha_i}\}$ follows by conditioning on pure states:

$$\begin{aligned} \mu_\beta^{\alpha_k}(\underline{\tau}) &\equiv \mathbf{P}_{\sim\mu_\beta}[\sigma = \underline{\tau} | \underline{\tau} \in \alpha_k] = \frac{\mathbf{P}_{\sim\mu_\beta}[\sigma = \underline{\tau}, \underline{\tau} \in \alpha_k]}{\mathbf{P}_{\sim\mu_\beta}[\alpha_k]} = \frac{\mathbf{1}\{\underline{\tau} \in \alpha_k\}.\mathbf{P}_{\sim\mu_\beta}[\sigma = \underline{\tau}]}{\mathbf{P}_{\sim\mu_\beta}[\sigma \in \alpha_k]} \\ &= \mathbf{1}\{\underline{\tau} \in \alpha_k\}.\frac{e^{-\beta\mathcal{H}(\underline{\tau})}}{\mathcal{Z}} \frac{1}{\sum_{\sigma\in\alpha_k} \frac{e^{-\beta\mathcal{H}(\sigma)}}{\mathcal{Z}}} = \frac{\mathbf{1}\{\underline{\tau} \in \alpha_k\}.e^{-\beta\mathcal{H}(\underline{\tau})}}{\mathcal{Z}_{\alpha_k}}. \end{aligned}$$

Under such conditions, we find that $\lim_{T\to\infty} \sum_{t=0}^{T} \mathcal{O}(\sigma_t)/T \approx \sum_{\underline{\sigma}\in\alpha_i} \mu_\beta(\underline{\sigma})\mathcal{O}(\underline{\sigma})$, where α_i is the pure state corresponding to the energy valley in which the system is found originally $\underline{\sigma}^{(0)}$ (Fischer and Hertz, 1991). This phenomenon is referred to as *ergodicity breaking.* In this case, we say that while the system is no longer ergodic on Σ, it is *ergodic within the pure state* α_i.

2.4 A new order parameter

As discussed in the previous chapter, the ferromagnetic phase transition is marked by a nonzero value of all local magnetizations. Recall that in the low temperature regime, the Gibbs measure is concentrated on the lowest energy configurations such that for very low temperatures $\beta^* \gg 1$, the marginal law of the i^{th} spin is

$$\mu^{(i)}_{\beta^*}(\tau_i) \approx \frac{1}{\mathcal{Z}} \sum_{\underline{\tau}_{[N]\setminus i}} e^{-\beta^* \mathcal{H}(\underline{\tau}_{[N]\setminus i}, \tau_i)}.\ \mathbf{1}\Big\{(\underline{\tau}_{[N]\setminus i}, \tau_i) \in \underset{\underline{\sigma}\in\Sigma}{argmin}\ \mathcal{H}(\underline{\sigma})\Big\}. \tag{2.11}$$

We say that the system is symmetric with respect to a set of outcomes if they are equally likely under μ_β. In a ferromagnet with the general Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\underline{\sigma}_i\underline{\sigma}_j - B\sum_{i\in[N]} \underline{\sigma}_i$, if $B \neq 0$, the ground state of the system is

$$\underset{\underline{\sigma}\in\Sigma}{argmin}\ \mathcal{H}(\underline{\sigma}) = \underset{\underline{\sigma}\in\Sigma}{argmin}\ \Big[-\sum_{1\leq i<j\leq N} J_{ij}\underline{\sigma}_i\underline{\sigma}_j - B\sum_{i\in[N]}\underline{\sigma}_i\Big] = \big\{sgn(B)\big\}^N, \tag{2.12}$$

Hence, as the temperature decreases below the critical temperature, the magnetization, which was zero at large temperatures by virtue of the symmetry of $\mu_{\beta\approx 0}(.)$ with regards to all states, takes a nonzero value. Since $\mu^{(i)}_{\beta^*\beta^*}(\tau_i) = \delta_{\tau_i, sgn(B)}$, we have $m_i = sgn(B)$, thus the ± symmetry of the high temperature regime is broken .

At zero magnetic field, the situation is complicated by the fact that there are two equally likely ground states:

$$min_{\underline{\sigma}\in\Sigma}\ \mathcal{H}_{B=0}(\underline{\sigma}) = \mathcal{H}_{B=0}(\{+1\}^N) = \mathcal{H}_{B=0}(\{-1\}^N). \tag{2.13}$$

From the dynamics point of view, as the temperature decreases, the energy landscape which was essentially flat at high temperatures, takes the shape of a double well separated by an energy barrier of order $\mathcal{O}(e^{\alpha N})$. Hence for $N \gg 1$, the landscape is essentially partitioned into two pure states $\{\alpha_1, \alpha_2\}$.

At the critical temperature, the system can go into either well with equal probability and assume the order of whatever pure state it finds itself in, meaning that all local magnetizations take the value of the corresponding ground state $m_k^{\alpha_i} \equiv \langle\sigma_k\rangle_{\alpha_i} = (argmin_{\sigma\in\alpha_i}\mathcal{H}(\sigma))_k$ and remain stable for arbitrarily long timescales. Considering that the system is symmetric with respect to either outcome and only one of them can occur, this phenomenon is referred to as *spontaneous symmetry breaking.*
Hence, in the case of the ferromagnet, ergodicity breaking simply corresponds to symmetry breaking of the overall magnetization. Therefore, we naively expect that the same order parameter should detect the spin glass transition.

As discussed previously, a popular choice for the disorder distribution $\mathcal{J} \overset{iid}{\sim} \Psi$ is the zero mean Gaussian or a symmetric binomial sometimes referred to as the $\pm\mathcal{J}$ model where $J_{ij} = \pm\gamma$ with equal probabilities. In both cases, we run into the same issue:

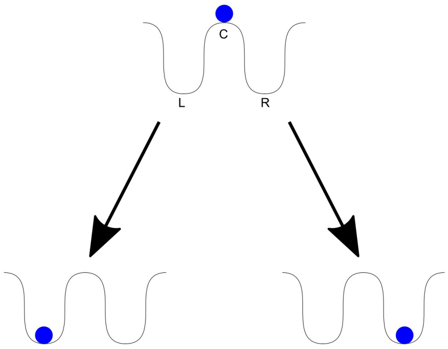


Figure 2.2: Spontaneous symmetry breaking (Wikipedia, n.d.)

the order parameter stays zero below the critical temperature, failing to detect the glassy transition (Castellani and Cavagna, 2005).

More precisely, we prove the following result, hinted at in (Mezard et al., 1987):

Theorem 3. *Consider an $N-$particle spin system with the simplified Hamiltonian: $\mathcal{H}(\sigma) = -\sum_{1 \leq i < j \leq N} J_{ij}\sigma_i\sigma_j$, where the couplings are either: i) Gaussian with variance $\Delta:\ J_{ij} \overset{iid}{\sim} \mathcal{N}(0,\Delta)$, or ii) symmetric binomial where $J_{ij} = \pm\gamma$ with equal probabilities. Then, the disorder-averaged local magnetization for any spin $k \in [N]$ in both cases satisfies:*

$$\overline{m_k} \equiv \mathbf{E}_{J_{ij} \overset{iid}{\sim} \mathcal{N}(0,\Delta)} \big[\sum_{\sigma_k \in \{-1,1\}} \sigma_k \mu_\beta^{(i)}(\sigma_k) \big] = 0, \tag{2.14}$$

where $\mu_\beta^{(k)}(\sigma_k = .)$ is the marginal probability of the k^{th} spin.

Proof. We will follow the conventional notation in spin glass theory (Mezard et al., 1987; Fischer and Hertz, 1991) by denoting the expectation with regards to the distribution of $\{J_{ij}\}$ with an overline:

$$\overline{m_k} = \sum_{\underline{\sigma}_{[N]\backslash\{k\}}} \overline{\mu_\beta(\underline{\sigma_{[N]\backslash k}}, \sigma_i = 1) - \mu_\beta(\underline{\sigma_{[N]\backslash k}}, \sigma_i = -1)} \tag{2.15}$$

$$\propto \sum_{\underline{\sigma}_{[N]\backslash\{k\}}} \prod_{(i,j) \in \mathcal{J}\backslash(.,k)} \int e^{\beta J_{ij}\underline{\sigma}_i\underline{\sigma}_j} dJ_{ij} \prod_{l \in [N]} \left(\int e^{\beta J_{lk}\underline{\sigma}_l} dJ_{lk} - \int e^{-\beta J_{lk}\underline{\sigma}_l} dJ_{lk} \right). \tag{2.16}$$

Throughout the thesis, a useful identity that we will often apply is the Gaussian identity:

$$\mathbf{E}_{\mathcal{X} \sim \mathcal{N}(0,\Delta)} \Big[e^{\Delta X} \Big] = \exp\Big(\frac{\Delta\lambda^2}{2}\Big).$$

In the Gaussian case, the term in parenthesis is zero

$$\int e^{\beta J_{lk}\underline{\sigma}_l} dJ_{lk} - \int e^{-\beta J_{lk}\underline{\sigma}_l} dJ_{lk} = e^{\frac{\beta^2}{2}\Delta} - e^{\frac{\beta^2}{2}\Delta} = 0 \quad \text{for all} \quad l \in [N], \tag{2.17}$$

as it is in the symmetric binomial case

$$\int e^{\beta J_{lk}\underline{\sigma}_l} dJ_{lk} - \int e^{-\beta J_{lk}\underline{\sigma}_l} dJ_{lk} = \frac{e^{+\gamma\beta\underline{\sigma}_l}}{2} + \frac{e^{-\gamma\beta\underline{\sigma}_l}}{2} - \frac{e^{-\gamma\beta\underline{\sigma}_l}}{2} - \frac{e^{+\gamma\beta\underline{\sigma}_l}}{2} = 0. \quad \blacksquare \quad (2.18)$$

Therefore, the local magnetizations are equally zero and consequently the overall magnetization stays zero throughout both the high and low temperature regimes.

It is thus apparent that the site average of the expectation of individual spins with respect to the Gibbs distribution can not be used as an order parameter when the interaction distribution is an even function.

However, as discussed in the second chapter of (Fischer and Hertz, 1991), we can still take inspiration from the ferromagnetic situation and make ad-hoc attempts at breaking the ± symmetry of the marginals (i.e. make $\mu_\beta^{(k)}(1) \neq \mu_\beta^{(k)}(-1)$) by skewing the thermal average in two (equally non-rigorous) ways. We denote the configuration resulting from flipping the i^{th} spin of σ by $\sigma^{(i)} \equiv (\sigma_1, .. - \sigma_i, ..\sigma_N)$.

- We can restrict the expectation w.r.t. $\mu_\beta : \langle . \rangle$ to a portion of the state space $\mathcal{P} \subset \Sigma$ with a relatively homogeneous general orientation, by defining a renormalized Gibbs measure measure $\nu_\beta(\sigma) \equiv \mathbf{1}\{\sigma \in \mathcal{P}\}.e^{-\beta\mathcal{H}(\sigma)}/\sum_{\sigma\in\mathcal{P}} e^{-\beta\mathcal{H}(\sigma)}$, with respect to which the expectation produces a zero total magnetization: $m^{\mathcal{P}} \equiv 1/N \sum_{i\in[N]}\langle\sigma_i\rangle_\nu \neq 0$, even if some of the local magnetizations $\langle\sigma_i\rangle_\nu$ stay zero.
- Or, suppose that the energy landscape, i.e. the graph of $\mathcal{H}$ with fixed β, B and $\{J_{ij}\}$, consists of a double well centered around two energetically equal minima σ and $\tau = (\sigma_1, \dots, -\sigma_i, \dots, \sigma_N)$. Considering that $m_i = \mathcal{O}(\sqrt{N})$ (Mezard et al., 1987), we can add an external magnetic field $B \gg 1/\beta\sqrt{N}$ such that $\exp\{\beta B(m_\sigma - m_\tau)\} \gg 1$ making $\mu_\beta^i(\tau_i)$ negligible compared to $\mu_\beta^i(\sigma_i)$, which would in turn make $m_i \approx sgn(\sigma_i)$, hence allowing us to make some local magnetizations nonzero, as suggested in chapter 2 of (Fischer and Hertz, 1991).

This is, however, far too simplistic as it does not account for the remanence effects observed in spin glass experiments (Binder and Young, 1986), which seem to suggest the existence of a large number of pure states that are stable on very long timescales, below the critical temperature (i.e. that their Glauber dynamics are exponentially slow, this hand-wavy statement will be made formal in the last chapter).

Still from the statics point of view, we can accommodate this picture, that is, impose the existence of many quasi-stable states, by formulating the free energy as a function of the local magnetizations $F : (m_1, ..m_N) \mapsto \mathbf{R}$ as in (Fischer and Hertz, 1991), such that we may have many local minima, each of them satisfying

$$\frac{\partial F}{\partial m_i} = 0 \quad \text{for at least one spin } i \in [N] \text{ and } \quad \frac{\partial^2 F}{\partial m_i \partial m_j} \geq 0 \quad \text{for all } j \in [N]. \quad (2.19)$$

This evidently prevents us from using simple methods for breaking the ± symmetry such as skewing μ_β in favour of one of two minima, considering that the energy landscape is now highly non-trivial.

It is thus apparent that a linear combination of local magnetizations is not a suitable order parameter for the glass transition, which leads us to consider higher orders such as the squared magnetization

$$q \equiv \langle \sigma_i \rangle^2. \tag{2.20}$$

Note that in the rigorous sense, the pure state decomposition is an asymptotic property of the Gibbs measure (as we shall see in the ch.5), where μ_β becomes equal to the convex combination above as $N \to \infty$. By sweeping things under the rug, we write it as an approximation. However, for the remainder of this chapter, since we are mainly following (Fischer and Hertz, 1991), we will take it to be an equality for finite N, as is custom in the physics literature (see also (Mezard et al., 1987)). We show the following result, whose proof was omitted from (Fischer and Hertz, 1991).

Lemma 1. *Consider an N-particle spin system displaying a pure state decomposition $\Sigma = \left(\bigsqcup_{i\in[\eta]} \alpha_i\right) \bigsqcup \mathcal{S}$, where η is the number of pure states $\{\alpha_i\}$ and $\mathcal{S}$ are the configurations in between, and let $\mathcal{O} : \Sigma \mapsto \mathbf{R}$, the expectation of $\mathcal{O}(\sigma)$ w.r.t μ_β that we denote by brackets, then satisfies*

$$\langle \mathcal{O}(\sigma) \rangle = \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i} \langle \mathcal{O}(\sigma) \rangle_{\alpha_i}, \quad \textit{where} \quad \langle \mathcal{O}(\sigma) \rangle_{\alpha_i} \equiv \sum_{\sigma\in\Sigma} \mu_\beta^{\alpha_i}(\sigma) \mathcal{O}(\sigma), \tag{2.21}$$

with

$$\mu_\beta^{\alpha_i}(\tau) \equiv \frac{\mathbf{1}\{\tau \in \alpha_i\} e^{-\beta\mathcal{H}(\tau)}}{\mathcal{Z}_{\alpha i}}, \; w_{\alpha_i} = \frac{\mathcal{Z}_{\alpha_i}}{\sum_{\alpha_i : i\in[\eta]} \mathcal{Z}_{\alpha i}}, \; \textit{and } \mathcal{Z}_{\alpha_i} \equiv \sum_{\underline{\sigma}\in\alpha_i} e^{-\beta\mathcal{H}(\underline{\sigma})}. \tag{2.22}$$

Proof. Since we assume that $\mu_\beta(\mathcal{S}) = 0$, we have:

$$\begin{aligned}
\langle \mathcal{O}(\sigma) \rangle &= \sum_{\sigma\in\Sigma} \mu_\beta(\sigma)\mathcal{O}(\sigma) = \sum_{\alpha_i : i\in[\eta]} \Big(\sum_{\sigma\in\alpha_i} \mu_\beta(\sigma)\mathcal{O}(\sigma) \Big) \\
&= \sum_{\alpha_i : i\in[\eta]} \Big(\sum_{\sigma\in\Sigma} \mathbf{1}\{\sigma \in \alpha_i\}.\mu_\beta(\sigma)\mathcal{O}(\sigma) \Big) \\
&= \sum_{\alpha_i : i\in[\eta]} \Big(\sum_{\sigma\in\Sigma} \frac{\mathbf{1}\{\sigma \in \alpha_i\}.e^{-\beta\mathcal{H}(\sigma)}}{\mathcal{Z}} \mathcal{O}(\sigma) \Big) \\
&= \sum_{\alpha_i : i\in[\eta]} \Big(\sum_{\sigma\in\Sigma} \frac{\mathbf{1}\{\sigma \in \alpha_i\}.e^{-\beta\mathcal{H}(\sigma)}}{\mathcal{Z}_{\alpha_i}} \frac{\mathcal{Z}_{\alpha_i}}{\mathcal{Z}} \mathcal{O}(\sigma) \Big) \\
&= \sum_{\alpha_i : i\in[\eta]} \frac{\mathcal{Z}_{\alpha_i}}{\mathcal{Z}} \Big(\sum_{\sigma\in\Sigma} \frac{\mathbf{1}\{\sigma \in \alpha_i\}.e^{-\beta\mathcal{H}(\sigma)}}{\mathcal{Z}_{\alpha_i}} \mathcal{O}(\sigma) \Big) \\
&= \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i} \Big(\sum_{\sigma\in\Sigma} \mu_\beta^{\alpha_i}(\sigma)\mathcal{O}(\sigma) \Big) = \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i} \langle \mathcal{O}(\sigma) \rangle_{\alpha_i}. \quad \blacksquare
\end{aligned}$$

It then follows (see ch.2 of (Fischer and Hertz, 1991)) that:

$$q \equiv \langle\sigma_i\rangle^2 = \Big(\sum_{\alpha_k : i\in[\eta]} w_{\alpha_k}\langle\sigma_i\rangle\Big)^2 = \sum_{\alpha_k,\alpha_l : 1\leq k<l\leq\eta} w_{\alpha_k}w_{\alpha_l}\langle\sigma_i\ \rangle_{\alpha_k}\langle\sigma_i\ \rangle_{\alpha_l}. \tag{2.23}$$

In (Edwards and Anderson, 1975), the authors formulated a slightly different order parameter

$$q_{EA} \equiv \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i}\langle\sigma_i\rangle^2_{\alpha_i}. \tag{2.24}$$

The difference being that while the sum in q compromises inter-valley contributions in $\mathcal{S}$, q_{EA} does not. In fact, it only makes sense to speak of q_{EA} if the ergodicity is broken. When the system is ergodic, i.e. when there exists a unique pure state compromising all of Σ, the two order parameters are equivalent (Fischer and Hertz, 1991). Hence, depending on the height of the energy barriers, the difference between the two order parameters fluctuates between zero and some value, and can thus serve as a measure of the extent to which ergodicity is broken:

$$\Delta_q \equiv q_{EA} - q. \tag{2.25}$$

The physical significance of Δ_q as a measure of broken ergodicity is most clearly exemplified from the dynamical point of view. Which leads to the dynamical formulation of the Edwards-Anderson order parameter (Fischer and Hertz, 1991):

$$q^*_{EA} \equiv \underset{t\longrightarrow\infty}{lim}\ \underset{N\longrightarrow\infty}{lim}\ \overline{\langle\sigma_i(t+t_0)\sigma_i(t_0)\rangle}, \tag{2.26}$$

where the overline signifies the expectation w.r.t. the distribution of the couplings $\{J_{ij}\}$. Since the ergodic system explores the entirety of the state with a frequency proportional to the energy spectrum, it decorrelates from its initial state in finite time, such that $q^*_{EA} = 0$ (Fischer and Hertz, 1991).

Below the critical temperature, that is, at the onset of ergodicity breaking, the infinite ($N = \infty$) system is trapped by infinite energy barriers within the pure state containing its initial value. As $t \to \infty$, the system stays within a close Hamming distance to its initial state to which it stays correlated, hence $q^*_{EA} \neq 0$ (Mezard et al., 1987).

Therefore, q_{EA} can effectively serve as an order parameter for the spin glass transition, considering that it is zero throughout the paramagnetic high temperature phase and takes a nonzero value at the beginning of the spin glass phase, when the individual spins' values stay correlated with their initial values indefinitely. A sharper discussion of how fast an N-particle spin system decorrelates can be formulated though *the relaxation time* (which will be discussed in ch.5), and which essentially describes how many Glauber iterations it takes for the time average: $\langle\mathcal{O}(t)\mathcal{O}(0)\rangle_{\underline{\sigma}_0}$ to become arbitrarily small, thus permitting us to provide a quantitative characterization of the dynamics in the spin glass phase.

The equivalence between the static and dynamic formulation of q_{EA} comes from the fact that while the system is trapped within a pure state $\alpha_k \ni \sigma_0$, the ergodic hypothesis is still valid within it (Fischer and Hertz, 1991). Meaning that the system explores each energy level in $\{\mathcal{H}(\tau) : \tau \in \alpha_k\}$ with frequencies proportional to their energy spectrum such that

$$\underset{T \longrightarrow \infty}{lim} \frac{\sum_{t=0}^{T} \mathcal{O}(\sigma^{(t)})}{T} = \langle \mathcal{O} \rangle_{\alpha_k}. \tag{2.27}$$

It can therefore be shown that q_{EA} actually measures the squared local mangetization within pure states (Fischer and Hertz, 1991):

$$\underset{t \longrightarrow \infty}{lim} \underset{N \longrightarrow \infty}{lim} \overline{\langle \sigma_i(t+t_0)\sigma_i(t_0) \rangle} = \sum_{\alpha_k : k \in [\eta]} w_{\alpha_k} \langle \sigma_i \rangle_{\alpha_k}^2. \tag{2.28}$$

Note that the order in which the limits are taken is crucial, since:

1. We are interested in the dynamics of the macroscopic system.
2. The decomposition of the state space due to infinite energy barriers is only realized in the thermodynamic limit.

And although proper ergodicity breaking only takes place at $N = \infty$ and $\beta \geq \beta_c$ (the critical temperature), the introduction of Δ_q and its dependence on N permits discussing intermediate ergodicity regimes and their associated time-to-equilibrium timescales.

A spin glass sample is constrained by its disorder $\underline{\mathcal{J}}$ and as the temperature is lowered the system essentially becomes more and more constrained. The important realization is that, at fixed N, we can control the height of the energy barriers by varying β.

From an algorithmic perspective, the system's evolution to equilibrium, along the lines of Glauber dynamics, is essentially a stochastic local search for lower energy configurations, guided by the slopes of the energy landscape. In the zero temperature limit the stochastic problem reduces to a deterministic highly non-convex one of finding the ground state, i.e. the global minimum of the Hamiltonian, subject to a set of constraints instantiated by the sample dependant disorder $\underline{\mathcal{J}}$

$$\underset{\underline{\sigma} \in \Sigma}{argmin}\, \beta \Big[- \sum_{1 \leq i < j \leq N} J_{ij} \underline{\sigma}_i \underline{\sigma}_j - B \sum_{i \in [N]} \underline{\sigma}_i \Big] \quad \text{subject to} \quad \{J_{ij}\} = \underline{\mathcal{J}}. \tag{2.29}$$

Since the state of the system obeys μ_β, the constraints are softened by the temperature parameter, such that for high enough temperatures the system is satisfied by settling in a less than optimal state. As the temperature is lowered the system becomes more constrained by $\underline{\mathcal{J}}$ and the energy landscape gets more complex forming valleys within bigger valleys each centered around one of the many critical points of $\mathcal{H}_{\underline{\mathcal{J}}}$.

That is to say, that the origin of broken ergodicity, the freezing of local magnetizations and the non-trivial correlation structure all relate back to the large number of critical points of $\mathcal{H}_{\underline{\mathcal{J}}}$, produced by the competing interactions (recall that positive interactions result in at most two critical points), and the resulting highly non-convex domain.

2.5 Intermediate ergodicity regimes

In a large but finite system ($1 \ll N < \infty$) displaying pure state decomposition, energy barriers are finite and therefore, the system is allowed to escape its original pure state in finite time even below the critical temperature.

Considering the dramatic slowdown of dynamics caused by the pure state decomposition, it can be useful to distinguish between short, intermediate and long size-dependent timescales in order to characterize a finite system's time-dependant behaviour. We distinguish the finite time versions of the order parameters Δ_q and q_{EA} by adding the subscripts $^{(a)}$,$^{(b)}$ and $^{(c)}$ for short, intermediate and long timescales respectively, as illustrated in the figure below by (Liu SQ et al., 2012).

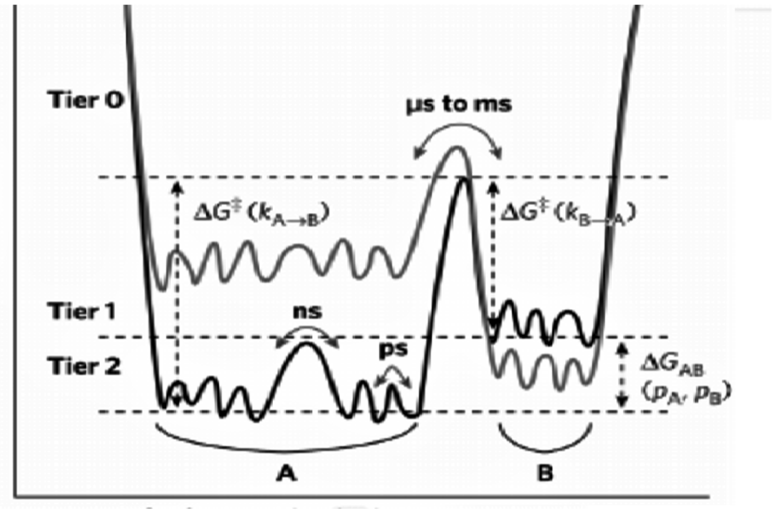


Figure 2.3: Subvalleys within valleys (Liu SQ et al., 2012)

In most interesting cases, such as the mean field models introduced below, as the temperature decreases the system transitions through a series of glassy phases where the energy landscape acquires a hierarchical structure with deep valleys nested within each pure state (Fischer and Hertz, 1991). These subvalleys are separated by high enough energy barriers that it traps the system on a certain timescale as to function as a *quasi-pure state.* If we assume a third-order hierarchy, then each quasi-pure state will have subvalleys nested within. And so on and so forth, we can define an n^{th} order pure state decomposition.

Geometrically, the density of the Gibbs measure in Hamming space displays clusters within clusters, as illustrated in the figure by (Berthier et al., 2019). Recall that the energy spectrum of a given energy level E is defined as: $\mathcal{N}_\epsilon(E) \equiv |\{\sigma \in \Sigma : E \leq \mathcal{H}(\sigma) \leq E + \epsilon\}|$, and that the ergodic system visits energy levels with frequencies proportional to their respective spectra, such that the time average is equal to the expectation w.r.t. μ_β.

In a two-level hierarchy, the state space breaks into pure states $\Sigma = \mathcal{S} \bigsqcup_{i\in[\eta]} \alpha_i$ and each pure states breaks into quasi-pure states $\alpha_i = \mathcal{Q}_i \bigsqcup_{k\in[\kappa^i]} \zeta_k^i$ where $\mathcal{S}$ and

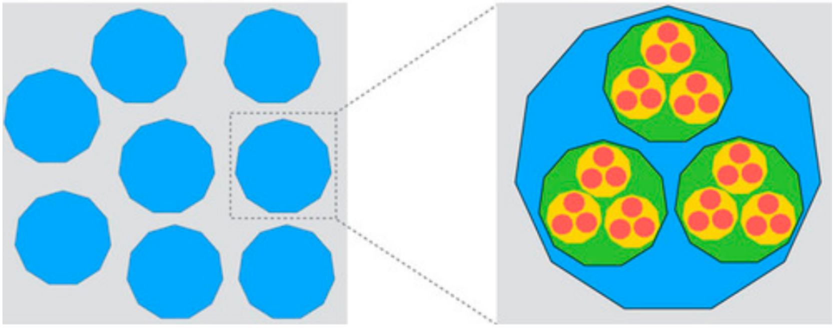


Figure 2.4: Hierarchical structure of the energy landscape (Berthier et al., 2019)

$\{\mathcal{Q}_{i\in[\eta]}\}$ are the spectra of the high energy barriers separating $\{\alpha_{i\in[\eta]}\}$ and $\{\zeta^i_{k\in[\kappa^i]}\}$ respectively with vanishing probability mass in the low temperature limit.

We can define the shortest timescale as the one during which the system roams in its original quasi-pure state ζ^i_k, visiting energy levels with frequencies proportional to their spectrum, that is to say ergodically, without having the time to climb the energy barriers surrounding ζ^i_k, in which case $\Delta^{(a)}_q \neq 0$ since $q^{(a)}_{EA} = \langle\sigma_i\rangle^2_{\zeta^i_k}$.

On an intermediate timescale, the system has enough time to climb the energy barriers within α_i (whose spectrum is $\{\mathcal{Q}_{i\in[\eta]}\}$) and roam its original pure state ergodically (i.e. such that the ergodic hypothesis holds within α_i), without having the time to climb over $\{\mathcal{H}_{\beta^*}(s) : s \in \mathcal{S}\}$ to visit the other pure states $\{\alpha_{j\in[\eta]\setminus i}\}$. In this intermediate regime, ergodicty is less broken than on the smaller timescale: $|\Delta^{(b)}_q| \leq |\Delta^{(a)}_q|$ and the Edwards-Anderson parameter measures the squared magnetization within pure states $q^{(b)}_{EA} = \langle\sigma_i\rangle^2_{\alpha_i}$.

Finally within the longest timescale, the system is given enough time to climb over the highest energy barriers $\{\mathcal{H}_{\beta^*}(s) : s \in \mathcal{S}\}$ statistically enough times that the time average is equal to the expectation w.r.t. μ_β on the entire state space Σ, the system is then completely ergodic and $\Delta^{(c)}_q = 0$, $q^{(c)}_{EA} = \langle\sigma_i\rangle^2_\Sigma$.

Chapter 3

The overlap and the birth of replica theory

Chapter organization

We start by showing some results concerning the distribution of *the overlap parameter* (**3.1**), and move on to a brief discussion on two competing candidate theories for spin glasses (**3.1.1**), and some history on the one which is relevant to our discussion (**3.1.2**). Then, we introduce the main tool from which a number of prediction concerning the low temperature correlation structure of the Gibbs measure follow, called *the Replica trick* (**3.2**), starting with a brief introduction to the preliminary notion of *large deviations* (**3.2.1**). Afterwards, we illustrate the Replica trick on a simple model called *the Random Energy model* (**3.3**), discuss its failure and introduce an extended version of the trick called *the Replica Symmetry Breaking scheme* (**3.4**). Then, after introducing a very general model of spin systems, that encompasses a number of constraint satisfaction problems, under the name of the *p-spin model* and relating it to the simpler model above (**3.5**), we use the extended tool to characterize what is called the *1-Replica symmetry breaking scenario* (**3.5.1**), which describes some important properties of constraint satisfaction problems that we will explore further in the next two chapters.

3.1 Introduction

Although the picture painted in chapter 2 gives a qualitative description of the pure state decomposition, it lacks the right formalism needed to distinguish between the various hierarchical scenarios or give a sharper quantitative characterization of the pure states in terms of size, that is the distribution of their Gibbs weights $\{w_{\alpha_{i\in[\eta]}}\}$.

The central realization leading to the new order parameter, is that we can probe the existence of pure states and their relative weight, simply by drawing two independent configurations and looking at how similar they are or *their overlap*:

Definition 15. *The (configuratinal) overlap between two states $\sigma, \tau \in \Sigma$ is defined as:*

$$q_{\alpha,\tau} \equiv \frac{\sum_{i\in[N]} \sigma_i \tau_i}{N}. \tag{3.1}$$

Given two independent states $\sigma, \tau \overset{iid}{\sim} \mu_\beta$, their overlap measures how correlated they are, in the extreme we distinguish 3 cases (Castellani and Cavagna, 2005):

$$q_{\alpha,\tau} \equiv \frac{\sum_{i\in[N]} \sigma_i \tau_i}{N} = \begin{cases} -1 & \text{if } \sigma, \tau \text{ are anti-correlated,} \\ \ \ 0 & \text{if they are uncorrelated,} \\ \ \ 1 & \text{if they are completely correlated} \\ & \text{(or identical).} \end{cases} \tag{3.2}$$

Along the same lines, we can measure the similarity between pure state by introducing the *states overlap* (whose definition is adapted from (Castellani and Cavagna, 2005)).

Definition 16. *Supposing that α, γ are non overlapping subsets of Σ (that satisfy some lengthy technical conditions listed in the last chapter permitting us to call them pure states), the states overlap between the two is defined as:*

$$q_{\alpha,\gamma} \equiv \frac{\sum_{i\in[N]} \langle \sigma_i \rangle_\alpha \langle \sigma_i \rangle_\gamma}{N}, \tag{3.3}$$

where

$$\langle \mathcal{O} \rangle_\alpha \equiv \sum_{\underline{\sigma}\in\alpha} \mathcal{O}(\underline{\sigma})\ \mu_\beta^\alpha(\underline{\sigma}) \quad \textit{with} \quad \mu_\beta^\alpha(\underline{\sigma}) \equiv \frac{\mathbf{1}\{\underline{\sigma} \in \alpha\}\ e^{-\beta\mathcal{H}(\underline{\sigma})}}{\sum_{\underline{\sigma}\in\alpha} e^{-\beta\mathcal{H}(\underline{\sigma})}},\ \forall \mathcal{O} : \Sigma \mapsto \mathbf{R}. \tag{3.4}$$

To avoid confusion, we will refer to $\sum_{i\in[N]} \sigma_i \tau_i / N$ as the *configurational overlap* or simply *the overlap* to distinguish it from the other kind that we always refer to as *states*-overlap. It is useful to write the states overlap in terms of configurational overlap, and simply focus characterizing the latter.

Theorem 4. *The states overlap between α and γ, satisfies: $q_{\alpha,\gamma} = \langle\ q_{\sigma,\tau}\ \rangle_{\alpha,\gamma}$.*

Proof.

$$q_{\alpha,\gamma} = \frac{1}{N} \sum_{i\in[N]} \left(\frac{\sum_{\underline{\sigma}\in\alpha} \underline{\sigma}_i \exp\big[-\beta\mathcal{H}(\underline{\sigma})\big]}{\mathcal{Z}_\alpha} \frac{\sum_{\underline{\tau}\in\gamma} \underline{\tau}_i \exp\big[-\beta\mathcal{H}(\underline{\tau})\big]}{\mathcal{Z}_\gamma} \right) \tag{3.5}$$

$$= \frac{1}{\mathcal{Z}_\alpha \mathcal{Z}_\gamma} \sum_{\underline{\sigma}\in\alpha} \sum_{\underline{\tau}\in\gamma} \exp\big[-\beta\mathcal{H}(\underline{\sigma})\big] \exp\big[-\beta\mathcal{H}(\underline{\tau})\big] \left\{ \frac{\sum_{i\in[N]} \underline{\sigma}_i \underline{\tau}_i}{N} \right\} \tag{3.6}$$

$$\equiv \langle\ q_{\sigma,\tau}\ \rangle_{\alpha,\gamma}. \quad \blacksquare \tag{3.7}$$

We can make sense of the size that a given pure state occupies in the state space Σ, by thinking of the typical overlap between two configurations drawn independently from said pure state. Note that the Hamming distance between two configurations satisfies $d(\sigma, \tau) = \frac{N}{2}(1 - q_{\sigma,\tau})$, and therefore, if a given pure state is relatively large, then the typical Hamming distance between $\sigma, \tau \overset{iid}{\sim} \mu_\beta^\alpha(.)$ won't be bounded away from N (Mezard and Montanari, 2009).

Thus, the smaller a pure state is, the closer two configurations draw from it are, the larger their overlap and vice versa; the larger $q_{\alpha,\alpha}$ is, the smaller is α.

Note that the self-overlap is given by the first moment of the configurational overlap independently of whether there is pure state decomposition, and is equal to the site average of the squared magnetization:

Theorem 5. *The configurational overlap between two arbitrary configurations $\sigma, \tau \in \Sigma$ satisfies: $\langle q_{\sigma,\tau} \rangle_\Sigma = \frac{1}{N} \sum_{i \in [N]} m_i^2$.*

Proof.

$$\langle q_{\sigma,\tau} \rangle_\Sigma \equiv \frac{1}{N} \sum_{\sigma \in \Sigma} \sum_{\tau \in \Sigma} \sum_{i \in [N]} \sigma_i \tau_i \mu_\beta(\sigma) \mu_\beta(\tau) \tag{3.8}$$

$$= \frac{1}{N} \sum_{i \in [N]} \left(\sum_{\sigma \in \Sigma} \sigma_i \mu_\beta(\sigma) \right) \left(\sum_{\tau \in \Sigma} \tau_i \mu_\beta(\tau) \right) = \frac{1}{N} \sum_{i \in [N]} m_i^2. \quad \blacksquare \tag{3.9}$$

The above two theorems were hinted at (although not proven) in the second chapter of (Fishcer and Hertz, 1991) where the authors expand on the physical significance of the different overlap parameters.

Since, the Edwards-Anderson order parameter has the implicit assumption that the local magnetization is site-independent (same across $i \in [N]$), the first moment of the overlap is also equal to the Edwards-Anderson order parameter:

$$\langle q_{\sigma,\tau} \rangle_\Sigma = \frac{\sum_{i \in [N]} m_i^2}{N} = q_{EA}. \tag{3.10}$$

To compute higher moments of the configurational overlap, it is useful to derive a more general result:

Theorem 6. *Let $\gamma_i \equiv \sigma_i \tau_i$, the r^{th} moment of the overlap is then given by:*

$$\langle q_{\sigma,\tau}^r \rangle = \sum_{k_1 \ldots k_r \in [N]^r} \frac{\langle\, \sigma_{k_1} \ldots \sigma_{k_r} \,\rangle^2}{N^r}, \tag{3.11}$$

where the sum $\sum_{k_1 \ldots k_r \in [N]^r}$ runs over all possible $\binom{N}{r}$ combinations of r spin indices from $\{1 \ldots N\}$ (without replacement).

Proof.

$$\langle q^r_{\sigma,\tau}\rangle = \frac{1}{N^r}\sum_{\sigma,\tau\in\Sigma^2}(\sigma_1\tau_1+\ldots\sigma_N\tau_N)\overset{r\ times}{\ldots}(\sigma_1\tau_1+\ldots\sigma_N\tau_N)\,\mu_\beta(\sigma)\mu_\beta(\tau) \tag{3.12}$$

$$= \frac{1}{N^r}\sum_{\sigma,\tau\in\Sigma^2}\left(\sum_{k_1\ldots k_r\in[N]^r}\prod_{l=1}^{r}\gamma_{k_l}\right)\mu_\beta(\sigma)\mu_\beta(\tau) \tag{3.13}$$

$$= \frac{1}{N^r}\sum_{k_1\ldots k_r\in[N]^r}\underbrace{\left(\sum_{\sigma\in\Sigma}\sigma_{k_1}\ldots\sigma_{k_r}\ \mu_\beta(\sigma)\right)}_{=\langle\ \sigma_{k_1}\ldots\sigma_{k_r}\ \rangle}\left(\underbrace{\sum_{\tau\in\Sigma}\tau_{k_1}\ldots\tau_{k_r}\ \mu_\beta(\tau)}_{same}\right) \tag{3.14}$$

$$= \sum_{k_1\ldots k_r\in[N]^r}\frac{\langle\ \sigma_{k_1}\ldots\sigma_{k_r}\ \rangle^2}{N^r}.\quad\blacksquare \tag{3.15}$$

This result was pointed out in chapter 12 of (Mezard and Montanari, 2009) although its proof was left as an exercise to the reader.

It follows then, that the variance of the overlap is equal to the rescaled spin glass susceptibility

$$Var(q_{\alpha,\tau}) = \langle q^2_{\sigma,\tau}\rangle - \langle q_{\sigma,\tau}\rangle^2 \tag{3.16}$$

$$= \frac{1}{N}\,\frac{\left(\sum_{k_1,k_2\in[N]^2}\langle\ \sigma_{k_1}\sigma_{k_2}\ \rangle^2\right) - \left(\sum_{k_1\in[N]}\langle\ \sigma_{k_1}\ \rangle^2\right)}{N} \tag{3.17}$$

$$\equiv \frac{\chi_{SG}}{N}, \tag{3.18}$$

noting that χ_{SG} is expected to diverge at the critical temperature of the glass transition (Mezard and Montanari, 2009).

3.1.1 Intractability of the Edwards-Anderson model and the replica symmetric solution

Still after more than four decades of the original paper by (Edwards and Anderson, 1975), the question of the existence of a phase transition at finite temperature is yet to be settled. The nearest neighbors model being analytically intractable, the only information available comes from principled numerical simulations.

After four decades, the consensus among physicists is that the two-dimensional EA model (at zero magnetic field) shows no phase transition at finite temperature but does in three dimensions and above.

Facing the intractability of the model on the grid, there are two competing theories;

a. The finite dimensional *droplet picture*, in which we analyse the grid model as it is, by assuming that the low temperature phase is governed by the excitations of finite blocks of the system, whose spins are reversed with respect to a given ground state. These excitations would thermodynamically satisfy: $\Delta\mathcal{H} \propto l^\theta$, where l is the block size and θ is a critical exponent and $\Delta\mathcal{H}$ is the induced change in energy. From there follows a methodology on how to go about computing the free energy and derive a physically sound picture of the spin glass phase, see chapter 4 of (Binder et al., 2008) for more details.

b. The second one, which has taken an enormous scope, way beyond spin glasses or even physics, is *Mean Field theory* (Mezard et al., 1987), where we forget about the geometry of the spins, and consider the system as a set of spins taking ± 1 values. These models are called *infinite range*, since every spin is assumed to interact with all of the others, and can be interpreted as the infinite dimensional limit of the grid model, where as $d \to \infty$, every point on the grid has an infinite number of nearest neighbors.

Hence, in the mean field approach, we relax the nearest neighbors condition, effectively enlarging $\mathcal{J}$ to encompass $\{J_{ij} : \forall \ (i,j) \in [N]^2\}$, we then proceed, as in the grid model, to identify the right order and figure out a way to solve it, then compute the free energy, and finally study the fluctuation around the mean field prediction.

This is where things get muddy. The validity of the mean field solution depends on a crucial assumption, namely that the critical dimension, above which fluctuations become negligible, needs to be finite (Fischer and Hertz, 1991). And while the veracity of this assumption in the case of the Edwards-Anderson model is a long debated question, the intricate hierarchy in the support of the Gibbs measure at low temperature in these mean field models, and its apparent universality in a wide range of disordered systems, is nothing short of extraordinary.

But what was even more remarkable than the accuracy of the non-rigorous predictions, that are now rigorously vindicated by the work of Guerra, Talagrand and then Panchenko, among others, was the method used to derive it, namely; *The Replica Symmetry Breaking scheme* (RSB) of Parisi (Mezard et al., 1987).

This general method, whose original intent was to derive an explicit expression of the free energy $F(\beta) = -\log \mathcal{Z}/\beta$, is a very powerful method that has been successfully used to predict phase transitions in a range of classical computational problems from the travelling salesman problem, to vertex covering, to the quintessential NP hard problem, namely *k-SAT*, that we will explore in depth in the next chapter. A very nice book that explores that surveys the use of the replica method in classical NP hard problems is (Hartmann and Weigt, 2005).

And while the replica method is not entirely rigorous, and quite strange as it involves the manipulation of matrices of half a row/column, the details of the computations are entirely specified, and is an essentially automatic analytical tool. However, the computations involved are quite cumbersome, and we have found that most references tend to skip key steps in deriving the end results, we have therefore attempted to derive the missing steps from (Mezard and Montanari, 2009), and included two important prerequisite notions, namely: large deviations and the saddle-point approximation, to make the exposition as self-contained as possible.

3.1.2 Some context and a summary for this chapter

A mean field variant of the Edwards-Anderson model was proposed by the authors in the seminal paper (Sherrington and Kirkpatrick, 1975) under the name of *the Sherrington-Kirkpatrick* (SK) model:

Definition 17 (Sherrington and Kirkpatrick, 1975)**.** *The Sherrington-Kirkpatrick model or SK model is an $N-$particle spin system with the Hamiltonian $\mathcal{H}(\sigma) = -\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j - B\sum_{i\in[N]}\sigma_i$.*

After proposing the model, the authors proceeded to solve it using the replica trick, that we describe in detail below.

The solution given by the replica method, however, turns out to display negative entropy at low temperature, and is thus physically nonsensical. Following this paper, numerous attempts were made to amend the low temperature behaviour by proposing variations of the static Edwards-Anderson order parameter, but have proven unsuccessful, until the celebrated Replica Symmetry Breaking solution by Parisi (Mezard et al., 1987).

In the next section, we will briefly go into some key steps in the replica trick without going into details, then introduce the technical machinery involved in the full computation, two key pieces are *the saddle-point approximation* and *the large deviation approach.* Afterwards, to illustrate the replica trick in a simple example, we will carry it on a toy model called *the random energy model*, and demonstrate the unphysical nature of the predicted low temperature free energy density.

Finally, we will present *Parisi's replica symmetry breaking scheme* as a solution to the low temperature erroneous prediction of the simple replica trick, and demonstrate it on a more complex model which generalizes the Sherrington-Kirkpatrick one, called the the p-spin model whose Hamiltonian is given by:

$$\mathcal{H}(\sigma) = -\sum_{1\leq i_1<\cdots<i_p\leq N} \underline{J}_{i_1\ldots i_p}\underline{\sigma}^a_{i_1}\cdots\underline{\sigma}^a_{i_p}, \tag{3.19}$$

and whose generality allows it to describe a large class of random constraint satisfaction problems that fall into the same universality class; the discontinuous 1RSB universality class.

The central aim of this thesis is to describe the physical meaning, as well as the algorithmic implications, of *pure state decomposition* in the $p-$spin model, the natural way to do this is through the overlap parameter. The first chapter introduces the necessary physical jargon, and the second chapter gives an informal qualitative description of (multiple) pure state decompositions in terms of overlap. This chapter revisits these same themes in the context of the replica method, which gives the right tools to characterize the distribution of the overlap and hence a clearer description of pure state decomposition.

3.2 The Replica trick

Recall that the entropy, internal energy, susceptibility and all other potentials of interest can be derived straightforwardly from the disorder-averaged free energy $\overline{F_{N,\beta}} \equiv -\frac{1}{N\beta}\overline{\log \mathcal{Z}}$.

However, integrating a logarithm over the disorder distribution is especially difficult and thus we have to resort to some kind of trick. The replica method, simply stated, consists of using a Taylor expansion to rewrite $\log \mathcal{Z}$ into a form that is easier to integrate over the disorder distribution. Before establishing, the replica identity, we have the following,

$$\lim_{n\to 0} \frac{x^n - 1}{n} = \lim_{n\to 0} \frac{e^{\log x^n} - 1}{n} \tag{3.20}$$

$$= \lim_{n\to 0} \frac{n \log x + 1/2!(n \log x)^2 + \dots}{n} \tag{3.21}$$

$$= \log x. \tag{3.22}$$

Now, since $\log(1 + nx) \approx nx$, when $n \approx 0$, we have for $x = \overline{\log \mathcal{Z}}$,

$$\overline{\log \mathcal{Z}} = \lim_{n\to 0} \frac{\log\left(1 + n\overline{\log \mathcal{Z}}\right)}{n} \quad \text{and since} \quad \overline{\log \mathcal{Z}} \xrightarrow[n\to 0]{} \frac{\overline{\mathcal{Z}^n} - 1}{n}, \tag{3.23}$$

$$= \lim_{n\to 0} \frac{\log\left(1 + n\frac{\overline{\mathcal{Z}^n} - 1}{n}\right)}{n} \tag{3.24}$$

$$= \lim_{n\to 0} \frac{\log\left(\overline{\mathcal{Z}^n}\right)}{n}. \tag{3.25}$$

The problem of deducing the free energy of an N particle system is then reduced to studying the n^{th} power of its partition function and then averaging over the disorder.

In the replica method, the central object of interest is $\overline{\mathcal{Z}^n}$, which can be interpreted as the partition function of a larger $n \times N$ particle system , often called *the replicated system*, a system of n copies or *replicas* $\{\sigma^a : a \in [n]\} \overset{iid}{\sim} \mu_{\beta,\underline{\mathcal{J}}}$, where the word *replica* is meant in the sense of having the same disorder $\underline{\mathcal{J}}$.

3.2.1 A preliminary: Large deviations

We start by introducing the saddle-point approximation, that we adapt from the wikipedia entry.

Theorem 7. *Consider an arbitrary function $f : \mathcal{X} \mapsto \mathbf{R}$ that is bounded and analytic around its maximum $x^* \equiv argmin_{x\in\mathcal{X}} f(x)$. The saddle-point approximation is an asymptotic approximation of the integral $\mathcal{I}_N \equiv \int_{\mathcal{X}} e^{Nf(x)}dx$ which satisfies*

$$\mathcal{I}_N = \sqrt{\frac{2\pi}{N|f''(x^*)|}}\, e^{Nf(x^*)} \quad \textit{for} \quad N \gg 1 \tag{3.26}$$

Proof. Since $f(x)$ is analytic around its maximum $x^* \equiv argmin_{x\in\mathcal{X}} f(x)$, we have the following Taylor approximation around x^*:

$$f(x) \approx f(x^*) + f^{'}(x^*)(x - x^*) + \frac{f^{''}(x^*)(x - x^*)^2}{2} + \frac{f^{'''}(x^*)(x - x^*)^3}{6} \tag{3.27}$$

Moreover, since x^* is a maximum of f, the second term is zero and $f^{''}(x^*) = -|f^{''}(x^*)|$. Now, consider the change of variables $y \equiv (x - x^*)\sqrt{N}$, the above integral then satisfies

$$\mathcal{I}_N = e^{Nf(x^*)} \int_{\mathcal{Y}} \exp\left\{ \frac{y^2}{2}|f^{''}(x^*)| + \frac{y^3}{6\sqrt{N}} f^{'''}(x^*) \right\} dy \tag{3.28}$$

For $N \gg 1$, we have $\frac{y^3}{6\sqrt{N}} f^{'''}(x^*) \approx 0$, and therefore, using the well known Gaussian integral $\int e^{-a(z+b)^2} dz = \sqrt{\frac{\pi}{a}}$, we have the following *saddle-point approximation*:

$$\mathcal{I}_N = \sqrt{\frac{2\pi}{N|f^{''}(x^*)|}}\, e^{Nf(x^*)} \quad \text{for} \quad N \gg 1. \quad \blacksquare \tag{3.29}$$

Definition 18 (Mezard and Montanari, 2009)**.** *We say that A_N is up to leading exponential order equal to B_N, when $\underset{N \longrightarrow \infty}{lim} \frac{1}{N} log\left(\frac{A_N}{B_N}\right) = 0$ and denote it: $A_N \doteq B_N$.*

Definition 19 (Mezard and Montanari, 2009)**.** *Suppose that the probability of a random variable $\mathcal{O}$ is given by $\mathbf{P}[.]$, we say that $\mathcal{O}$ satisfies a large deviation principle with a rate function $\mathcal{I} : \mathcal{X} \mapsto \mathbf{R}^+$, if*

$$\mathbf{P}[\mathcal{O} = \overline{\mathcal{O}}] \doteq e^{-N\mathcal{I}(\overline{\mathcal{O}})}. \tag{3.30}$$

Now assuming that $\mathcal{O}$ is a random variable that does follow a large deviation principle, the next step is to derive the *rate function $\mathcal{I}$*. To this end, we introduce *the logarithmic cumulant generating function* $\Psi : \mathbf{R} \mapsto \mathbf{R}$:

$$\Psi_N(t) \equiv \frac{1}{N} \log\left(\mathbf{E}\left[e^{Nt\overline{\mathcal{O}}} \right] \right). \tag{3.31}$$

We then have the following result from (Mezard and Montanari, 2009) of which we complete the proof:

Theorem 8. *Suppose $\mathcal{O}$ is a random variable that follows a large deviation principle with the rate function $\mathcal{I}$, then the large N limit of the logarithmic moment generating function of $\mathcal{O}$, $\Psi_N(t) \equiv \log\left(\mathbf{E}\left[e^{Nt\overline{\mathcal{O}}}\right]\right)/N$ is given by the Legendre transform of the rate function:*

$$\underset{N \longrightarrow \infty}{lim} \Psi_N(t) = \underset{\mathcal{O} \in \mathbf{R}}{sup} \left\{ t\mathcal{O} - \mathcal{I}(\mathcal{O}) \right\}. \tag{3.32}$$

Proof. To make use of the saddle-point approximation, relying on the large deviation assumption, we rewrite the expectation as an integral over $e^{Ng(.)}$:

$$\underset{N \longrightarrow \infty}{lim} \Psi_N(t) = \underset{N \longrightarrow \infty}{lim} \frac{1}{N} \log\left(\int \exp\left\{ N(t\overline{\mathcal{O}} - \mathcal{I}(\overline{\mathcal{O}})) \right\} d\overline{\mathcal{O}} \right) \tag{3.33}$$

Let $g : \mathbf{R} \times \mathbf{R} \mapsto \mathbf{R}$ be

$$g(t, \mathcal{O}) \equiv t\mathcal{O} - \mathcal{I}(\mathcal{O}). \tag{3.34}$$

The limiting moment generating function then becomes the logarithm of a simple Gaussian integral

$$\underset{N \longrightarrow \infty}{lim} \Psi_N(t) = \underset{N \longrightarrow \infty}{lim} \frac{1}{N} \log \left(\int e^{Ng(t, \overline{\mathcal{O}})} d\overline{\mathcal{O}} \right) \tag{3.35}$$

Assuming that g is analytic in $\mathcal{O}$ around its maximum $\mathcal{O}^* \equiv \underset{\mathcal{O} \in \mathbf{R}}{argmax}\ g(t, \mathcal{O})$, we have the saddle-point approximation

$$\underset{N \longrightarrow \infty}{lim} \Psi_N(t) = \underset{N \longrightarrow \infty}{lim} \frac{c}{N} log \left[e^{Ng(t, \mathcal{O}^*)} \right] \quad \text{for some constant } c, \tag{3.36}$$

which leads us to *the Legendre transform* of the rate function $\mathcal{I}$:

$$\underset{N \longrightarrow \infty}{lim} \Psi_N(t) = \underset{\mathcal{O} \in \mathbf{R}}{sup} \Big\{ t\mathcal{O} - \mathcal{I}(\mathcal{O}) \Big\}. \quad \blacksquare \tag{3.37}$$

The use the aforementioned techniques come up in the last steps of the replica trick. Before we go into details of the computation as carried in (Mezard and Montanari, 2009), we give a brief summary of the key steps:

1). We start by expanding the product of replicas (identical partitions) in $\mathscr{Z}^n$ to rewrite as a sum over 2^{Nn} elements:

$$\mathscr{Z}^n = \prod_{k=1}^{n} \Big(\sum_{i_k \in [2]^N} e^{-\beta E_{i_k}} \Big) = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \prod_{k=1}^{n} e^{-\beta E_{i_k}}.$$

2). We use the *Gaussian identity* $\mathbf{E}e^{\lambda X} = e^{\Delta \lambda^2/2}$ for zero mean Gaussians with Δ^2 variance, to compute the average over the randomness of the energy to get an explicit formulation of $\overline{\mathscr{Z}^n}$ as a function of an overlap matrix $\mathcal{Q}_{ab} \equiv \mathbf{1}\{i_a = i_b\}$ between the n replicas $(i_1, \dots, i_n)$, yielding:

$$\overline{\mathscr{Z}^n} = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \exp \Big\{ \frac{N\beta^2}{4} \sum_{a,b \in [n]^2} Q_{ab} \Big\}. \tag{3.38}$$

3). We then define *the spectrum of the overlap* $\mathcal{N}(\underline{Q})$ to be the cardinality of the set of n replicas: $\{i_a\}_n$ whose overlap is equal to a specific value: $\underline{Q}$:

$$\mathcal{N}(\underline{Q}) \equiv \Big| \Big\{ (i_1 \dots i_n) \in [2]^{Nn} : \{Q_{ab}(i_1 \dots i_n)\} = \underline{Q} \Big\} \Big|, \tag{3.39}$$

and make the change of variables inside the summands from $\sum_{(i_1,\dots i_n)}$ to $\sum_{\underline{Q}} \mathcal{N}(\underline{Q})$. Then, we assume that the overlap follow a large deviation principle such that $\mathcal{N}(\underline{Q}) \doteq e^{Ns(\underline{Q})}$, to get:

$$\overline{\mathscr{Z}^n} \doteq \sum_{\underline{Q}} e^{Ng(\underline{Q})}.$$

4). Finally, now that we have $\overline{\mathcal{Z}^n}$ written in saddle-point friendly form, we compute the saddle-point of $g(Q)$, and take the limit when $n \to 0$ to verify the accuracy of the replica method in predicting the free energy density $f(\beta) \equiv \lim_{N\to\infty} -\overline{\log(\mathcal{Z})}/N\beta$.

Note that the precise definition of large deviation principles is given by the Gartner-Ellis theorem, since we find that it is not directly relevant to understanding the replica method more deeply, we have omitted it, a very nice survey of its use in statistical mechanics is given in (Touchette, 2009).

3.3 The Random Energy Model

In (Derrida, 1980), the author introduced a toy model of spin glass, which simplifies the SK model by replacing the Hamiltonian with a set of Gaussian energy levels $E_j \overset{iid}{\sim} \mathcal{N}(0, N/2)$ for all $j \in |\Sigma|$, where E_j represents the energy level of the j^{th} state in $\Sigma \equiv \{-1,1\}^N$.

For this section, we follow the simpler notation of (Derrida, 1980; Mezard and Montanari, 2009), and denote an arbitrary configuration in the state space by $i_k \in \Sigma$ instead of the usual σ, such that $(i_k)_p$ refers to the value of the p^{th} spin in the $N-$dimensional binary vector of the configuration $i_k \in \{-1,1\}^N \equiv \Sigma$. Moreover, since we are not interested in the particular value of a given state i_k but just its index (to whom we associate the corresponding energy level E_{i_k}) which spans $|\Sigma| = 2^N$, by abuse of notation, we will replace the states in $\{-1,1\}^N$ by their indices to write $i_k \in \Sigma \equiv [2]^N$.

Definition 20. *The Random Energy Model (REM) is an $N-$particle system whose state $i_k \in \{-1,1\}^N \equiv \Sigma$ is distributed according to:*

$$\mu_\beta(i_k) \equiv \frac{e^{-\beta E_{i_k}}}{\sum_{i_l \in [2]^N} e^{-\beta E_{i_l}}}, \tag{3.40}$$

where $E_{i_k} \overset{iid}{\sim} \mathcal{N}(0, N/2)$ for all $i_k \in [2]^N$.

Theorem 9 (Mezard and Montanari, 2009)**.** *The expectation w.r.t. $\{E_{i_k}\}$ of the n^{th} power of the partition function of an $N-$particle REM satisfies:*

$$\overline{\mathcal{Z}^n} \equiv \mathbf{E}_{E_{i_k} \overset{iid}{\sim} \mathcal{N}(0,N/2)}[\mathcal{Z}^n] = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \exp\left\{ \frac{N\beta^2}{4} \sum_{a,b \in [n]^2} Q_{ab} \right\}, \tag{3.41}$$

where $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$ for all $a, b \in [n]$ are the entries the overlap matrix.

Proof. By developing the product it is easy to see that:

$$\mathcal{Z}^n = \prod_{k=1}^{n} \Big(\sum_{i_k \in [2]^N} e^{-\beta E_{i_k}} \Big) = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \prod_{k=1}^{n} e^{-\beta E_{i_k}}. \tag{3.42}$$

To get rid of the dependence on the indices inside the product, we can encode $\{i_k\}$

using identity functions such that for all $i_k \in [2]^{Nn}$ we have

$$e^{-\beta E_{i_k}} = \prod_{j=1}^{2^N} \exp\{-\beta E_j\} \, \mathbf{1}\{i_k = j\} \tag{3.43}$$

Hence,

$$\mathcal{Z}^n = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \prod_{j=1}^{2^N} \exp\Big\{ -\beta E_j \sum_{k=1}^{n} \mathbf{1}\{i_k = j\}\Big\} \tag{3.44}$$

Let $\lambda_j \equiv -\beta \sum_{k=1}^{n} \mathbf{1}\{i_k = j\}$, recalling that $\mathbf{E} e^{\lambda X} = e^{\Delta \lambda^2/2}$ for zero mean Gaussians with Δ^2 variance, we get

$$\mathbf{E} e^{\lambda E_j} = \exp\Big\{ \frac{N}{4} \Big(-\beta \sum_{k=1}^{n} \mathbf{1}\{i_k = j\} \Big)^2 \Big\} \tag{3.45}$$

Hence the disorder average of the replicated partition is

$$\begin{aligned} \overline{\mathcal{Z}^n} &= \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \prod_{j=1}^{2^N} \exp\Big\{ \frac{N\beta^2}{4} \Big(\sum_{k=1}^{n} \mathbf{1}\{i_k = j\} \Big)^2 \Big\} & (3.46) \\ &= \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \exp\Big\{ \frac{N\beta^2}{4} \sum_{j \in [2]^N} \Big(\sum_{k=1}^{n} \mathbf{1}\{i_k = j\} \Big)^2 \Big\} & (3.47) \\ &= \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \exp\Big\{ \frac{N\beta^2}{4} \sum_{j \in [2]^N} \sum_{a,b \in [n]^2} \mathbf{1}\{i_a = j\} \mathbf{1}\{i_b = j\} \Big\}. & (3.48) \end{aligned}$$

Since the indices $\{i_k\}$ and their assignments $j \in [2]^N$ are analogous to configuration in an Ising state space, i.e. $i_k \in \{\pm 1\}^N \; \forall k \in [n]$, we have

$$\sum_{j \in [2]^N} \mathbf{1}\{i_a = j\} \mathbf{1}\{i_b = j\} \Big\} = \mathbf{1}\{i_a = i_b\} \tag{3.49}$$

defines an *overlap matrix* $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\};$, that is symmetric $(Q_{ab} = Q_{ba})$ with unit elements in its diagonal $Q_{aa} = 1, \; \forall a \in [n]$. The averaged replicated partition can thus be written as a function of the overlap

$$\overline{\mathcal{Z}^n} = \sum_{(i_1 \dots i_k) \in [2]^{Nn}} \exp\Big\{ \frac{N\beta^2}{4} \sum_{a,b \in [n]^2} Q_{ab} \Big\}. \quad \blacksquare \tag{3.50}$$

Note that while i_k and i_l are probabilistically independent conditional on the sample realization $\{E_{i_k}\}$, once we average over the energy distribution, the replicas are no longer independent as illustrated by the overlap parameter. In fact, as observed in chapter 8 of (Mezard and Montanari, 2009), the exponent in the sum effectively defines a Hamiltonian $\mathcal{H}_{REM}(i_1 \dots i_k) \equiv \frac{N\beta}{4} \sum_{a,b \in [n]^2} Q_{ab}$ of a nN particle system who is no longer disordered (deterministic energy), whose ground state is given by

$$\min_{(i_1 \dots i_n) \in [2]^{Nn}} \mathcal{H}_{REM}(i_1 \dots i_k) = \mathcal{H}_{REM}(i_1^* \dots i_n^*) = \frac{-N\beta n^2}{4}, \tag{3.51}$$

where the ground states of system formed by the replicated system given by $\{i_a\}_{a\in[n]}$ are the all equal assignments $\{(i_1^* \dots i_n^*) : i_1^* = \dots = i_n^*\}$. Moreover, after averaging, the replicas as no longer independent as they were in the product measure in $\mathcal{Z}^n = \sum_{(i_1 \dots i_n)} \prod_{a\in[n]} e^{-\beta E_{i_k}}$, since the Hamiltonian has a unique minimum, as the temperature is lowered μ_β concentrates on the $(i_1^* = \dots = i_n^*)$ state where the replicas are locked together in the same configuration with high probability and are thus highly correlated.

The main strategy behind the replica method following (Mezard and Montanari, 2009) is to formulate $\overline{\mathcal{Z}^n}$ as a function of a single variable; the overlap, and to find a suitable saddle point $\{Q_{ab}^*\}$ approximation to derive an explicit expression. As we will see, the crux of the problem lies in this key step, where the energy landscape determines the suitability of the candidate saddle point.

In order to make the overlap matrix the only independent variable in $\overline{\mathcal{Z}^n}$, we need to make a change of variable where we replace the sum over the nN particle binary state space $\sum_{(i_1 \dots i_n)}$ by a sum over all possible values of the overlap with each term multiplied by its frequency or *spectrum* $\sum_{\underline{Q}} \mathcal{N}(\underline{Q})$:

Definition 21. *We define the spectrum of the overlap: $\mathcal{N}(\underline{Q})$ to be the cardinality of the set of n replicas: $\{i_a\}_n$ whose overlap is equal to a specific value $\underline{Q}$:*

$$\mathcal{N}(\underline{Q}) \equiv \left|\left\{(i_1 \dots i_n) \in [2]^{Nn} : \{Q_{ab}(i_1 \dots i_n)\} = \underline{Q}\right\}\right|. \tag{3.52}$$

Let the overlap matrix be such that its entries are given by: $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$ and let $f(\{Q_{ab}\}) \equiv \exp\left\{\frac{N\beta^2}{4} \sum_{a,b\in[n]^2} Q_{ab}\right\}$, we make the change of variable:

$$\sum_{(i_1 \dots i_n)\in[2]^{Nn}} f(\{Q_{ab}(i_1 \dots i_n)\}) = \sum_{\underline{Q}\in\mathcal{B}_{n\times n}} \mathcal{N}(\underline{Q})\, f(\{\underline{Q}_{ab}\}), \tag{3.53}$$

where $\mathcal{B}_{n\times n}$ is the set of symmetric $\{0,1\}$ matrices with unit diagonal.

Notice that in the right hand side, the sum goes over $|\mathcal{B}_{n\times n}| = 2^{n(n-1)/2}$ matrices while the one on the left sums over 2^{nN} elements, we can thus expect $\mathcal{N}(\underline{Q})$ to follow a large deviation principle, where most of the probability mass of the distribution of the overlap values is carried by a very small ($\ll 2^{nN}$) subset of configurations $(i_1 \dots i_n)$:

$$\mathcal{N}(\underline{Q}) \doteq \exp\left\{Ns(\underline{Q})\right\}. \tag{3.54}$$

The replicated partition is then up to leading exponential order equal to

$$\overline{\mathcal{Z}^n} \doteq \sum_{\underline{Q}} \exp\left\{Ng(\underline{Q})\right\} \quad \text{where} \quad g(\underline{Q}) \equiv \frac{\beta^2}{4} \sum_{a,b\in[n]^2} \underline{Q}_{ab} + s(\underline{Q}). \tag{3.55}$$

Now, consider the permutation group of n elements: S_n , for any permutation $\pi \in S_n$, π takes as input an ordered set of replica indices and permutes their order, π :

$(i_1 \dots i_n) \mapsto (\pi(i_1)\dots\pi(i_n))$. The central observation underlying the replica method is that the function $g : \mathcal{B}_{n\times n} \mapsto \mathbf{R}$ is invariant under permutation of replicas. We make this observation formal and prove in the following lemma that was left implicit in (Mezard and Montanari, 2009).

Lemma 2. *Let $\mathcal{B}_{n\times n}$ denote the set of $n \times n$ symmetric matrices with binary entries and unit diagonal, and suppose that the overlap $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$ (where $i_a, i_b \overset{iid}{\sim} \mu_\beta$ with fixed $\{E_{i_k}\}$) follows a large deviation principle with rate function $s(.)$. Then, the function $g : \mathcal{B}_{n\times n} \mapsto \mathbf{R}$ given by $g(\underline{Q}) \equiv \frac{\beta^2}{4}\sum_{a,b\in[n]^2} \underline{Q}_{ab} + s(\underline{Q})$ is invariant under any permutation $\pi \in S_n$ where S_n is the permutation group of n elements: $g \circ \pi(Q) = g(Q)$, $\forall \pi \in S_n$ for all $Q \in \mathcal{B}_{n\times n}$.*

Proof. Since the replicas were assigned indices ($a \in [n]$) arbitrarily from the very beginning when we rewrote the product (that is invariant under S_n) as a sum: $\mathcal{Z}^n = \prod_{k=1}^n \left(\sum_{i_k\in[2]^N} e^{-\beta E_{i_k}}\right) = \sum_{(i_1\dots i_k)\in[2]^{Nn}} \prod_{k=1}^n e^{-\beta E_{i_k}}$, the summands $(i_1 \dots i_n)$ are invariant under any permutation $\pi \in S_n$:

$$\overline{\mathcal{Z}^n} = \sum_{(i_1\dots i_k)} \exp\left\{\frac{N\beta^2}{4}\sum_{a,b\in[n]^2} \mathbf{1}\{i_a = i_b\}\right\} = \sum_{(\pi(i_1)\dots\pi(i_k))} \exp\left\{\frac{N\beta^2}{4}\sum_{a,b\in[n]^2} \mathbf{1}\{\pi(i_a) = \pi(i_b)\}\right\}. \tag{3.56}$$

That is to say, the value of $\overline{\mathcal{Z}^n}$ is unchanged by any permutation of $(i_1 \dots i_n)$, and since $\overline{\mathcal{Z}^n}$ depends on the replica indices ($a \in [n]$) only through $g(Q_{ab})$ (where $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$), $g : \mathcal{B}_{n\times n} \mapsto \mathbf{R}$ is also invariant under replica permutation. ■

For a fixed permutation $\pi \in S_n$, let Q^π denote the permuted matrix whose elements are $Q_{\pi(a)\pi(b)}$ for all $1 \le a < b \le n$,. Since the overlap matrix is defined as $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$, the permuted matrix Q^π is obtained simply by permuting pairs of rows and columns of Q simultaneously.

To arrive at an explicit expression for $\overline{\mathcal{Z}^n}$, our strategy is to approximate the sum though a single dominant term, namely its saddle-point: $Q^{sp} = argmax_{\underline{Q}\in\mathcal{B}_{n\times n}}\, g(\underline{Q})$. Because of the asymmetry in the number of summands between $\sum_{(i_1\dots i_n)}$ and $\sum_{\underline{Q}}$, where $|\mathcal{B}_{n\times n}| = 2^{n(n-1/2)} \ll 2^{nN} = |(i_1 \dots i_n)|$, we are lead to assume that the spectrum of the overlap $\mathcal{N}$ is highly concentrated on a relatively very small subset of $\mathcal{B}_{n\times n}$, with exponential decay away from $argmax_Q s(Q)$.

One could argue that using a saddle-point for a discrete sum is inappropriate, hence the importance of the order of the limits of the replica method where we take the $n \longrightarrow 0$ is only taken after taking the thermodynamic limit $N \longrightarrow \infty$:

$$\overline{log(\mathcal{Z})} = \lim_{n\longrightarrow 0}\lim_{N\longrightarrow\infty} \frac{1}{n} log(\overline{\mathcal{Z}^n}). \tag{3.57}$$

With this caveat, as $N \longrightarrow \infty$, the overlap's limiting support is in the continuum $Q_{ab} \in [-1, 1]$, $\forall\, 1 \le a < b \le n$, such that:

$$\overline{\mathcal{Z}^n} \doteq \int_{[-1,1]^{n(n-1)/2}} e^{Ng(Q)} \prod_{a<b} dQ_{ab} \tag{3.58}$$

However, since we don't have an explicit expression for the rate function, we are left to guess which overlap matrix has the largest spectrum and work back from there.

The insight leading to the replica symmetric solution can be summarized in a simple argument: Considering g is replica symmetric, we can try to guess that $s(Q^{sp})$ is also invariant under any $\pi \in S_n$, that is to say, that $Q^{sp,\,\pi} = Q^{sp}$, $\forall \pi \in S_n$, which restricts the pool of saddle-point candidates from $\mathcal{B}_{n\times n}$ to a tiny subset $\mathcal{A} \equiv \{Q \in \mathcal{B}_{n\times n} : Q^{\pi} = Q,\, \forall \pi \in S_n\}$:

$$\overline{\mathcal{Z}^n} \doteq \exp\Big\{N \max_{Q\in\mathcal{A}} g(Q)\Big\}. \tag{3.59}$$

If the saddle-point matrix is invariant under any simultaneous permutation of rows and their corresponding columns, then $Q_{ab} = \underline{q}$, $\forall a \neq b$. Since $Q \in \mathcal{B}_{n\times n}$, this leaves us with only two options:

- $Q_{ab} = 1$, for all $a \neq b$ and $Q_{aa} = 1$ for all $a \in [n]$.
- $Q_{ab} = 0$, for all $a \neq b$ and $Q_{aa} = 1$ for all $a \in [n]$.

We will denote the first item by Q_1^{sp} and the second by Q_0^{sp}. The above saddle-point then becomes:

$$\overline{\mathcal{Z}^n} \doteq \exp\Big\{N \max\big[g(Q_0^{sp}),\; g(Q_1^{sp})\big]\Big\}. \tag{3.60}$$

As a reassurance for our initial guess, the following result proven in (Mezard and Montanari, 2009) shows that both saddle-point matrices have a large spectrum $\mathcal{N}(Q)$, that is at least exponential in the size of individual system.

Lemma 3 (Mezard and Montanari, 2009)**.** *Suppose that the overlap $Q_{ab} \equiv \mathbf{1}\{i_a = i_b\}$ follows a large deviation principle with rate function $s(.)$, and let Q_0^{sp}, Q_1^{sp} be as defined above and $g(\underline{Q}) \equiv \frac{\beta^2}{4}\sum_{a,b\in[n]^2} \underline{Q}_{ab} + s(\underline{Q})$, we then have:*

$$g(Q_0^{sp}) = \frac{n^2\beta}{4} + \log 2, \quad \textit{and} \quad g(Q_1^{sp}) = \frac{n^2\beta}{4} + n\log 2. \tag{3.61}$$

Proof. For replicated systems with zero overlap, each individual $i_{k\in[n]}$ need to differ by at least one spin from the others. To count the number of such replicated systems, we can choose i_1 from 2^N possible states, for i_2 , excluding the configuration i_1 , we have $2^N - 1$ remaining choices and so on, we arrive at:

$$\mathcal{N}(Q_0^{sp}) = 2^N(2^N - 1)\ldots(2^N - (n-1)). \tag{3.62}$$

Recalling the large deviation assumption on the overlap spectrum, we have

$$\mathcal{N}(Q) \doteq e^{Ns(Q)} \iff \lim_{N\longrightarrow\infty} \frac{1}{N}\log\left[\frac{\mathcal{N}(Q_0^{sp})}{e^{Ns(Q_0^{sp})}}\right] = 0, \tag{3.63}$$

hence

$$\lim_{N\longrightarrow\infty} \frac{1}{N}\left[\log\Big(2^{nN} + \mathcal{O}(1)\Big) - Ns(Q_0^{sp})\right] = 0 \tag{3.64}$$

such that

$$\lim_{N\longrightarrow\infty} \frac{nN\log 2 - Ns(Q_0^{sp})}{N} = n\log 2 - s(Q_0^{sp}) = 0 \tag{3.65}$$

and therefore

$$s(Q_0^{sp}) = n\log 2. \tag{3.66}$$

And since $Q_{0,\,ab}^{sp} = 0$ forall $a \neq b$, $\sum_{ab} Q_{0,\,ab}^{sp} = n$, hence

$$g(Q_0^{sp}) = \frac{n^2\beta}{4} + \log 2. \tag{3.67}$$

As for the Q_1^{sp} case, we choose one configuration from 2^N options, identical among all individual systems, such that

$$\lim_{N\longrightarrow\infty} \frac{1}{N}\log\left[\frac{\mathcal{N}(Q_1^{sp})}{e^{Ns(Q_1^{sp})}}\right] = \lim_{N\longrightarrow\infty} \frac{1}{N}\log\left[\frac{2^N}{e^{Ns(Q_1^{sp})}}\right] = \lim_{N\longrightarrow\infty} \frac{N\log 2 - Ns(Q_1^{sp})}{N} = 0 \tag{3.68}$$

hence, $s(Q_1^{sp}) = \log 2$. And since Q_1^{sp} has all unit elements, $\sum_{ab} Q_{1,\,ab}^{sp} = n^2$, which culminates in

$$g(Q_1^{sp}) = \frac{n^2\beta}{4} + n\log 2. \quad \blacksquare \tag{3.69}$$

By abuse of notation, we have thus far omitted the temperature dependence in $\overline{\mathcal{Z}^n}$, when the original intent of the whole exercise was to show the existence of a non-trivial phase transition at a certain critical temperature. Now that we have laid out explicit formulas for computing the free energy, we can address this central question.

As previously discussed, phase transitions can be recognized by dramatic changes in the (thermodynamic) free energy. Let Q^* denote the correct saddle-point, since $\overline{\mathcal{Z}^n}$ is up leading exponential order exponential in $g(Q^*)$, we can safely assume that the phase transition happens at the critical point of $g : \beta, n \mapsto \mathbf{R}$ with fixed overlap.

To get rid of the overlap dependence and further highlight the dependence on the temperature parameter, we define

$$g_0(\beta, n) \equiv \frac{n\beta^2}{4} + n\log 2 \quad \text{and} \quad g_1(\beta, n) \equiv \frac{n^2\beta^2}{4} + \log 2. \tag{3.70}$$

Since we don't know a priori which of the two quantities is the correct one, we choose rather arbitrarily $\beta_c(n) \equiv 2\sqrt{\log 2}/n$ as a point of comparison between g_0 and g_1 in different scenarios. By this, we mean to delineate two distinct cases; the integer one $n \geq 1,$ and its analytic continuation $n < 0$, which, as we shall see, display quite different features.

In chapter 2 of (Mezard and Montanari, 2009), the authors introduce a result which stipulates that if there exists a function $s : \mathbf{R} \mapsto \mathbf{R}^+$ such that the spectrum of a given energy level $\mathcal{N}_\Delta(E) \equiv \left|\{i_k : E \leq E_{i_k} \leq E{+}\Delta\}\right|,$ satisfies $\mathcal{N}_\Delta(E) \doteq \exp\Big[Ns(E/N)\Big]$, then the free energy density is given by

$$-\beta f(\beta) = \max_e \Big[s(e) - \beta e\Big]. \tag{3.71}$$

A short computation shows that in the case of the random energy model, the free energy density $f(\beta)$ $(\equiv \lim_{N\longrightarrow\infty} -\overline{\log \mathcal{Z}}/N\beta)$ satisfies

$$f(\beta) = -\frac{\beta}{4} - \frac{\log 2}{\beta} \quad \text{if } \beta \leq \beta_c(1) \quad \text{and} \quad f(\beta) = -\sqrt{\log 2} \quad \text{otherwise.} \tag{3.72}$$

We start with a qualitative description of the physical picture painted by the replica symmetric solution starting with integer $n \geq 1$ above and below $\beta_c(n)$, and then consider the trickier case of the analytic continuation $n \longrightarrow 0$. In each case, we begin by sketching out the physical meaning of having one saddle-point dominate rather than the other. If the predicted picture is physically sound, we take the analytic continuation to zero and derive the free energy density to check if the solution is exact, i.e. agrees with the equation just above.

In the integer $n \geq 1$ case, it is easy to see that $g_1(\beta, n) \leq g_0(\beta, n)$ for all $\beta \leq \beta_c(n)$ and $g_1(\beta, n) > g_0(\beta, n)$ otherwise. Hence, the correct saddle-point shifts from being Q_0^{sp} in the high temperature regime to Q_1^{sp} below the critical temperature.

This picture agrees with the physical expectations outlined in previous sections. Indeed, viewing the collection of replicas of the same $N-$particle system as a single larger $nN-$particle system with no disorder and an effective Hamiltonian given by $\mathcal{H}_{REM}(i_1 \dots i_k) \equiv \frac{N\beta}{4} \sum_{a,b\in[n]^2} Q_{ab},$, for high enough temperatures (small β), we expect to see its particles point in independently random directions.

Hence, for $N \gg 1$, the nN Ising state space, which is exponential in nN , becomes so large that replicas $\{i_{a\in[n]}\}$ point in completely different directions with high probability and are therefore expected to have near zero overlap. The replica symmetric solution then predicts the correct saddle-point Q_0^{sp} in the high temperature phase for all integer $n \geq 1$.

On the other hand, below the critical temperature μ_β concentrates on the minimum energy configuration of the $\mathcal{H}_{REM}$ which has replicas locked together in the same $N-$particle configuration $(i_1^* = \cdots = i_n^*)$, and therefore completely overlap with each other. Hence, the replica symmetric solution predicts the right saddle-point Q_1^{sp} which gives the correct physical picture, at least qualitatively, even if it is non exact as we shall see below.

Once we consider the analytic continuation of $\overline{\mathcal{Z}^n}$ to non integer numbers of replicas as $n \longrightarrow 0$, things start getting tricky. For one, the order of comparison between g_0 and g_1 , below and above $\beta_c(n)$ are reversed, and therefore the system becomes physically nonsensical with long range correlation at large temperatures and independent particles in the low temperature phase. The replica symmetric solution neither make sense mathematically; for $\beta \leq \beta_c(n)$, $g_1(\beta, n)$ dominates and thus $\overline{\mathcal{Z}}$ becomes nonlinear in n and hence does not go to 1 as $n \longrightarrow 0$.

In light of these observations, we can add the ad-hoc prescription, though mathematically unjustified, of choosing $argmin_{Q\in\mathcal{A}}\ g(Q)$ instead of its maximum as the correct saddle-point $n < 1$.

With this caveat, the replica symmetric procedure paints a physically sound qualitative picture for all n and we can therefore go on to verify the exactitude of the resulting free energy density. Recall that the saddle-point approximation only makes sense for $N \gg 1$, it is then imperative to start with the thermodynamic limit

$$\lim_{N \longrightarrow \infty} \overline{\mathcal{Z}_N^n} \doteq \lim_{N \longrightarrow \infty} \int e^{Ng(Q)} \prod_{a<b} dQ_{ab} = \exp \left\{ Ng(Q^{sp}) \right\}. \tag{3.73}$$

and only then apply the replica trick $f \propto \overline{\log \mathcal{Z}_\infty} = \lim_{n \to 0} \log(\overline{\mathcal{Z}_N^n})/n$ where $\mathcal{Z}_\infty$ is the partition function of the infinite system.

With that said, the replica method does rest on the (unjustified) assumption that the two limits $\lim_{n\to 0}$, $\lim_{N\to\infty}$ commute.

For all $\beta > \beta_c(1)$, $g_0(\beta, 1) > g_1(\beta, n)$ and therefore $\overline{\mathcal{Z}^n} \doteq e^{Ng_0(\beta,n)}$. In the high temperature regime, the replica symmetric solution then predicts that the free energy density satisfies

$$\begin{aligned} -\beta f(\beta) &\equiv \lim_{N \longrightarrow \infty} \frac{\overline{\log \mathcal{Z}}}{N} &(3.74)\\ &= \lim_{n \longrightarrow 0} \lim_{N \longrightarrow \infty} \frac{\log\left(\overline{\mathcal{Z}^n}\right)}{nN} &(3.75)\\ &= \lim_{N \longrightarrow \infty} \lim_{n \longrightarrow 0} \frac{\log\left(\overline{\mathcal{Z}^n}\right)}{nN} &(3.76)\\ &= \lim_{n \longrightarrow 0} \frac{g_0(\beta, n)}{n} &(3.77)\\ &= \frac{\beta^2}{4} + \log 2. &(3.78) \end{aligned}$$

Hence, the replica symmetric solution is exact for all $\beta < \beta_c(n)$. In the low temperature regime, since $g_1(\beta, n)$ dominates, by an identical computation we get

$$-\beta f(\beta) = \lim_{n \longrightarrow 0} \frac{n^2 \frac{\beta^2}{4} + \log 2}{n} \neq -\sqrt{\log 2}. \tag{3.79}$$

And therefore, the replica symmetric solution while qualitatively sound, is only exact in the high temperature phase.

3.4 Parisi's Replica Symmetry Breaking scheme

As the temperature decreases, a disordered system typically goes through a series of phase transitions, each characterized by a subtly different correlation structure. For simplicity, we will only consider the first three, and denote their respective critical points by $\beta_{RS} < \beta_{1RSB} < \beta_{2RSB}(or\ \beta_{cond})$, whose subscripts will be explained below;

0. $\beta \leq \beta_{RS}$: The high temperature paramagnetic or *liquid* phase is characterized by nearly independent spins and very low energy barriers, such that the system is strongly ergodic; in the sense of having a very short mixing time w.r.t. Glauber dynamics, spins are therefore uncorrelated in time, hence the liquid quality.

1. $\beta \in [\beta_{RS}, \beta_{1RSB})$: The first glassy phase is induced by the forming of a deep well in the energy landscape, delimited by infinite energy barriers, the system is rapidly mixing within the well.

2. $\beta \in [\beta_{1RSB}, \beta_{2RSB/cond})$: The second glassy phase, marked by the forming of subvalleys or *pure states*, delimited by $\mathcal{O}(e^N)$ energy barriers within the well, is accompanied by the emergence of short range correlations between spins. Moreover, the system is no longer rapidly mixing on the entire state space (i.e. has an $\mathcal{O}(e^N)$ mixing time), but is rapidly mixing when restricted to a pure state.

3. $\beta \geq \beta_{2RSB/cond}$: The third glassy phase can either consist of:

   i. The appearance of "sub-subvalleys" within sub-valleys, i.e. a further breaking of configuration space into short-correlated rapidly-mixing quasi-pure states, grouped within well separated clouds corresponding to pure states, thus defining a 2-level hierarchical landscape.

   ii. **Or**, the onset of a condensation phenomenon, where the equally sized 1RSB-pure states diverge in size, such that a linear number of clusters carry most of the probability mass, and where spins become long range correlated, in *the correlation viewpoint* subsection of chapter 4 we go into more details about what is meant by "short/long range" correlations.

Recall that a sample of the Random Energy model (REM) is given by a collection of realizations of zero mean Gaussian energy levels; one for each configuration in an $N-$particle Ising state space $\Sigma \equiv \{\pm 1\}^N$. For each pair of configurations $(i_k, i_l) \in \Sigma^2$, the overlap between the two is as previously defined $q_{i_k,i_l} \equiv \sum_{p\in[N]} i_{k_p} i_{l_p}/N$.

The replica symmetric (RS) framework, more generally conceived, assumes that the overlap between any two replicas $i_k, i_l \overset{iid}{\sim} \mu_\beta$ must either be zero, if they are identical, or equal to a certain value q_1 , naturally, for $N \gg 1$, they are unlikely to be identical and will therefore have an overlap equal to q_1 with high probability.

In essence, the replica symmetric assumption is a statement about the correlation structure of μ_{β^*} for all $\beta_{RS} \leq \beta^* < \beta_{1RSB}$, which stipulates that the correlation or *susceptibility* of the system χ^{SG} remains bounded as $N \to \infty$, such that

$$\lim_{N\longrightarrow\infty} \frac{\chi^{SG}}{N} = \lim_{N\longrightarrow\infty} \frac{\beta^2}{N^2} \sum_{x,y\in[N]^2} [\langle (i_l)_x (i_k)_y \rangle - \langle (i_l)_x \rangle \langle (i_k)_y \rangle]^2 = 0, \quad \forall\, \beta < \beta_{1RSB}. \tag{3.80}$$

What follows, is that the overlap distribution is tightly concentrated around a certain value $q_1 \in [-1, 1]$ such that; $\lim_{N\to\infty} P_{\sim\mu_{\beta,N}}[q_{i_k,i_l} = q] = \delta_{q_1,q}$.

Recall the pure state decomposition described above and the hierarchical energy landscape that ensues. The simplest possible scenario (case 1.) is one where there is no pure state decomposition, such that the support of the Gibbs measure consists of a large dense cluster, close in Hamming space, carrying almost all of the probability mass with no high energy barriers within.

This simple scenario is precisely the one being described by the replica symmetric solution, where $\forall\, \beta^* \in [\beta_{RS}, \beta_{1RSB})$, as $N \to \infty$, the state space $|\Sigma| = 2^N$ growing exponentially in N together with the vanishing correlation will make it such that any two configurations under $\mu_{\beta^*,N}$ will typically be very far in Hamming space, i.e. differing by a large number of spins, and will consequently have zero overlap with high probability.

As the temperature is lowered below the second critical point $\beta^* \geq \beta_{1RSB}$, the state space breaks into an exponential number of pure states $\Sigma = \mathcal{S} \bigsqcup_{i\in[\eta]} \alpha_i$ with $\eta = \mathcal{O}(e^N)$ and $\mathcal{S}$ being the configuration space separating clusters with vanishing probability mass below the critical temperature, allowing us to write μ_{β^*} as a convex combination of pure state weights:

$$\mu_{\beta^*}(\tau) = \sum_{\alpha_i : i\in[\eta]} w_{\alpha_i} \mu^{\alpha_i}_{\beta^*}(\tau) \quad \text{where} \quad \mu^{\alpha_i}_{\beta^*}(\tau) \equiv \frac{\mathbf{1}\{\tau \in \alpha_i\}\ e^{-\beta \mathcal{H}(\tau)}}{\mathcal{Z}_{\alpha i}}, \quad w_{\alpha_i} = \frac{\mathcal{Z}_{\alpha_i}}{\sum_{\alpha_i : i\in[\eta]} \mathcal{Z}_{\alpha i}}. \tag{3.81}$$

In this temperature regime, clusters are assumed to be asymptotically at equal distance from each other as $N \to \infty$, and of equal size, i.e. having equal Gibbs weights $w_{\alpha_i} = w_{\alpha_j}\ \forall\, i, j \in [\eta]$.

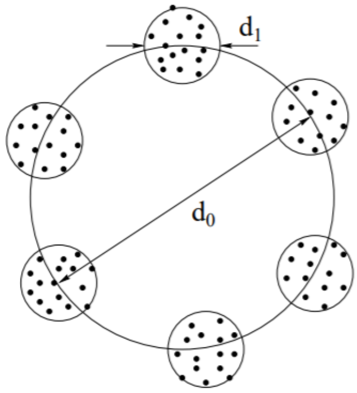


Figure 3.1: Clusters (Semerijan, 201X)

Hence, if we draw $i_k, i_l \overset{iid}{\sim} \mu_{\beta^*}$, the two configuration will be either in the same cluster or a different one, and both event will have non negligible probability as $N \to \infty$, i.e. there is no value of the overlap around which fluctuation under $\mu_{\beta^* \geq \beta_{RSB}}$ are vanishing in the thermodynamic limit, but rather two values that carry almost all of the probability mass for N large enough such that $\lim_{N\to\infty} P_{\sim\mu_{\beta,N}}(q_{i_k,i_l} = q) = \delta_{q1,q}\mathrm{x}(\beta) + \delta_{q0,q}(1 - \mathrm{x}(\beta))$, where $\mathrm{x}(\beta)$ is *the Parisi 1RSB parameter*, to be detailed below.

In a 2-level hierarchical landscape, each cluster decomposes into quasi-pure states $\alpha_i = \mathcal{Q}_i \bigsqcup_{k\in[\kappa^i]} \zeta^i_k$ with $\{\mathcal{Q}_{i\in[\eta]}\}$ having vanishing probability mass at low temperature. The RS assumption is supposed to hold within quasi-pure states, such that the distribution of the overlap when restricted to any quasi-pure state has vanishing fluctuation around a certain value q_ζ, as $N \to \infty$. Thus, any two independently random configurations from μ_β are most likely to be:

- In the same quasi-pure state $(i_k, i_l) \in \zeta^i_k$, with overlap q_ζ.

- In the same pure-state $(i_k, i_l) \in \alpha_i$ but different quasi-sates $i_k \in \zeta_k^i,\ i_l \in \zeta_l^i$ and hence typically have a smaller overlap $q_\alpha < q_\zeta$.
- In altogether different pure states and thus have the smallest overlap of the likely three values; $q_\Sigma < q_\alpha < q_\zeta$.

It is thus evident that to describe a $k-$level hierarchical landscape, we need $k-1$ overlap values.

The failure of the replica symmetric solution in the case of the REM at low temperature $\beta \geq \beta_c = 2\sqrt{\log 2}$, points to the fact that the sum $\overline{\mathcal{Z}^n} \doteq \sum_{\underline{Q}} e^{Ng(\underline{Q})}$ is dominated by non-replica symmetric overlap matrices.

The issue then lies within the assumption that the saddle-point matrix is invariant w.r.t. to permutation of indices, thus restricting the candidate pool from $\mathcal{B}_{n\times n}$ to matrices with equal off-diagonal elements; $\mathcal{A} \equiv \{Q_0^{sp}, Q_1^{sp}\}$, which amounts to assuming that a single overlap value suffices to describe each of the two phases delimited by β_c, effectively ignoring the possibility of having more than one possible overlap value for the low temperature phase, as is the case in a 1-level hierarchical energy landscape.

Parisi's ingenious *Replica Symmetry Breaking* (RSB) scheme is an iterative procedure which fixes this issue by recursively enlarging $\mathcal{A}$ to include non-replica symmetric overlap matrices by considering one additional possible overlap value for off diagonal elements at each iteration, as to account for the higher multiplicity of overlap values inherent to higher level hierarchies.

In practice, given a disordered system with a specified Hamiltonian, we start by assuming the simplest energy landscape for the low temperature phase, with no pure state decomposition, we write $\overline{\mathcal{Z}^n}$ in a saddle-point friendly form and assume that the overlap distribution concentrates around a single value, apply the RS solution, then carry out self-consistency checks. If the solution is not exact, we assume that the low temperature phase displays a 1-level hierarchical energy landscape, best described by an additional overlap (so 2 in total), this added overlap candidate breaks the symmetry of Q w.r.t S_n, since $\{Q_{ab} : a \neq b\}$ can now take more than one value.

This first iteration of the RSB scheme is the so called *1-step Replica Symmetry Breaking scheme* (1RSB), where we divide the n replicas into equally sized subsets and assume that replica symmetry $(Q_{ab} = Q_{\pi(a)\pi(b)})$ still holds within subgroups. Intuitively, we can think of each subgroup as a sample from a different pure state, which naturally defines two overlaps in the low temperature phase. Since the RS picture is valid within pure states, the overlap between same-subgroup configurations should be equal to one.

If the free energy density is still not exact for $\beta \geq \beta_c$, we assume a 2-level hierarchy, add a third candidate overlap, thus breaking the replica symmetry within subgroups, divide subgroups into "sub-subgroups" and assume that replica symmetry holds within "sub-subgroups", which defines the 2^{nd} step of RSB, and so on. Along the same lines, we can define kRSB for all $k \in \mathbf{N}$ with ∞RSB as the solution for infinitely nested clusters with in the limiting case.

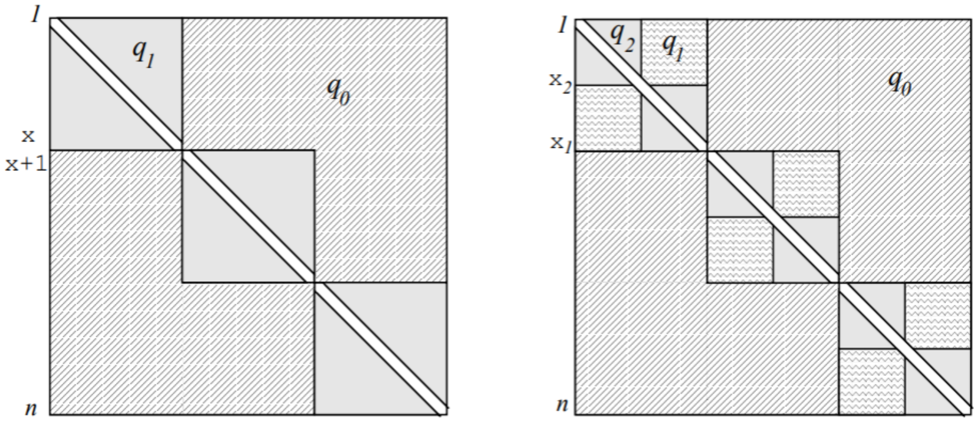


Fig. 8.4 Structure of the matrix Q_{ab} when replica symmetry is broken. *Left*: 1RSB Ansatz. The $n(n-1)/2$ values of Q_{ab} are the non-diagonal elements of a symmetric $n \times n$ matrix. The n replicas are divided into $n/\mathbf{x}$ blocks of size $\mathbf{x}$. When a and b are in the same block, $Q_{ab} = q_1$; otherwise, $Q_{ab} = q_0$. *Right*: 2RSB Ansatz: an example with $n/\mathbf{x}_1 = 3$ and $\mathbf{x}_1/\mathbf{x}_2 = 2$.

Figure 3.2: The saddle-point overlap matrix for the 1RSB (left) and 2RSB (right) schemes (Mezard and Montanari, 2009)

Suppose that n is a multiple of x, such that we can partition the replicas into n/x subsets of x elements each.

Back to the REM, in light of the failure of the RS solution, we assume that the energy landscape displays a 1-level hierarchy with a single pure state decomposition and consider as a saddle-point candidate, the following overlap matrix $Q \in \mathcal{B}_{n\times n}$:

- $Q_{aa} = 1, \ \forall \, a \in [n]$.
- $Q_{ab} = q_1$ if a and b are in the same subgroup.
- $Q_{ab} = q_0$ if a and b are in different subgroups.

Note that the 1-step RSB scheme compromises the replica symmetric solution as a special case for when $\mathrm{x} = 1$.

As previously stated, each subgroup of replicas can be seen as sampled from a different 1RSB cluster, within which the RS assumption holds. More precisely, the 1-RSB scenario presupposes that, if we restrict ourselves to a given pure state α_i, the correlation within said pure state is bounded, such that the overlap between any two configurations $i_k, i_l \overset{iid}{\sim} \mu_\beta^{\alpha_i}$ has vanishing fluctuation in the thermodynamic limit, and therefore the distribution of the overlap distribution restricted to α_i becomes a $\delta-$ function at one; $\lim_{N\to\infty} P_{\sim\mu_\beta^{\alpha_i}}(q_{i_k,i_l} = q) = \delta_{q,1}$.

Moreover, since clusters stay are supposed to be at equal distance in a Hamming space that grows exponentially in N; $|\Sigma| = 2^N$, any two configuration belonging to different pure states will have zero overlap with high probability.

Which leads us to two overlap candidates for the low temperature phase: $q_1 = 1$ and $q_0 = 0$.

Since the end goal is to approximate $\overline{\mathcal{Z}^n}$ up to leading exponential order, the next step is to compute the spectrum of the above saddle-point, i.e. the number of configurations $(i_1 \dots i_n)$ with the above overlap. Considering that $q_1 = 1$ and $q_0 = 0$, the n replicas would have to satisfy

$$i_1 = \dots = i_{\mathrm{x}} \neq i_{\mathrm{x}+1} = \dots = i_{2\mathrm{x}} \neq \dots \tag{3.82}$$

Hence,

$$\mathcal{N}_{RSB} = \underbrace{2^N}_{\#options\ for\ 1^{st}\ group} \underbrace{(2^N - 1)}_{2^{nd} group} \dots \underbrace{(2^N - \frac{n}{\mathrm{x}} + 1)}_{\mathrm{x}^{th} group} . \tag{3.83}$$

Using the same argument as before, the drastic drop in the number of summands after the change of variable from $\overline{\mathcal{Z}^n} = \sum_{(i_1 \dots i_n)}$ to $\sum_{\underline{Q} \in \mathcal{B}_{n\times n}} \mathcal{N}_{RSB}(\underline{Q})$ leads us to assume a large deviation principle of the form $\mathcal{N}_{RSB}(\underline{Q}) \doteq e^{Ns(\underline{Q})}$, such that

$$\lim_{N \longrightarrow \infty} \frac{1}{N} \Big(\frac{nN}{\mathrm{x}} \log 2 - Ns(Q) \Big) = 0 \quad \Longleftrightarrow \quad s(Q) = \frac{n}{\mathrm{x}} \log 2 \tag{3.84}$$

Hence, the saddle-point satisfies $g_{RSB}(\beta, n, \mathrm{x}) = \frac{\beta^2}{4} n\mathrm{x} + \frac{n}{\mathrm{x}} \log 2$.

To summarize the steps taken thus far, using the Gaussian identity $\mathbf{E}_{E \sim \mathcal{N}(0,\Delta)}[e^{\lambda E}] = e^{\lambda^2 \Delta/2}$, we compute the average of $\mathcal{Z}^n$ over the Gaussian energy levels $\{E_{i_k}\}$, thus obtaining

$$\begin{aligned}
\overline{\mathcal{Z}^n} &= \sum_{i_1 \dots i_n} \exp \Big\{ \frac{N\beta^2}{4} \sum_{1 \le a < b \le n} \mathbf{1}\{i_a = i_b\} \Big\}, \quad \mathcal{N}(\underline{Q}) \equiv \Big| \Big\{ (i_1 \cdots_n) : \underbrace{Q_{ab}}_{\mathbf{1}\{i_a = i_b\}} = \underline{Q}_{ab} \Big\} \Big| \\
&\doteq \sum_{\underline{Q}} \mathcal{N}(\underline{Q})\ e^{\frac{N\beta^2}{4} \sum_{a<b} \underline{Q}_{ab}} \quad \text{and suppose} \quad \mathcal{N}(\underline{Q}) \doteq e^{Ns(\underline{Q})} \\
&\doteq \sum_{\underline{Q}} \exp \Big\{ N \underbrace{\Big(\frac{\beta^2}{4} \sum_{a<b} \underline{Q}_{ab} + s(\underline{Q}) \Big)}_{\equiv g(\underline{Q})} \Big\} = \int_{[-1,1]} e^{Ng(\underline{Q})} \prod_{a<b} \underline{Q}_{ab} \quad \text{for} \quad N \gg 1.
\end{aligned}$$

In the Replica Symmetric procedure we assumed Q to be symmetric w.r.t any permutation $\pi \in S_n$ where S_n is the permutation group of n elements, we now relax this symmetry condition by assuming that replica symmetry holds only within subgroups, i.e $Q_{\pi(a)\pi(b)} = Q_{ab}$, for all $\pi \in S_{\mathrm{x}}$; $a, b \in \{p\mathrm{x} + 1, \dots (p+1)\mathrm{x}\}$ and $p \in [n/\mathrm{x}]$.

What is missing now, is how many subgroups should there be for $g_{RSB}(\beta, n, \mathrm{x}^*)$ to dominate the above sum, in other words; for what value of x does g_{RSB} provide the correct saddle-point, in the low temperature phase; $\beta \geq \beta_c$.

Following the same argument used to remediate the non-physically sensical case of non-integer $n < 1$ in the RS picture, we take the minimum rather than the maximum

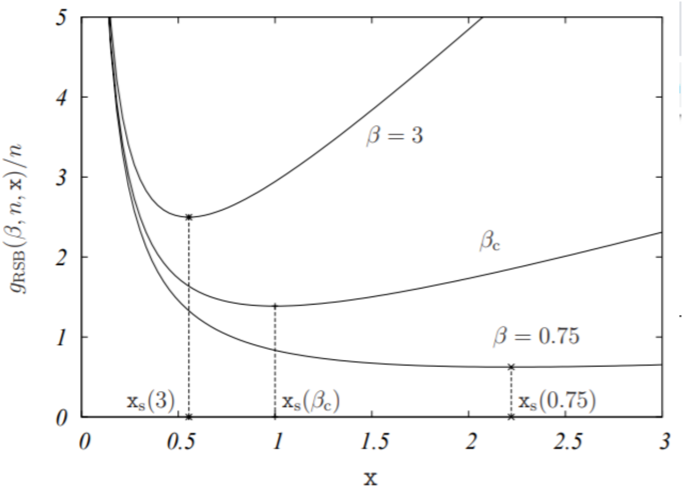


Figure 3.3: $argmin_{\mathrm{x}}\ g_{RSB}(\beta, n, \mathrm{x})$ at different temperatures (Mezard, Montanari, 2009)

as the saddle-point approximation of $\int_{[-1,1]} e^{Ng(\underline{Q})} \prod_{a<b} \underline{Q}_{ab}$, wich yields

$$\frac{\partial g_{RSB}}{\partial \mathrm{x}} = \frac{\beta}{4} n - \frac{n}{\mathrm{x}^2} \log 2 = 0 \quad \Longleftrightarrow \quad \mathrm{x}^*(\beta) = \underbrace{2\sqrt{\log 2}}_{=\beta_c(n=1)} / \beta. \tag{3.85}$$

Again, assuming $\lim_{n\to 0}$ and $\lim_{N\to\infty}$ commute, we verify the exactitude of the RSB saddle-point for the glassy low temperature phase $\beta \geq \beta_c(1)$,

$$-\beta f = \lim_{n\to 0} \frac{1}{n}\, g_{RSB}(\beta, n, \mathrm{x}) = \lim_{n\to 0} \frac{2\sqrt{\log 2}\; n\beta^2}{4n\beta} + \frac{\log 2\; n\beta}{2n\sqrt{\log 2}} = \beta\sqrt{\log 2}, \tag{3.86}$$

hence, the 1RSB solution is exact for all $\beta \geq \beta_c(1) = 2\sqrt{\log 2}$.

Note that, while the RS free energy density is in fact correct for $\beta < \beta_c$, it may feel unsatisfying to have to use the RS scheme for high temperatures, then switch to 1RSB for the low temperature phase. This problem can be mended by fixing $\mathrm{x} = 1$ for the high temperature phase, such that for all $\beta < \beta_c$, we have:

$$-\beta f = \lim_{n\to 0} \frac{1}{n}\, g_{RSB}(\beta, n, 1) = \lim_{n\to 0} \frac{\frac{\beta}{4} n.1 - \frac{n}{1}\log 2}{n} = \frac{\beta^2}{4} + \log 2, \tag{3.87}$$

thus, recovering the correct value of the free energy density for both the high and low temperature regimes.

3.5 The p-spin model

Considering the failure of the Replica Symmetric solution to predict the low temperature free energy density of the Sherrington Kirkpatrick (SK) model, (Derrida, 1980) proposed a generalization of the SK model at zero magnetic field, which allows interactions between any $p-$tuple of spins, called *the p-spin model*:

Definition 22. *The p-spin model is an $N-$particle system where the set of couplings*

$\mathcal{J}$ consists of all $\binom{N}{p}$ possible $p-$tuple of spin-interaction, and whose Hamiltonian is given by:

$$\mathcal{H}(\sigma) = -\sum_{1\leq i_1<\cdots<i_p\leq N} \underline{J}_{i_1\dots i_p}\underline{\sigma}^a_{i_1}\cdots\underline{\sigma}^a_{i_p}, \tag{3.88}$$

where $J_{i_1\dots i_p} \overset{iid}{\sim} \mathcal{N}(0,\Delta^2)$.

The SK model is then a p-spin model with $p=2$.

After setting up this model, Derrida bypasses this difficult looking Hamiltonian, by considering a much simpler object; the *Random Energy Model*, which turns out to be the limiting case of the p-spin model when $p\to\infty$.

To see this, the trick is to consider the probability distribution of the pair of values that a couple of Hamiltonian can take:

$$\mathbf{P}_{\sim\mu_{\beta,p}}[E_1,E_2] = \langle \delta_{\mathcal{H}(\sigma),E_1}.\,\delta_{\mathcal{H}(\sigma),E_2}\rangle, \tag{3.89}$$

and introduce an overlap-like parameter, namely the Hamming distance between two configurations in the Ising $N-$particle state space: $x_{i_k,i_l} \equiv |\{p:\,(i_k)_p=(i_l)_p\}|$. After some standard manipulations, it follows that

$$\mathbf{P}_{\sim\mu_{\beta,p}}[E_1,E_2] \propto\ \exp\Big\{-\frac{(E_1+E_2)^2}{2N[1+(2x-1)^p]\Delta^2}-\frac{(E_1-E_2)^2}{2N[1-(2x-1)^p]\Delta^2}\Big\} \tag{3.90}$$

Depending on the value of this parameter, we distinguish two cases:

i. $\mathbf{P}_{\sim\mu_{\beta,p}}[E_1,E_2] \underset{x\to1}{\longrightarrow} \mathbf{P}_{\sim\mu_{\beta,p}}[E_1].\delta_{E_1,E_2}$, hence E_1,E_2 become strongly correlated.

ii. If $x\in(0,1/2]$, $\lim_{p\to\infty}(2x-1)^p=0$ and therefore: $\lim_{p\to\infty}\mathbf{P}_{\sim\mu_{\beta,p}}[E_1,E_2] \propto e^{-E_1^2/2N\Delta^2}e^{-E_2^2/2N\Delta^2}$. Hence, $\lim_{p\to\infty}\mathbf{P}_{\sim\mu_{\beta,p}}[E_1,E_2] = \mathbf{P}_{\sim\mu_{\beta,p}}[E_1].\,\mathbf{P}_{\sim\mu_{\beta,p}}[E_2]$, the two energies are independent.

It is thus, apparent that when the number of interacting spins in the $p-$spin model goes to infinity, we recover the REM. And while the REM is very insightful as a toy model, the canonical mean field model for spin glasses is in fact the $p-$spin.

3.5.1 Breaking replica symmetry in the p-spin case

Definition 23. *A set of n replicas is a set of configurations $\{\sigma^a\}_{a\in[n]}$ drawn independently from the same distribution (i.e. with fixed couplings).*

Theorem 10 (Mezard an Montanari, 2009)**.** *The expectation w.r.t. the distribution of the couplings of the n^{th} power of the partition function of an $N-$particle $p-$spin model satisfies:*

$$\overline{\mathcal{Z}^n} \doteq \int e^{\frac{N\beta^2 n}{4}} \sum_{\underline{\sigma}^1\dots\underline{\sigma}^n} \exp\Big\{\frac{N\beta^2}{4}\sum_{a,b}\underline{Q}^p_{ab}\Big\}.\delta(q_{ab}-\underline{Q}_{ab})\prod_{a<b}dq_{ab}, \tag{3.91}$$

where $\underline{Q}_{ab} \equiv \sum_i \underline{\sigma}_i^a \underline{\sigma}_i^b / N$.

Proof. As we did before, to ease notation, we denote the expectation w.r.t. the distribution of the couplings with an overline. We have

$$\mathcal{Z}^n = \prod_{a=1}^{n} \Big(\sum_{\underline{\sigma}^a \in \Sigma} \exp\Big\{\beta \sum_{i_1 < \cdots < i_p} \underline{J}_{i_1 \dots i_p} \underline{\sigma}_{i_1}^a \dots \underline{\sigma}_{i_p}^a \Big\}\Big) \tag{3.92}$$

$$\sum_{\underline{\sigma}^1 \dots \underline{\sigma}^n} \prod_{a=1}^{n} \exp\Big\{\beta \sum_{i_1 < \cdots < i_p} \underline{J}_{i_1 \dots i_p} \underline{\sigma}_{i_1}^a \dots \underline{\sigma}_{i_p}^a \Big\} \tag{3.93}$$

$$= \sum_{\underline{\sigma}^1 \dots \underline{\sigma}^n} \prod_{i_1 < \cdots < i_p} \exp\Big\{ \underline{J}_{i_1 \dots i_p} \beta \underbrace{\Big(\sum_{a=1}^{n} \underline{\sigma}_{i_1}^a \dots \underline{\sigma}_{i_p}^a \Big)}_{\equiv \gamma_{i_1 \dots i_p}} \Big\} \tag{3.94}$$

$$= \sum_{\underline{\sigma}^1 \dots \underline{\sigma}^n} \prod_{i_1 < \cdots < i_p} \exp\Big\{ \underline{J}_{i_1 \dots i_p} \Big(\beta \gamma_{i_1 \dots i_p} \Big) \Big\} \tag{3.95}$$

Recall the Gaussian identity $\mathbf{E}_{Z \sim \mathcal{N}(0,\Delta^2)} e^{\lambda Z} = e^{\frac{\Delta \lambda^2}{2}}$. Since $\overline{J_{i_1 \dots i_p}^2} = \frac{p!}{2N^{p-1}}$ and $\lambda \equiv \beta \gamma_{i_1 \dots i_p}$, we have

$$\overline{\mathcal{Z}^n} = \sum_{\underline{\sigma}^1 \dots \underline{\sigma}^n} \prod_{i_1 < \cdots < i_p} \exp\Big\{ \frac{p!}{2N^{p-1}} \frac{\beta^2 \gamma_{i_1 \dots i_p}^2}{2} \Big\} \tag{3.96}$$

$$= \sum_{\underline{\sigma}^1 \dots \underline{\sigma}^n} \exp\Big\{ \frac{N \beta^2}{4} \frac{p!}{N^p} \sum_{i_1 < \cdots < i_p} \gamma_{i_1 \dots i_p}^2 \Big\} \tag{3.97}$$

And since

$$\gamma_{i_1 \dots i_p}^2 = \Big(\sum_{a=1}^{n} \underline{\sigma}_{i_1}^a \dots \underline{\sigma}_{i_p}^a \Big)^2 = \sum_{a,b\ \in [n]^2} \underline{\sigma}_{i_1}^a \underline{\sigma}_{i_1}^b \underline{\sigma}_{i_2}^a \underline{\sigma}_{i_2}^b \dots \underline{\sigma}_{i_p}^a \underline{\sigma}_{i_p}^b , \tag{3.98}$$

we have

$$p! \sum_{i_1 < \cdots < i_p} \gamma_{i_1 \dots i_p}^2 = p! \underbrace{\sum_{i_1 < \cdots < i_p}}_{\binom{N}{p}\ terms} \sum_{a,b\ \in [n]^2} \underline{\sigma}_{i_1}^a \underline{\sigma}_{i_1}^b \underline{\sigma}_{i_2}^a \underline{\sigma}_{i_2}^b \dots \underline{\sigma}_{i_p}^a \underline{\sigma}_{i_p}^b \tag{3.99}$$

$$\doteq \sum_{(i_1 \dots i_p) \in [N]^p} \sum_{a,b\ \in [n]^2} \underline{\sigma}_{i_1}^a \underline{\sigma}_{i_1}^b \underline{\sigma}_{i_2}^a \underline{\sigma}_{i_2}^b \dots \underline{\sigma}_{i_p}^a \underline{\sigma}_{i_p}^b . \tag{3.100}$$

The next step is to formulate $\overline{\mathcal{Z}^n}$ as a function of the *overlap* between fixed assignments of replicas $(\underline{\sigma}^a, \underline{\sigma}^b)$: $\underline{Q}_{ab} \equiv \sum_i \underline{\sigma}_i^a \underline{\sigma}_i^b / N$:

$$N^p \underline{Q}_{ab}^p \equiv \Big(\sum_{i \in [N]} \underline{\sigma}_i^a \underline{\sigma}_i^b \Big)^p = \big(\underline{\sigma}_1^a \underline{\sigma}_1^b + \cdots + \underline{\sigma}_N^a \underline{\sigma}_N^b \big) \overset{p\ times}{\cdots} \big(\underline{\sigma}_1^a \underline{\sigma}_1^b + \cdots + \underline{\sigma}_N^a \underline{\sigma}_N^b \big) \tag{3.101}$$

$$= \sum_{i_1 \dots i_p} \underline{\sigma}_{i_1}^a \underline{\sigma}_{i_1}^b \dots \underline{\sigma}_{i_p}^a \underline{\sigma}_{i_p}^b . \tag{3.102}$$

Thus,

$$\sum_{a,b} \underline{Q}_{ab}^{p} \doteq \frac{p!}{N^p} \sum_{i_1<\cdots<i_p} \gamma^2_{i_1 \ldots i_p} \tag{3.103}$$

And since, $Q_{aa} = \sum_i (\sigma_i^a)^2/N \ = 1$, we have

$$\sum_{(a,b)\in[n]^2} \underline{Q}_{ab}^{p} = n + 2 \sum_{1\leq a<b\leq b} Q_{ab}^{p}. \tag{3.104}$$

Such that,

$$\overline{\mathcal{Z}^n} \doteq \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \exp\Big\{\frac{N\beta^2}{4} \sum_{a,b} \underline{Q}_{ab}^{p}\Big\} = e^{\frac{N\beta^2 n}{4}} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \exp\Big\{\frac{N\beta^2}{2} \sum_{a<b} \underline{Q}_{ab}^{p}\Big\}. \tag{3.105}$$

Finally, to rid the exponential of its dependence (through $\underline{Q}_{ab}$) on the assignment of the n replicas $\{\underline{\sigma}^a\}$, we use the Dirac delta function to obtain:

$$\overline{\mathcal{Z}^n} \doteq \int e^{\frac{N\beta^2 n}{4}} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \exp\Big\{\frac{N\beta^2}{4} \sum_{a,b} \underline{Q}_{ab}^{p}\Big\}.\delta(q_{ab} - \underline{Q}_{ab}) \prod_{a<b} dq_{ab}. \quad \blacksquare \tag{3.106}$$

Now, the next step is to write the above integral in a saddle-point friendly form (like $\int e^{Ng(q)}dq$). To do this, we make use of the Laplace transform of the Dirac delta function $\delta(s-t) = \frac{1}{2\pi}^{-i\zeta(s-t)}d\zeta$ such that, for all $1 \leq a < b \leq n$ we have

$$\begin{aligned} 1 &= \int \delta\Big(q_{ab} - \frac{\sum_i \underline{\sigma}_i^a \underline{\sigma}_i^b}{N}\Big) dq_{ab} && (3.107) \\ &= \int \Bigg[\frac{1}{2\pi} \int \exp\Big\{ - i\lambda_{ab}\Big(Nq_{ab} - \sum_i \underline{\sigma}_i^a \underline{\sigma}_i^b\Big)\Big\} d\lambda_{ab}\Bigg] N dq_{ab}, && (3.108) \end{aligned}$$

hence

$$\begin{aligned} \overline{\mathcal{Z}^n} &\doteq e^{\frac{N\beta^2 n}{4}} \int \Bigg[\prod_{a<b} \Big\{ e^{\frac{N\beta^2}{4} q_{ab}^p} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \delta(q_{ab} - \underline{Q}_{ab})\Big\}\Bigg] \prod_{a<b} dq_{ab} && (3.109) \\ &\doteq e^{\frac{N\beta^2 n}{4}} \int \Bigg[\prod_{a<b} \Big\{ e^{\frac{N\beta^2}{4} q_{ab}^p} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \int e^{-i\lambda_{ab}(Nq_{ab} - \sum_i \underline{\sigma}_i^a \underline{\sigma}_i^b)} d\lambda_{ab}\Big\}\Bigg] \prod_{a<b} dq_{ab} && (3.110) \\ &\doteq e^{\frac{N\beta^2 n}{4}} \int \Bigg[\prod_{a<b} \Big\{ e^{\frac{N\beta^2}{4} q_{ab}^p}\, e^{-iN\lambda_{ab}\, q_{ab}} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \int e^{i\lambda_{ab}\, \overrightarrow{\underline{\sigma}^a}\overrightarrow{\underline{\sigma}^b}} d\lambda_{ab}\Big\}\Bigg] \prod_{a<b} dq_{ab} && (3.111) \\ &\doteq e^{\frac{N\beta^2 n}{4}} \int\int e^{\frac{N\beta^2}{4} \sum_{a<b} q_{ab}^p - iN\sum_{a<b} \lambda_{ab}\, q_{ab}} \sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} e^{i\sum_{a<b} \lambda_{ab}\, \overrightarrow{\underline{\sigma}^a}\overrightarrow{\underline{\sigma}^b}} \prod_{a<b} dq_{ab}\, d\lambda_{ab}. && (3.112) \end{aligned}$$

Let G be a function of $2 \times \frac{n(n-1)}{2}$ variables, mapping $(\{q_{ab}\}, \{\lambda_{ab}\}) \mapsto \mathbf{R}:$

$$G(\{q_{ab}\}, \{\lambda_{ab}\}) \equiv -\frac{n\beta^2}{4} - \frac{\beta^2}{2} \sum_{a<b} Q_{ab}^p + i \sum_{a<b} \lambda_{ab} Q_{ab} - \log \Big(\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} e^{\sum_{a<b} i\lambda_{ab} \overrightarrow{\sigma^a} . \overrightarrow{\sigma^b}} \Big), \tag{3.113}$$

such that

$$\overline{\mathcal{Z}^n} \doteq \int e^{-NG(\{q_{ab}\},\{\lambda_{ab}\})} \prod_{a<b} dq_{ab}\, \lambda_{ab}. \tag{3.114}$$

We can now approximate the integral by a saddle-point. Let $w_{ab} \equiv i\lambda_{ab}$, since the saddle-point must be a critical point of G, we have

$$\frac{\partial G}{\partial q_{ab}} = -\frac{\beta^2}{2} p\; q_{ab}^{p-1} + w_{ab} \Longleftrightarrow w_{ab}^* = \frac{\beta^2}{2} p\; q_{ab}^{p-1}. \tag{3.115}$$

As for $\{q_{ab}^*\}$,

$$\frac{\partial G}{\partial w_{ab}} = q_{ab} - \frac{\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \overrightarrow{\underline{\sigma}}^a \overrightarrow{\underline{\sigma}}^b\; e^{\sum_{a<b} w_{ab} \overrightarrow{\underline{\sigma}}^a \overrightarrow{\underline{\sigma}}^b}}{\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} e^{\sum_{a<b} w_{ab} \overrightarrow{\underline{\sigma}}^a \overrightarrow{\underline{\sigma}}^b}}. \tag{3.116}$$

The second term on the right hand side defines a probability measure whose support is $(\underline{\sigma}^1, \ldots \underline{\sigma}^n)$, with expectation

$$\langle \mathcal{O} \rangle_n \equiv \frac{\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} \mathcal{O}(\underline{\sigma}^1, \ldots \underline{\sigma}^n)\; e^{\sum_{a<b} w_{ab} \overrightarrow{\underline{\sigma}}^a \overrightarrow{\underline{\sigma}}^b}}{\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} e^{\sum_{a<b} w_{ab} \overrightarrow{\underline{\sigma}}^a \overrightarrow{\underline{\sigma}}^b}}, \quad \text{for all} \quad \mathcal{O} : \Sigma^n \mapsto \mathbf{R}. \tag{3.117}$$

The saddle-point for $\{q_{ab}^*\}$ then satisfies

$$q_{ab}^* = \langle \sum_i \sigma_i^a \sigma_i^b \rangle_n, \quad \text{with} \quad \sigma^a, \sigma^b \overset{iid}{\sim} \mu_{\beta, \underline{\mathcal{J}}}. \tag{3.118}$$

The next step is to find a candidate saddle-point satisfying the conditions above. As a first step, we consider a RS solution in the set of symmetric matrices with unit diagonal $\mathcal{B}_{n \times n}$, as sketched in the REM case,

$$q_{ab} = q, \quad w_{ab} = w \quad \forall a \neq b \tag{3.119}$$

which yields

$$w = \frac{\beta^2 p\; q^{p-1}}{2} \tag{3.120}$$

and since, $q = \langle \sum_i \sigma_i^a \sigma_i^b \rangle_n$, using the Gaussian identity we get

$$q = \mathbf{E}_z \tanh^2(z\sqrt{w}) \quad \text{where} \quad z \sim \mathcal{N}(0,1). \tag{3.121}$$

One possible solution for the above identities, is to take $q = w = 0,$ which yields

$$\lim_{n \to 0} \frac{1}{n} G(\{q_{ab} = 0\}, \{\lambda_{ab} = 0\}) = \lim_{n \to 0} \frac{1}{n} \Big[\frac{n\beta^2}{4} - \log \Big(\sum_{\underline{\sigma}^1 \ldots \underline{\sigma}^n} e^0 \Big) \Big] = -\frac{\beta}{4} - \frac{\log 2}{\beta}. \tag{3.122}$$

An important detail that we omitted thus far, is the $p-$dependence of $G,$ which, as can be seen in the two graphs above, distinguishes two qualitatively different cases;

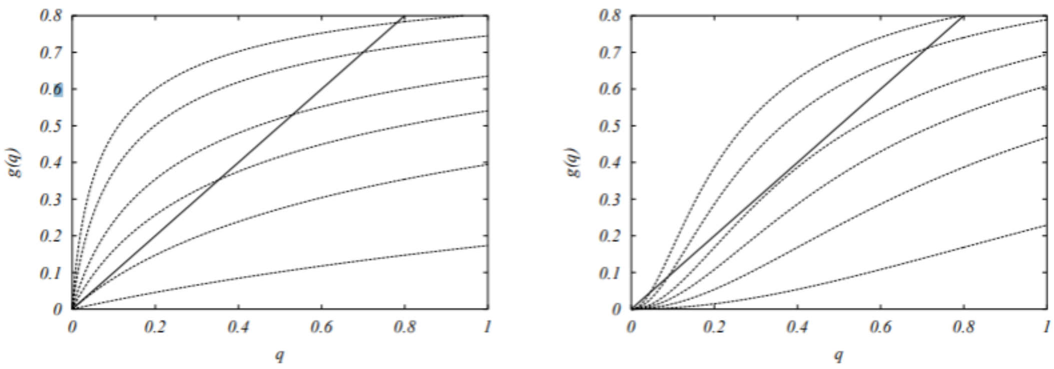


Figure 3.4: The graph of $r(q) \equiv \mathbf{E}_{z\sim\mathcal{N}(0,1)} \tanh^2(z\sqrt{p\beta^2 q^{p-1}/2})$ at $\beta =$ 4, 3, 2, ,1.5, 1 and 0.5 for the SK model (left) and $p = 3$ (right), from (Mezard and Montanari, 2009).

$p = 2$. which corresponds to the SK model, where G sees the appearance of a second critical point at $\beta = 1$, which then depart continuously (as $\beta = 1 + \epsilon$) from the first at $q = 0$.

$p \geq 3$. for each p above 3, there exists a $p-$dependent critical temperature $\beta_c(p)$, above which, a second critical point appears, and as $\beta \downarrow \beta_c(p)$, the two critical points merge into one.

Unfortunately, all of these RS saddle-points turn out to produce a highly inaccurate physical picture, where the system's entropy is allowed to decrease, which is actually one of the biggest sins in physics.

Unlike in the REM case, which was a toy model whose main purpose was to show when the RS(B) scheme does/fails to predict the correct value of $f(\beta)$, that we a priori knew, we do not have an explicit result for the free energy density to check the veracity of the RS predictions, and hence we need to resort to a more fundamental principle.

Recall that the entropy of a probability measure with discrete support is a function of the temperature: $S(\beta) = \sum_{\underline{\sigma}} \mu_\beta(\underline{\sigma}) \log[\mu_\beta(\underline{\sigma}])$, in the large temperature limit it is easy to see that $S(\beta) = \log|\mathcal{X}| + \Theta(\beta)$, where $\mathcal{X}$ is the state space of a single spin, and is equal to $\{\pm 1\}$ in the Ising case.

If we define *the energy gap*, i.e. the height of the energy of a given state $\underline{\sigma}$, w.r.t a ground state E_0, as $\Delta E = \min_{\underline{\sigma}\in\Sigma}\{\mathcal{H}(\underline{\sigma}) - E_0\}$, we can rewrite the entropy in a low temperature expansion as $S(\beta^*) = \log|\mathcal{X}| + \Theta(e^{-\beta\Delta E})$ for small β^*. Since $\Delta E \geq 0$ by definition, we see that, that at low temperature, the entropy essentially counts the number of degrees of freedom $|\mathcal{X}|$ of the system.

This is known as the *positivity of the entropy*, and is a crucial condition to be satisfied for any theory to be satisfactory. A direct consequence of which;

$$F(\beta) = U(\beta) - \frac{S(\beta)}{\beta} = E_0 - \frac{\log|\mathcal{X}|}{\beta} + \Theta(-\beta\Delta E) \overset{\beta\gg 1}{\approx} E_0 - \frac{\log|\mathcal{X}|}{\beta}. \qquad (3.123)$$

Hence, the free energy density $f(\beta) \equiv \lim_{N\to\infty} F_N(\beta)/N$, must also be a increasing function of β, which is not true in the RS prediction above.

The RS prediction is thus fundamentally flawed, hence the 1RSB solution. Just as in the REM case, we partition the replicas into n/x subsets and consider the 1RSB saddle-point:

- $q_{aa} = w_{aa} = 1, \; \forall\, a \in [n]$.
- $q_{ab} = q_1, \; w_{ab} = w_1$ if a and b are in the same subgroup.
- $q_{ab} = q_0, \; w_{ab} = w_0$ if a and b are in different subgroups.

By repeated use of the Gaussian identity we get,

$$\begin{aligned} G_{1RSB}(\{q_{ab}\},\{w_{ab}\}) = &-\frac{n\beta^2}{4} + \frac{n\beta^2}{4}\Big((1-\mathrm{x})q_1^p + \mathrm{x}q_0^p\Big) - \frac{n}{2}\Big((1-\mathrm{x})q_1w_1 + \mathrm{x}q_0w_0\Big) \\ &+ \frac{n}{2}w_1 - \log\Big(\mathbf{E}_{z_0\sim\mathcal{N}(0,1)}\left[\mathbf{E}_{z_1\sim\mathcal{N}(0,1)}2^{\mathrm{x}}\cosh^{\mathrm{x}}\left(\sqrt{w_0}z_0 + \sqrt{w_1 - w_0}z_1\right)\right]^{\frac{n}{\mathrm{x}}}\Big), \end{aligned}$$

where z_0, z_1 are, as suggested, two independent zero mean Gaussians with unit variance.

Proceeding as before, we set $q_0 = w_0 = 0$, and look for the critical points of G_{1RSB}, we find that

$$w_1 = \frac{1}{2}p\,\beta^2 q_1^{p-1}, \quad q_1 = \frac{\mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[2^{\mathrm{x}}\cosh^{\mathrm{x}}(\sqrt{w_1}z)\tanh^2(\sqrt{w_1}z)\right]}{\mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[2^{\mathrm{x}}\cosh^{\mathrm{x}}(\sqrt{w_1}z)\right]}. \tag{3.124}$$

One obvious solution to these equations is the replica symmetric $q_1 = w_1 = 0$ overlap matrix, which as we said is erroneous, hence the need to look for other saddle-points.

Just like in the RS case, for $p \geq 3$, the above identity (on the right) admits two solutions away from $q_1 = 0$. A local stability analysis, whose details are expanded upon in the last section of chapter 8 in (Mezard and Montanari, 2009), reveals that the larger one of the two solutions $q_1^* > q_1^{**}$; is the suitable one.

With both $(q_1 = q_1^*, q_0 = 0)$ fixed, we minimize G_{1RSB} w.r.t x, and what we find is that; the unicity of a critical point $\mathrm{x}^* \in [0,1]$ holds only in the low temperature regime $\beta > \beta_c(p)$.Notice, the strict inequality for β, more precisely, at the critical temperature $\beta = \beta_c(p)$, we have

$$f_{1RSB} = \lim_{n\to 0}\frac{1}{n\beta_c(p)}G_{1RSB} = -\frac{\beta}{4} - \frac{\log 2}{\beta} = f_{RS}, \tag{3.125}$$

which effectively means that, for all $p \geq 3$, there exists a phase transition $\beta = \beta_c(p)$, from the RS 1-level hierarchical picture to the 1RSB one characterized by a pure state decomposition.

By some involved computations that we omit for concision, we find that for $p \gg 1$, we have

$$\beta_c(p) = 2\sqrt{\log 2} + e^{-\Theta(p)}, \quad \mathrm{x}^*(\beta) = \frac{\beta_c(p)}{\beta} + e^{-\Theta(p)}, \quad q_1 = 1 - e^{-\Theta(p)}. \tag{3.126}$$

Moreover, for $\beta = \beta_c(p) + \epsilon$, $f_{1RSB}(\beta) = -\sqrt{\log 2}$, thus painting essentially the same picture as the REM at low temperature, hence reaffirming the validity of the REM as a large p limit for the more complicated $p-$spin model.

A further stability analysis, by Gardner, reveals that for $p \geq 3$, the 1RSB solution is only stable up to a certain $\beta_u(p)$, above which any kRSB for finite $k < \infty$ turn out to be unstable. In fact, at $\beta_u(p)$, the thermodynamic overlap distribution becomes a combination of an infinite number of delta functions, thus signaling an infinitely nested hierarchical structure requiring an $\infty-$RSB scheme. As for the $p = 2$ case of the SK model, one finds that at the appearance of the second critical point of G at $\beta = 1$, signals a phase transition, where $\forall\ \beta > 1$, the system is full (∞) Replica Symmetry Breaking.

Part II

Statistical physics of random CSPs

Chapter 4

Statics or the study of the equilibrium measure

Chapter organization

We start by introducing a well known constraint satisfaction problem under the name of k-SAT and its factor graph representation (**4.1**). Then, we introduce a crucial result concerning the connectivity of random factor graphs (associated with random k-SAT instances), that states that they are locally a tree (**4.1.1**). Afterwards, we introduce the message passing approach to computing the normalization constant and approximating marginals of high dimensional probability distributions, and show that it is exact on tree factor graphs (**4.1.2**). Then, we introduce some definitions that allow us to define a notion of a "good marginal approximation", and show that a message passing algorithm under the name of *Belief propagation* satisfies it under some conditions (**4.2**, **4.2.1**). Subsequently, we describe the significance of the 1RSB scenario discussed in the previous chapter in the case of k-SAT from two perspectives (**4.3.1** and **4.3.2**), and move on to discuss a message passing approach that works well beyond some critical variable/clause density threshold (where Belief propagation fails), called *Survey propagation* (**4.4**). Finally, we discuss an inherent sampling bias in Survey propagation, and the problem of sampling *uniformly* from the set of satisfying assignments (**4.5**), that we address in the last remaining chapter.

4.1 The k-SAT problem

Definition 24. *A logical conjunction over a set of binary variables $c_i \in \{0,1\}$ for $i \in [M]$, is an* ***and*** *operation between them that we denote by $\bigwedge_{i\in[M]} c_i$. A logical disconjunction over a set of binary variables $x_i \in \{0,1\}$ for $i \in [k]$, is an* ***or*** *operation between them, that we denote by $\bigvee_{i\in[k]} x_i$.*

Definition 25. *A k-SAT formula is a logical conjunction between M binary valued clauses $\bigwedge_{i\in[M]} c_i$, where each clause is given by a logical disjunction between a subset of k variables (and/or their negations $\neg$) from $\{x_i\}_{i\in[N]}$: $c_i = \bigvee_{i\in[k]}(\neg)x_i$, for some $k, N, M \in \mathbf{N}$.*

Following the notation in (Mezard and Montanari, 2009), we denote the set of variables involved in clause a by ∂a to mean $a = \bigvee_{i\in\partial a} x_i$, and the set of clauses in which the variables x_i occurs by ∂i. A $k-$SAT formula is then uniquely determined by the sets $\partial c_i, i \in [M]$ and can thus be written as: $\Phi(\{x_i\}) = \bigwedge_{j\in[M]} \bigvee_{i\in\partial c_j} (\neg) x_i$.

Note that we will often abuse notation to use interchangeably x_i and i to refer to the i^{th} variable. Here is a simple example of a $2-$SAT formula with $N = 3, M = 2$:

$$\Phi(x_1, x_2, x_3) = \underbrace{(x_1 \vee x_2)}_{clause\ a} \wedge \underbrace{(x_1 \vee \neg x_3)}_{b}. \tag{4.1}$$

To study random k-SAT formulas it is very useful to represent them as factor graphs and study their corresponding random graph ensembles, since these have been studied thoroughly in random graph theory. A nice introduction to the subject is the book of (Bollobas, 2001).

Definition 26 (Mezard and Montanari, 2009)**.** *For a given k-SAT formula, we associate a factor graph:* $\mathbf{F} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, *where* $\mathcal{V} \equiv [N]$, $\mathcal{F} \equiv [M]$ *are vertices (or nodes) representing the variable and function nodes respectively, and* $\mathcal{E}$ *are the edges relating elements across the two sets.*

To simplify notation, we use $i, j \ldots$ and $x_i, x_j \ldots$ interchangeably to denote variable nodes, and $a.b \ldots$ to denote function nodes. Consider the factor graph in the figure below, we draw a full edge between a variable vertex i and a clause vertex a whenever $x_i \in \partial a$, and a dashed edge is drawn whenever $\neg x_i \in \partial a$.

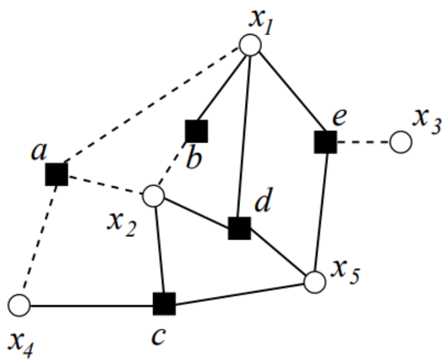


Fig. 10.1 Factor graph representation of the formula $(\overline{x}_1 \vee \overline{x}_2 \vee \overline{x}_4)\wedge(x_1 \vee \overline{x}_2)$ $\wedge(x_2 \vee x_4 \vee x_5)$ $\wedge(x_1 \vee x_2 \vee x_5)\wedge(x_1 \vee \overline{x}_3 \vee x_5)$.

Figure 4.1: Mézard, Montanari 2009

Definition 27 (Mezard and Montanari, 2009)**.** *A random k-SAT instance is generated as follows; let the number of clauses* M *be a Poisson random variable with parameter* $\pi \equiv N\alpha$, *where* α *is the clause density. Conditioned on* M, *each clause can be independently generated by sampling* k *elements uniformly from* $\{x_1, \neg x_1, ..x_N, \neg x_N\}$.

Given a k-SAT written in conjunctive normal form (CNF), i.e. written as a mapping $\Phi : \Sigma \mapsto \{0, 1\}$ like the example above, the *satisfiability problem* consists of determining whether the formula is satisfiable, i.e. whether $\mathcal{S} \equiv \{\underline{x} : \Phi(\underline{x} = 1)\} \neq \emptyset$, and if so, to give the set of solutions $\mathcal{S}$. Unfortunately this problem is Polynomial for $k = 2$ and NP-hard for $k \geq 3$ (Mezard and Montanari, 2009).

However, as we will see towards the end of this chapter and in more details in the next one, the $k-$SAT problem can be reformulated as a special case of the $p-$spin model, where the RSB picture can be used to shed light on the correlation structure and geometry of the solution space of random k-SAT instances, thus inspiring some clever message passing algorithms which prove to be able to solve large k-SAT instances surprisingly fast.

4.1.1 Local convergence to a tree

To be more precise about the randomness of the k-SAT instances, we define a generative model of their factor graph distribution, following chapter 9 of (Mezard and Montanari, 2009). Given, the number of variable nodes and a density parameter α, we generate a number of clauses according to a Poisson r.v. with intensity parameter $\pi \equiv \alpha N$. The next step is then to determine the connectivity of the graph, to this end, it is useful to describe two equivalent random graph distributions; $\mathcal{G}_N(k, M)$ and $\mathcal{D}_N(\mathcal{P}, \mathcal{Q})$;

a. Given a fixed number of variables N and a realization of the random number of clauses $M \sim Poiss(\alpha)$, for every clause $a \in [M]$, we draw a fixed number of variables $|\partial a| = k$ uniformly at random, from the set of possible $\binom{2N}{k}$ $k-$tuples in $\{x_1, \neg x_1 \ldots x_N, \neg x_N\}$, and connect their corresponding variable nodes to the function node representing clause a, this defines the random factor graph ensemble $\mathcal{G}_N(k, M)$.

b. The second ensemble generalizes the above by introducing the notion of a *degree profile*, describing the connectivity of the factor graph. More precisely, given a set of variable nodes $\mathcal{V}$ drawn uniformly from $\{x_i : i \in [N]\}$, let q_k, p_k be the fractions of variable and function nodes with degree k respectively, such that a fixed set of fractions $\underline{\mathcal{P}} \equiv \{\underline{p}_k : k \in \mathbf{N}\}$, $\underline{\mathcal{Q}} \equiv \{\underline{q}_k : k \in \mathbf{N}\}$ defines a distribution over variable (and function) node degree profiles: $\mathbf{Pr}[|\partial x_i| = k] = \underline{\mathcal{P}}(deg_{x_i} = k) = \underline{p}_k$ and $\mathbf{Pr}[|\partial a| = k] = \underline{\mathcal{Q}}(deg_a = k) = \underline{q}_k$. This naturally leads to the definition of a *degree constrained factor graph* $\mathcal{D}_N(\underline{\mathcal{P}}, \underline{\mathcal{Q}})$; a random factor graph, with fixed degree profile $\underline{\mathcal{P}}, \underline{\mathcal{Q}}$. Note that we can give more weight to certain degrees, for example; to generate an *l-regular* k-SAT instance, i.e. one where each variable appears in exactly l clauses: $|\partial i| = l, \; \forall i \in [N]$, we fix $p_l = 1, p_{j\neq l} = 0$ and $q_k = 1, q_{j\neq k} = 0$.

Both of these ensemble have been studied thoroughly in the mathematical literature (Bender and Canfield, 1978), more specifically their limiting local tree-like structure has been developed in the *theory of local weak convergence* of (Aldous and Steele, 2003).

Consider a randomly chosen edge in a factor graph, $(i, a) \overset{unif.}{\sim} \mathcal{E}$, we can describe the degree distribution of both of its ends using the *generating functions* $\mathcal{P}(x) \equiv \sum_{k\geq 0} p_k x^k$ and $\mathcal{Q}(x) \equiv \sum_{k\geq 0} q_k x^k$, this approach was developed by (Flajolet and Sedgewick, 2008) to enumerate trees but we will only make superficial use of it, to show the tree like structure of the limiting object of $\mathcal{D}_N(\mathcal{P}, \mathcal{Q})$.

More precisely, let $\mathcal{I}(x) \equiv \mathcal{P}^{'}(x)/\mathcal{P}^{'}(1),\ \mathcal{J}(x) \equiv \mathcal{Q}^{'}(x)/\mathcal{Q}^{'}(1),$ we have

$$\mathbf{E}_{\sim\mathcal{P}}[deg(x_i)] = \sum_{k\geq 0} k\ p_k\ 1^k = \mathcal{P}^{'}(1), \tag{4.2}$$

the same goes for $\mathcal{Q}^{'}(1)$. Moreover, we can see that the set $\mathcal{I} \equiv \{k.p_k/\mathcal{P}^{'}(1) :\ k \geq 0\}$ $(resp.\ \mathcal{J} \equiv \{k.q_k/\mathcal{Q}^{'}(1) :\ k \geq 0\})$ has positive elements that sum to one, thus defining a probability distribution over the degree of the variable (resp. function) node at the end of a randomly chosen edge.

To be more specific about the limiting object, consider the following random tree ensemble:

Definition 28 (Mezard and Montanari, 2009). *Given $\mathcal{V}, \mathcal{F}$ the set of variable and function nodes respectively, we define the random tree ensemble $\mathcal{T}_r(\mathcal{P}, \mathcal{Q})$ as follows. Let the distance between two variables nodes be the number of function nodes along the shortest path between the two. Given this definition, let $\mathcal{B}_r(x_i)$ be the ball of radius r around x_i, i.e. the set of variable nodes whose distance from x_i is less or equal to r. The generative model for the tree ensemble is recursive, starting from the root $r = 0$ consisting of a single variable node, to generate $\mathcal{T}_r(\mathcal{P}, \mathcal{Q})$, for each $(r-1)^{th}$ generation variable node x_{r_k}, we draw independently its degree ∂x_{r_k} from the edge function connectivity distribution $\mathrm{j} \sim \mathcal{J}$, and connect it to $\underline{\mathrm{j}}$ added function nodes below it. Then, for each of the $\underline{\mathrm{j}}$ function nodes, we independently draw their degree from the edge variable connectivity distribution $\mathrm{i} \sim \mathcal{I}$, and connect each to their descendent $\underline{\mathrm{i}}$ variable nodes, thus defining the r^{th} generation of the random tree.*

We then have the following result:

Theorem 11 (Mezard Montanari, 2009). *Consider a random factor graph with fixed degree profile $\mathbf{F} \sim \mathcal{D}_N(\underline{\mathcal{P}}, \underline{\mathcal{Q}})$ and let x_i be a uniformly drawn variable node from $\mathcal{V}$, then $\mathcal{B}_r(x_i) \underset{N\to\infty}{\longrightarrow} \mathcal{T}_r(\underline{\mathcal{P}}, \underline{\mathcal{Q}})$ in distribution.*

4.1.2 Message passing on trees is exact

Many complex systems, such as social networks, biological neurons or genomes, depend on a large number of variables that depend on each other in complex ways, these dependencies are often known to domain experts, e.g. biologists. Therefore, in order to model these systems accurately, it is very useful to encode the experts beliefs about the dependency structure of these high dimensional probability distributions. Throughout this section, we will write $x_i \perp x_j$ to mean that the two random variables are independent.

A very useful way to encode these dependencies is through *Markov Networks*:

Definition 29. *A Markov Network is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices representing the variables $\{x_i\}_{i\in\mathcal{V}}$ of a high dimensional probability distribution that we denote by $\nu(x_1, \ldots x_{|\mathcal{V}|})$, and $\mathcal{E}$ is the set of edges encoding the dependency structure of ν, such that $x_i \perp x_j \mid \mathcal{M}_{\mathcal{G}}(i)$ where $\mathcal{M}_{\mathcal{G}}(i) \equiv \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$, the Markov blanket of x_i.*

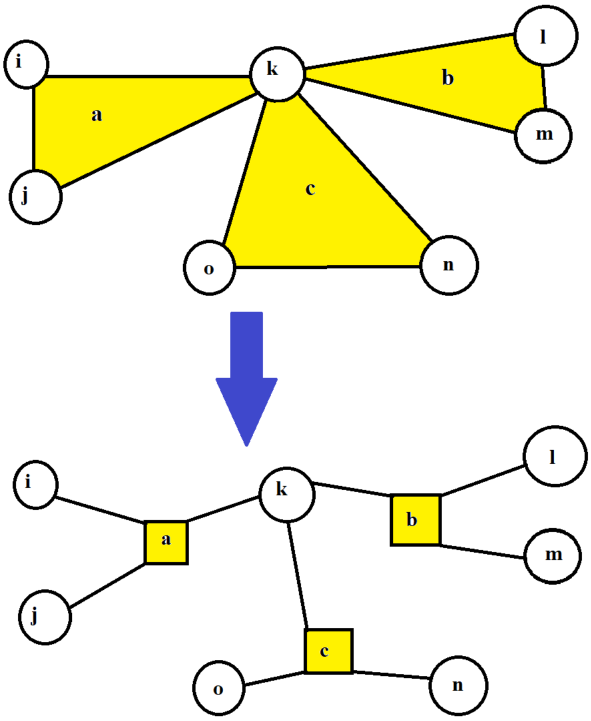


Figure 4.2: Transforming Markov Networks into factor graphs.

We say that a set of variables $\mathcal{S}_{\mathcal{G}}(i,j)$ *separates* i from j, if by deleting all variables in $\mathcal{S}_{\mathcal{G}}(i,j)$ (and all edges attached to said variables), there does not remain any path from i to j. Markov blankets are a special case of "separation", more generally, in a Markov Network, any two variables $i, j \in \mathcal{V}$ are independent conditional on any set of nodes that separates them: $x_i \perp x_j \mid \mathcal{S}_{\mathcal{G}}(i,j)$.

A recurring feature in biological systems and genotypes, is that certain subsets of $\mathcal{V}$ tend to have variables that are highly correlated with those belonging to the same subset, but largely independent with the others, in such cases it is very useful to use a factor graph representation. Fortunately, Graph theory does provide the right language for these notions, so the transition is very natural. A *clique* $\mathcal{C}$ is a connected component of a graph $\mathcal{C} \subseteq \mathcal{V}$ such that $\forall\, k, l \in \mathcal{C}$, $(i,j) \in \mathcal{E}$.

To transform a Markov network into a factor graph, for each clique $\mathcal{C}_a \subseteq \mathcal{V}$, we add a function node. Conventionally, we should also add a factor for each leaf, however for simplicity and all practical purposes relevant to constraint satisfaction type problems, we can do without.

Once the dependency structure is assumed, we assign a compatibility measure between states or *potentials*, which are essentially unnormalized probability measures over possible states within cliques $\psi_a : \mathcal{X}^a \mapsto \mathbf{R}^+$ where $\mathcal{V} = \bigsqcup_{a \in [M]} \mathcal{C}_a$.

For example, if a certain function node a encodes the phenotype of a hidden trait and i, j are the binary variables indicating the presence of some genetic information correlated with the expression of said phenotype and likewise for another phenotype b and its observed variables, given some expert information, e.g. the frequency of finding the hidden trait a or b in individuals, given the presence of genes i, j, k ;

(i,j,k)	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$	$(1,1,0)$	$(0,1,1)$	$(1,0,1)$	$(1,1,1)$
$\psi_a(i,j)$	1	2	2	2	3	3	3	5
$\psi_b(j,k)$	4	3	3	3	2	2	2	0

Table 4.1: Potentials as unnormalized probabilities

If the distance between any two variables belonging to different cliques is larger than one, i.e. more than one function node along the shortest path, then the two variables are independent conditioned on their respective clique, which allows to factorize $\nu(x_1, \dots x_{|\mathcal{V}|})$ into

$$\nu(\underline{x}) \propto \prod_{a\in[M]} \psi_a(\underline{x}_{\partial a}), \tag{4.3}$$

where M is the number of cliques in $\mathcal{G}$ and $\{\psi_a\}_{a\in[M]}$ are the unnormalized marginal probabilities of the set of variables $x_{\partial a}$.

The next step then is to estimate marginal probabilities of individual variables $\nu_i(x_i)$, but since the potential need not to sum to one, the target distribution ν is determined only up to the normalization constant $\mathcal{Z} = \sum_{(x_1,\dots,x_N)} \prod_{a\in[M]} \psi_a(x_{\partial a})$, thus to be able to make inference we need to normalize it.

Suppose all variables share the same support $\mathcal{X}$, such that to compute the normalization constant $\mathcal{Z} \equiv \sum_{\underline{x}_1 \cdots \underline{x}_N} \prod_{a\in[M]} \psi_a(\underline{x}_{\partial a})$ where $N \equiv |\mathcal{V}|$, and $M = |\{\mathcal{C} : \mathcal{C} \subset \mathcal{V}\}|$ is the number of non overlapping cliques spanning all the vertices of $\mathcal{G}$, we need to sum over an exponential number of terms: $|\mathcal{X}|^N$.

If the resulting factor graph is a tree, the following identity holds

$$\mathcal{Z} = \sum_{\underline{x}_1 \cdots \underline{x}_N} \prod_{a\in[M]} \nu(\underline{x}_{\partial a}) = \prod_{a\in[M]} \Big(\sum_{\underline{x}_{\partial a}} \nu(\underline{x}_{\partial a}) \Big) \tag{4.4}$$

such that we can start summing variables at the leaves $x_{\partial a}$ and taking the product, all up to an arbitrary root.

We then have the following result that was left as an exercise in ch 14 of (Mezard and Montanari, 2009), of which we sketch the proof on the simple case below to illustrate it.

Lemma 4. *Given a tree factor graph* $\mathbf{F} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, *suppose that all variables* $\{x_i\}_{i\in\mathcal{V}}$ *share the same support* $\mathcal{X}$, *and let* $\mathcal{L}_\mathcal{F} \equiv \{f \in \mathcal{F} : \ |\partial f| = 1\}$ *and* $\mathcal{L}_\mathcal{V} \equiv \{x_i \in \mathcal{V} : \exists a \in \partial i, \ a \in \mathcal{L}_\mathcal{F}\}$ *denote the leaf function-nodes and leaf variable-nodes respectively. The normalization constant* $\mathcal{Z}$ *then satisfies:*

$$\mathcal{Z} = \sum_{\underline{x}_l \in \mathcal{X}} \psi_f(\underline{x}_l) \mathcal{Z}_{l\to f}(\underline{x}_l), \tag{4.5}$$

for all $l \in \mathcal{L}_\mathcal{V}, \ f \in \partial l,$ *where* $\mathcal{Z}_{l\to f}(x_l) = \prod_{b\in\partial l\setminus f} \Big(\sum_{\underline{x}_{\partial b \setminus l}} \psi_b(\underline{x}_{\partial b}) \prod_{k\in\partial b\setminus l} \mathcal{Z}_{k\to b}(x_k) \Big)$.

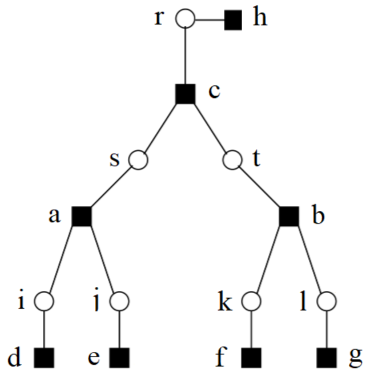


Figure 4.3: Factor graph adapted from (Mezard and Montanari, 2009)

Proof. Consider the factor graph in the figure above, where r is chosen to be the root, such that the factor graph is a 2-generation tree. The clever way to compute $\mathcal{Z}$ is to do so in a depth first search fashion. We have $\mathcal{L}_{\mathcal{V}} = \{r, i, j, k, l\}$, since r is the root, it is also the base case in the recursive computation that follows from the definition of $\mathcal{Z}_{r\to h}(x_r)$, we then have:

$$
\begin{aligned}
\mathcal{Z}_{r\to h}(x_r) &= \sum_{\underline{x}_s,\underline{x}_t} \psi_c(\underline{x}_s, \underline{x}_t, x_r)\ \mathcal{Z}_{s\to c}(\underline{x}_s)\mathcal{Z}_{t\to c}(\underline{x}_t) \\
&= \sum_{\underline{x}_s,\underline{x}_t} \Bigg\{ \psi_c(\underline{x}_s, \underline{x}_t, x_r)\ . \Bigg(\sum_{\underline{x}_i,\underline{x}_j} \psi_a(\underline{x}_i, \underline{x}_j, \underline{x}_s)\ \mathcal{Z}_{i\to a}(\underline{x}_i)\ \mathcal{Z}_{j\to a}(\underline{x}_j) \Bigg) \\
&\qquad . \Big(\sum_{\underline{x}_k,\underline{x}_l} \psi_b(\underline{x}_k, \underline{x}_l, \underline{x}_t)\ \mathcal{Z}_{k\to b}(\underline{x}_k)\ \mathcal{Z}_{l\to b}(\underline{x}_l) \Big) \Bigg\} \\
&= \sum_{\underline{x}_s,\underline{x}_t} \Bigg\{ \psi_c(\underline{x}_s, \underline{x}_t, x_r)\ . \Bigg(\sum_{\underline{x}_i,\underline{x}_j} \psi_a(\underline{x}_i, \underline{x}_j, \underline{x}_s)\ \psi_d(\underline{x}_i)\ \psi_e(\underline{x}_j) \Bigg) \\
&\qquad . \Big(\sum_{\underline{x}_k,\underline{x}_l} \psi_b(\underline{x}_k, \underline{x}_l, \underline{x}_t)\ \psi_f(\underline{x}_k)\ \psi_g(\underline{x}_l) \Big) \Bigg\},
\end{aligned}
$$

hence

$$
\begin{aligned}
\sum_{\underline{x}_r\in\mathcal{X}} \psi_h(\underline{x}_r)\mathcal{Z}_{r\to h}(\underline{x}_r) &= \sum_{\underline{x}_r}\sum_{\underline{x}_s,\underline{x}_t}\sum_{\underline{x}_i,\underline{x}_j}\sum_{\underline{x}_k,\underline{x}_l} \Bigg(\psi_h(\underline{x}_r)\ \psi_c(\underline{x}_s, \underline{x}_t, x_r)\ \psi_a(\underline{x}_i, \underline{x}_j, \underline{x}_s) \\
&\qquad \psi_b(\underline{x}_k, \underline{x}_l, \underline{x}_t)\psi_d(\underline{x}_i)\ \psi_e(\underline{x}_j)\psi_f(\underline{x}_k)\ \psi_g(\underline{x}_l) \Bigg) \\
&= \sum_{\underline{x}_r,\underline{x}_s,\underline{x}_t,\underline{x}_i,\underline{x}_j,\underline{x}_k,\underline{x}_l} \ \prod_{c\in\{a,b,c,d,e,f,g,h\}} \psi_c(\underline{x}_{\partial c}) \\
&= \mathcal{Z}. \quad \blacksquare
\end{aligned}
$$

A clearer description of the iterative marginalization of the probability distribution $\nu(x_1, \ldots, x_{|\mathcal{V}|})$ associated with $\mathbf{F}$, is described by the *Belief Propagation* (BP) equations:

Definition 30 (Mezard and Montanari, 2009)**.** *Given a factor graph* $\mathbf{F} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$*, suppose that all variables* $\{x_i\}_{i\in\mathcal{V}}$ *share the same support* $\mathcal{X}$*, the BP equations or updates are given by:*

$$\nu^t_{a\to j}(x_j) \propto \sum_{\underline{x}_{\partial a\backslash j}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\backslash j} \nu^t_{k\to a}(x_k), \tag{4.6}$$

$$\nu^{t+1}_{j\to a}(x_j) \propto \prod_{b\in\partial j\backslash a} \nu^t_{a\to j}(x_j). \tag{4.7}$$

Theorem 12 (Mezard and Montanari, 2009)**.** *Given a tree factor graph* $\mathbf{F} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$*, let the maximum distance between any two variable nodes define its diameter:* $d^* \equiv \max_{i,j\in\mathcal{V}} dist(i, j)$ *where* $dist(i, j)$ *is the number of function nodes along the shortest path in* $\mathbf{F}$ *from* i *to* j*. Then, for any initial messages in the space of probability measures in the* $|\mathcal{X}|-$*probability simplex:* $\nu^0_{i\to a}(.) \in \mathcal{M}(\mathcal{X})$*, the BP equations converge to a fixed point giving the correct marginals after at most* d^* *iterations, such that for any* $t \geq d^*$, $\nu^{t+1}_i(x_i) \equiv \prod_{a\in\partial i} \nu^t_{a\to i}(x_i) = \mu(x_i)$*, where* $\mu(x_1, \ldots, x_N) \equiv \prod_{a\in[M]} \psi_a(x_{\partial a})/\mathcal{Z}$ *is the target distribution.*

Instead of a technical proof, we will give expand on the remarks in ch 14 of (Mezard and Montanari, 2009), as to why the BP equations yield the correct marginals. The central idea is that the BP updates have a natural interpretation as *marginal probabilities in disconnected components of the original factor graph* $\mathbf{F} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, where:

1. $\nu_{j\to a}(x_j)$ is the marginal probability of x_j in the connected component that contains x_j, of a modified factor graph obtained by deleting the function node a: $\mathbf{F}^* \equiv (\mathcal{V}, \mathcal{F}\backslash\{a\}, \mathcal{E}\backslash\{(a, i) : \forall i \in \partial a\})$.
2. $\nu_{a\to j}(x_j)$ is the marginal probability of x_j in the connected component obtained by deleting all function nodes in $\partial j\backslash\{a\}$ in the original factor graph $\mathbf{F}$.

1. The interpretation of $\nu_{j\to a}$:
More precisely, it is easy to see in the computation of $\mathcal{Z}_{r\to h}(x_r)$ in the lemma above, that $\sum_{\underline{x}_r\in\mathcal{X}} \mathcal{Z}_{r\to h}(\underline{x}_r)$ yields the normalization constant $\mathcal{Z}$ of a modified factor graph where the function node h has been deleted: $\mathbf{F}^* \equiv (\mathcal{V}, \mathcal{F}\backslash\{h\}, \mathcal{E}\backslash\{(r, h)\})$.

Moreover, as pointed out in ch 14 of (Mezard and Montanari, 2009), it can be shown that $\nu_{j\to a}(x_j) \equiv \mathcal{Z}_{j\to a}(x_j)/\sum_{\underline{x}_j} \mathcal{Z}_{j\to a}(\underline{x}_j)$. When x_j is not a leaf node $x_j \notin \mathcal{L}_{\mathcal{V}}$, deleting a cuts the factor graph into $|\partial a|$ connected components, and $\nu_{j\to a}$ is the marginal probability of x_j in the connected component containing it (x_j).

2. The interpretation of $\nu_{a\to j}$:
As for $\nu_{a\to j}$, the interpretation of the messages as marginals, is a little more subtle. Consider the tree factor graph illustrated in the figure below, since $\mathbf{F}$ is a tree, it contains no loops and therefore $\forall\ a \in \mathcal{F}$, $\forall\ k, l \in \partial a$, $\{\partial k\backslash a\} \cap \{\partial l\backslash a\}$. It is easy to see that by deleting any function node, e.g. $\{a\}$, $\mathbf{F}$ becomes a collection of $|\partial a|$ non-connected trees each rooted at one of the variable nodes in ∂a.

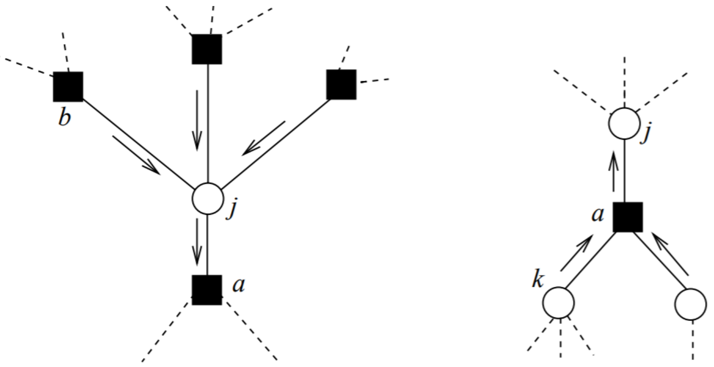


Figure 4.4: Messages from (Mezard and Montranari, 2009)

Hence, in the modified factor graph $\mathbf{F}^* \equiv \bigsqcup_{k \in \partial a \setminus j} \mathcal{B}_k, \ \forall\, k, l \in \partial a, \ k \perp l$. Moreover, in each of these branches $\{\mathcal{B}_k\}_{:k \in \partial a \setminus j}$, $\nu_{a \to k}(x_k) = \mu^k_{\mathcal{B}_k}(x_k)$ the marginal probability of x_k in $\mathcal{B}_k$. Therefore, the product $\psi_a(\underline{x}_{\partial a \setminus j}, x_j) \prod_{k \in \partial a \setminus j} \nu_{a \to k}(\underline{x}_k) = \mu_{\mathbf{F}'}(\underline{x}_{\partial a \setminus j}, x_j)$ gives the marginal probability of the set of variables in $\partial a \setminus j$ in the tree obtained by connecting the $|\partial a| - 1$ branches rooted at each of the variables in $\partial a \setminus \{j\}$ to the function node a, thereby obtaining a third factor graph $\mathbf{F}^{'} \equiv \bigsqcup_{k \in \partial a \setminus j} \mathcal{B}_k \bigcup \{(j), (a), (j, a)\}$.

Thus, by marginalizing out all variables of $\mu_{\mathbf{F}'}(\underline{x}_{\partial a})$ except x_j we get the marginal probability of x_j in $\mathbf{F}^{'}$; $\nu_{a \to j} \equiv \sum_{\underline{x}_{\partial a \setminus j}} \psi_a(\underline{x}_{\partial a \setminus j}, x_j) \prod_{k \in \partial a \setminus j} \nu_{a \to k}(\underline{x}_k)$. In other words, $\nu_{a \to j}$ is the marginal law of x_j in the connected component obtained by deleting all function nodes in $\partial j \setminus \{a\}$ in the original factor graph $\mathbf{F}$.

4.2 The Bethe approximation

In a tree factor graph, computing the normalization constant $\mathcal{Z}$ (a task that is computationally equivalent to computing marginals (Jordan and Wainwright, 2008)), amounts to cleverly choosing the order in which we sum over the variables, just like we do in other dynamical programs s.a. the Viterbi algorithm. However, when the factor graph contains at least one loop, the recursive computation of $\sum_{\underline{x}_l \in \mathcal{X}} \ \psi_f(\underline{x}_l) \mathcal{Z}_{l \to f}(\underline{x}_l)$ for $l \in \mathcal{L}_{\mathcal{V}}, \ f \in \partial l,$ no longer yields $\mathcal{Z}$, as can be seen in the example of the tree factor graph above when adding the edge (s, b).

Hence the need for a formalism in which we can assess the exactness of the BP fixed points as approximations of $\mu(x) \propto \prod_{a \in [M]} \psi_a(x_{\partial a})$, when the factor graph is no longer a tree. The *cavity method*, which originated from the theory of spin glasses as an attempt to put the replica method on a firm probabilistic footing, does just that.

Definition 31 (Mezard and Montanari, 2009). *Let $U \subset \mathcal{V}$ be a connected subset of variable nodes, the cavity: $\mathcal{U} \equiv (U, \mathcal{F}_U, \mathcal{E}_U)$ is its induced subgraph, whose factor nodes are precisely those whose all adjacent variables are in U, i.e. $\mathcal{F}_U \equiv \{a \in \mathcal{F} : \partial a \subset U\}$, and $\mathcal{E}_U \equiv \{(a, j) \in \mathcal{E} : a \in \mathcal{F}_U, j \in \mathcal{U}\}$. Moreover, we define the boundary of the cavity as $\partial\mathcal{U} \equiv \{(b, k) \in \mathcal{E} \setminus \mathcal{E}_U : k \in U, \ b \in \mathcal{F} \setminus \mathcal{F}_U\}$.*

Roughly speaking, the main idea goes as follows. Given a factor graph, consider some cavity $\mathcal{U} \equiv (U, \mathcal{F}_U, \mathcal{E}_U)$, and suppose we want to approximate the marginal probability of the variables in U from the joint target distribution associated with the original factor graph $\mathbf{F}$. When $\mathbf{F}$ is a tree, the subgraph obtained by deleting the cavity would be a collection of non-connected trees (or branches) $\mathbf{F} \backslash \mathcal{U} = \bigsqcup_{k:(a,k)\in\partial\mathcal{U}} \mathcal{B}_k$, each rooted at the variable node at the end of every edge around the cavity's boundary $\partial\mathcal{U}$. In (Mezard and Montanari, 2009) the authors observe that the marginal probability of the variables in U satisfies:

$$\hat{\mu}_U(x_U) \propto \prod_{a\in\mathcal{F}_U} \psi_a(x_{\partial a}) \prod_{(j,a)\in\partial\mathcal{U}} \nu_{a\to j}(x_j). \tag{4.8}$$

Recalling (**2. The interpretation of** $v_{a\to j}$), it is easy to see that the incoming messages at the boundary $\{v_{a\to i}(x_i) : (a,i) \in \partial\mathcal{U}\}$ yield the marginal probability of the boundary variable nodes in a modified factor graph where the cavity has been erased. Therefore, we approximate the joint probability of the variables in U by taking a product of the factors of its induced subgraph: $\{\psi_a : a \in \mathcal{F}_U\}$ and multiply these by the incoming boundary messages as an approximation of the joint distribution of boundary variables, which results in the *Bethe equation* (or *Bethe measure*):

Definition 32 (Mezard and Montanari, 2009)**.** *μ is a Bethe measure if there exists a set of messages $\{v_{a\to i} : (a,i) \in \mathcal{E}\}$ such that, for 'almost all' of the 'finite size' cavities U, we have:*

$$\mu_U(x_U) \propto \prod_{a\in F_U} \psi_a(x_{\partial a}) \prod_{(i,a)\in} v_{a\to i}(x_i) + err(x_U),$$

where $err(x_U)$ is a small error term.

For a more formal definition of the Bethe approximation, we refer the reader to (Dembo and Montanari, 2010) page 53.

The condition that a set of BP messages satisfies the Bethe equations for all finite cavities $U \in \mathcal{V}$ in the large N limit, enforces very strong constraints on the messages, leading to them being solution of the BP equations. The following result illustrating this, was left as an exercise in (Mezard and Montanari, 2009).

Theorem 13. *Consider 2 cavities $U, W \subseteq \mathcal{V}$ and let $\mathcal{F}_W = \mathcal{F}_U \cup \{a\}$ and $U \cap \partial a = \{(j,a)\}$, then the consistency of the Bethe measure for these two cavities implies the BP equation for $\nu_{a\to j}(x_j)$ for any $j \in \partial a \cap (\mathcal{V}\backslash U)$.*

Proof. Since $U, W \subseteq \mathcal{V}$, $\mathcal{F}_W = \mathcal{F}_U \cup \{a\}$ and $U \cap \partial a = \{(j,a)\}$, we have:

$$\partial\mathcal{W} = \Big\{\partial\mathcal{U}\backslash\{(j,a)\}\Big\} \bigcup \Big\{(f,k) : f \in \partial k\backslash\{a\} \text{ for all } k \in \overbrace{\partial a \cap (\mathcal{V}\backslash U)}^{=W\backslash U}\Big\} \tag{4.9}$$

Following the definition of a Bethe measure we have:

$$\hat{\mu}_W(x_W) \propto \psi_a(x_{\partial a}) \prod_{c\in F_U} \psi_c(x_{\partial c}) \prod_{(i,b)\in\partial W} v_{b\to i}(x_i), \tag{4.10}$$

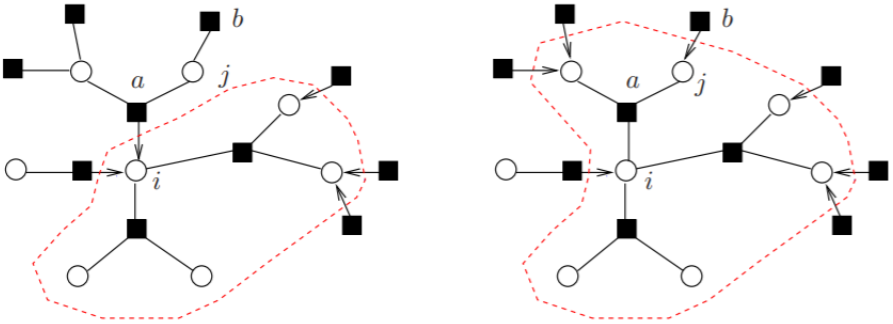


Figure 4.5: Two cavities, from (Mezard and Montanari, 2009)

and

$$\hat{\mu}_U(x_U) \propto \frac{v_{a\to j}(x_j)}{\prod_{k\in\partial a\setminus U}\prod_{b\in\partial k\setminus a} v_{c\to k}(x_k)} \prod_{c\in F_U} \psi_c(x_{\partial c}) \prod_{(i,b)\in\partial W} v_{b\to i}(x_i) \tag{4.11}$$

$$\propto \frac{v_{a\to j}(x_j)}{\prod_{k\in\partial a\setminus U} v_{k\to a}(x_k)} \frac{\hat{\mu}_W(x_W)}{\psi_a(x_{\partial a})}. \tag{4.12}$$

And since the above approximation can only be consistent between cavities if marginalizing μ_W over the variables in $W\setminus U$ yields μ_U, we obtain:

$$\hat{\mu}_U(x_U) = \sum_{x_{W\setminus U}} \hat{\mu}_W(x_{W\setminus U}, x_U) \tag{4.13}$$

$$\propto \Big(\sum_{x_{W\setminus U}} \psi_a(x_{\partial a}) \prod_{k\in\partial a\setminus U} v_{k\to a}(x_k)\Big) \frac{\hat{\mu}_U(x_U)}{v_{a\to j}(x_j)}, \tag{4.14}$$

thus recovering the correct BP equation:

$$v_{a\to j}(x_j) \propto \sum_{x_{W\setminus U}} \psi_a(x_{\partial a}) \prod_{k\in\partial a\setminus U} v_{k\to a}(x_k). \quad \blacksquare \tag{4.15}$$

4.2.1 Belief propagation in $k-$SAT

Definition 33 (Talagrand, 2011)**.** *A diluted p-spin model is an $N-$particle spin system $(\sigma_1, \dots \sigma_N) \sim \mu_\beta$, with the Hamiltonian $\mathcal{H}(\sigma) = \sum_{i_1\dots i_p} \theta_{i_1\dots i_p}(\sigma_{i_1}\dots\sigma_{i_p})$, where $\theta_{i_1\dots i_p} : [N]^p \mapsto \{0,1\}$ is a $\{0,1\}-$valued random function deciding whether the set of spins $\{\sigma_{i_k}\}_{i_k\in[p]}$ interact or not.*

A large class of constraint satisfaction problems, compromising q-coloring, independent set, XOR-SAT and many others, that come under the umbrella term of *sparse graphical models*, can be formulated as diluted p-spin models, and have been shown to display essentially the same qualitative behaviour corresponding to the discontinuous 1RSB scenario described in the previous chapter.

In the case of the random k-SAT, the corresponding $p-$spin model has $p = k$ and $\theta_{i_1\dots i_k}$ are given by the random factors determined by the k-SAT formula. More precisely,

Definition 34 (Mezard and Montanari, 2009). *A random k-SAT instance with N variables and α as a clause density parameter, is generated as follows;*

1. *Generate the number of clauses M according to a Poisson distribution with parameter $\pi \equiv N\alpha$.*
2. *For each of the M clauses, we independently generate each clause's variables ∂a by sampling k elements uniformly at random from $\{x_1, \neg x_1, ..x_N, \neg x_N\}$, which results in the set of variables $J_{a,i} = \mathbf{1}\{x_i \text{ is negated in } a\}$, $\forall i \in \partial a, \forall a \in [M]$.*

We can then explicitly express the factors in terms of the variables $\{J_{a,i}\}$, as illustrated below:

Proposition 3 (Mezard and Montanari, 2009). *Given a realized $k-$SAT instance, let $\mathcal{S} \equiv \{\underline{x} \in \{0,1\}^N : \psi_a(\underline{x}_{\partial a}) = 1,\ \forall a \in \mathcal{F}\}$ be the set of satisfying assignments, the uniform distribution on satisfying assignments is then given by $\mu(\underline{x}) \propto \prod_{a\in\mathcal{F}} \psi_a(\underline{x}_{\partial a})$ with the normalization constant being equal to the number of satisfying assignments: $|\mathcal{S}|$.*

Proof. Any unsatisfying assignment $\underline{x}_{\partial a}$ for clause a must have $x_i \neq J_{a,i}$, hence $\{\partial a :\ a \in \mathcal{F}\}$, $\psi_a : \underline{x}_{\partial a} \mapsto 1 - \prod_{i\in\partial a} \delta_{\underline{x}_i, J_{a,i}}$, $\forall\ a \in \mathcal{F}$. ■

Proposition 4 (Mezard and Montanari, 2009). *Given a k-SAT instance and its set of negation constants $\{J_{a,i}\}$, the BP updates (as previously defined) satisfy:*

$$\nu_{i\to a}(x_i = J_{a,i}) = \zeta_{ia}, \quad \nu_{a\to i}(x_i = J_{a,i}) \equiv \hat{\zeta}_{ai}, \tag{4.16}$$

where

$$\zeta_{ia} = \frac{\left(\prod_{b\in\partial_{\sim a}} \hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\not\sim a}} (1-\hat{\zeta}_{bi})\right)}{\left(\prod_{b\in\partial_{\sim a}} \hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\not\sim a}} (1-\hat{\zeta}_{bi})\right) + \left(\prod_{b\in\partial_{\not\sim a}} \hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\sim a}} (1-\hat{\zeta}_{bi})\right)},$$
$$\hat{\zeta}_{ai} = \frac{1 - \prod_{j\in\partial a\setminus i} \zeta_{ja}}{2 - \prod_{j\in\partial a\setminus i} \zeta_{ja}}.$$

Proof. If an assignment to variable x_i satisfies a , it still might not satisfy other clauses. More precisely, let $\partial^+ i \equiv \{a \in \mathcal{F} : J_{a,i} = 0\}$, $\partial^- i \equiv \partial i\setminus\ \partial^+ i$, $\partial_{\sim a} i \equiv \{b \in \partial i\setminus\ a :\ J_{a,i} = J_{b,i}\}$, $\partial_{\not\sim a} i \equiv \{b \in \partial i\setminus\ a :\ J_{a,i} \neq J_{b,i}\}$, it is easy then to see that if $J_{a,i} = 0$ then $\partial_{\sim a} i = \partial^+ i\setminus\ a$, $\partial_{\not\sim a} i = \partial^- i$, and if $J_{a,i} = 1$ then $\partial_{\sim a} i = \partial^- i\setminus\ a$, $\partial_{\not\sim a} i = \partial^+ i$.

Since the support of the BP messages are binary $x_i \in \{0,1\}$, each message is can parameterized by a single value;

$$\nu_{i\to a}(x_i = J_{a,i}) \equiv \zeta_{ia}, \quad \text{hence} \quad \nu_{i\to a}(x_i = 1 - J_{a,i}) = 1 - \zeta_{ia}, \tag{4.17}$$
$$\nu_{a\to i}(x_i = J_{a,i}) \equiv \hat{\zeta}_{ai}, \quad \nu_{a\to i}(x_i = 1 - J_{a,i}) = 1 - \hat{\zeta}_{ai}. \tag{4.18}$$

Moreover, since $\nu_{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a} \nu_{b\to i}(x_i)$ up to $\mathcal{Z}_{i\to a} \equiv \sum_{\underline{x}_i\in\{J_{a,i}, 1-J_{a,i}\}} \nu_{i\to a}(\underline{x}_i) =$

$\left(\prod_{b\in\partial_{\sim a}}\hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\not\sim a}}(1-\hat{\zeta}_{bi})\right) + \left(\prod_{b\in\partial_{\not\sim a}}\hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\sim a}}(1-\hat{\zeta}_{bi})\right)$, it follows that:

$$\zeta_{ia} = \frac{\left(\prod_{b\in\partial_{\sim a}}\hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\not\sim a}}(1-\hat{\zeta}_{bi})\right)}{\left(\prod_{b\in\partial_{\sim a}}\hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\not\sim a}}(1-\hat{\zeta}_{bi})\right) + \left(\prod_{b\in\partial_{\not\sim a}}\hat{\zeta}_{bi}\right)\left(\prod_{b\in\partial_{\sim a}}(1-\hat{\zeta}_{bi})\right)}, \quad (4.19)$$

and $\nu_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)$. Thus, for $\underline{x}_i = 1 - J_{a,i}$, we get $\psi_a(\underline{x}_{\partial a}) = 1 - \prod_{i\in\partial a}\delta_{\underline{x}_i, J_{a.i}} = 1$ independently of $\underline{x}_{\partial a\setminus i}$, with some straightforward algebra, we get

$$\hat{\zeta}_{ai} = \frac{1 - \prod_{j\in\partial a\setminus i}\zeta_{ja}}{2 - \prod_{j\in\partial a\setminus i}\zeta_{ja}}, \quad (4.20)$$

As for the leaf variables, where $\forall i \in \mathcal{L}_{\mathcal{V}}$, $|\partial i| = 1$, the product of zero terms is taken to be equal to one. ■

4.3 Evolution of the uniform measure on satisfying assignments

Recall that the BP equations rest on the implicit assumption that short range correlations vanish in the large N limit, such that any two close variables $\forall\, i, j \in \partial a$ for some factor $a \in \mathcal{F}$ become locally separated by deleting a, such that the marginal of x_j in the factor graph induced by deleting a is a product measure on branches $\nu_{j\to a}(x_j) \propto \prod_{b\in\partial j\setminus a}$, as discussed in the previous section.

4.3.1 The correlation-length viewpoint

We distinguish between two different types of correlations, each characterizing the clustering and condensation thresholds. In the replica symmetric regime, both short and longer range correlation are absent, beyond the clustering transition the shorter range one no longer decays in the thermodynamic limit while the other one does, hence failure of BP, and finally above the condensation threshold both of them are present. We begin by describing the long range type.

Recall the fluctuation-dissipation relation in the spin glass discussion, relating sensitivity w.r.t. perturbations and the correlation between spins in the unperturbed system given by the spin glass susceptibility:

$$\chi^{SG} = \frac{\beta^2}{N}\sum_{ij}[\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle]^2 = \frac{d}{dB_i}\langle\sigma_j\rangle \quad (4.21)$$

As previously discussed, the emergence of long range correlations characteristic of the spin glass phase can be detected from a divergence in the susceptibility, i.e. when $\lim_{N\to\infty}\chi^{SG} = \infty$ the system transitions to the glassy phase. More generally, in a (diluted) $p-$spin model, we distinguish the following two glassy transitions:

Definition 35 (Mezard and Montanari, 2009). *Given a $k-SAT$ instance, we say that the uniform distribution on the set of its satisfying assignments μ is stable w.r.t. small* *perturbation if for all* $l < \infty$, $\lim_{N\to\infty} \chi^{(l)}/N = 0$, *where*

$$\chi^{(l)} \equiv \frac{1}{N^{l-1}} \sum_{i_1,\ldots,i_l \in [N]^l} ||\mu_{i_1,\ldots,i_l}(\ldots) - \mu_{i_1}(.)\ldots\mu_{i_l}(.)||, \tag{4.22}$$

is the l-point correlation function.

The second type of correlation is a weaker condition, and can be best understood as a statistical thought experiment.

Definition 36 (Mezard and Montanari, 2009)**.** *Suppose we draw a satisfying assignment* $\underline{x} \sim \mu$, *and that all of the values of its variables are revealed except for the variables in the ball of radius r centered around* x_i, *that we denote* $\mathcal{B}_l(i)$ *(and with an overline for its complement in* $\mathcal{V}$*). The uniform distribution on satisfying assignments* μ *is said to be extremal or satisfy the non-reconstructibility condition if for any variable* $x_i \in \mathcal{V}$, *we have*

$$\lim_{l\to\infty} G_i(l) = 0, \quad \textit{where} \quad G_i(l) \equiv ||\mu_{i,\bar{\mathcal{B}}_l(i)}(.,.) - \mu_i(.)\mu_{\bar{\mathcal{B}}_l(i)}(.)||. \tag{4.23}$$

is the point-to-set correlation function. If there exists a variable for which G_i *remains bounded away from zero for arbitrarily large distances,* μ *is said to be reconstructible.*

The rationale behind the reconstructability condition is that, if the correlation become arbitrarily small beyond a certain distance l, then knowing the assignment of all variables in $\bar{\mathcal{B}}_l(i) \equiv \mathcal{V} \setminus \mathcal{B}_r(i)$ does not provide any information on the value of $\underline{x}_i$, the assignment is then said to be *non-reconstructible.*

In (Montanari and Semerjian, 2006), the authors give a number of asymptotically (in l) equivalent correlation vanishing criteria, and prove rigorous inequalities relating correlation length to mixing time, some of which will be reviewed in the next chapter.

4.3.2 The complexity function viewpoint

Another way to characterize the clustering and condensation thresholds is to look at the evolution of the number of clusters containing an up to leading exponential order number of satisfying assignments. To do this, we derive a large deviation result, whose rate function is called the *complexity function* that we discuss below. Recall the large deviation principles of the spectrum of overlap values in the replica trick, the complexity-function approach follows a similar vein by focusing on the distribution of the free energy per cluster.

Consider the general case of the p-spin model where: $\mu_\beta(\sigma) \propto e^{\beta \sum_{i_1\ldots,\imath_p} J_{i_1,\ldots,i_p}\sigma_{i_1}\ldots\sigma_{i_p}}$, and let the free energy of a given cluster be $F_{\alpha_i} \equiv -\log(\mathcal{Z}_{\alpha_i})/\beta$, for $\mathcal{Z}_{\alpha_i} \equiv \sum_{\sigma\in\alpha_i} e^{-\beta\mathcal{H}(\sigma)}$. The pure state decompostition then implies $\mu(\underline{x}) = \sum_{\alpha_k : k\in[\eta]} w_k\, \mu^{\alpha_k}(\underline{x})$, where $w_k \equiv e^{\beta F_k}/\sum_{\alpha_l : l\in[\eta]} e^{\beta F_l}$ and $\mu^{\alpha_k}(\underline{x}) = \mathbf{1}\{\underline{x} \in \alpha_k\}/\mathcal{Z}_{\alpha_k}$.

To get an idea of the distribution of cluster sizes, we study the cluster-spectrum of $\mathcal{Z}_{\alpha_k}$ as $\mathcal{N}(\underline{f}) \equiv |\{k : F_k = N\underline{f} \pm \delta\}|$ for small some $\delta > 0$, and derive a large deviation principle, for $N \gg 1$, such that

$$\mathcal{Z} = \sum_{\alpha_k : k \in [\eta]} e^{-\beta F_k} = \sum_{\underline{f}} \mathcal{N}(\underline{f})\, e^{-\beta N \underline{f}} \underset{N \gg 1}{=} \int e^{N(\Sigma(\beta,\underline{f}) - \beta \underline{f})}\, d\underline{f} \doteq e^{N(\Sigma(\beta,\underline{f}^*) - \beta \underline{f}^*)}, \tag{4.24}$$

where $f^*(\beta)$ is the Legendre transform of some function $\Sigma(\beta, f)$ called *the complexity function.*

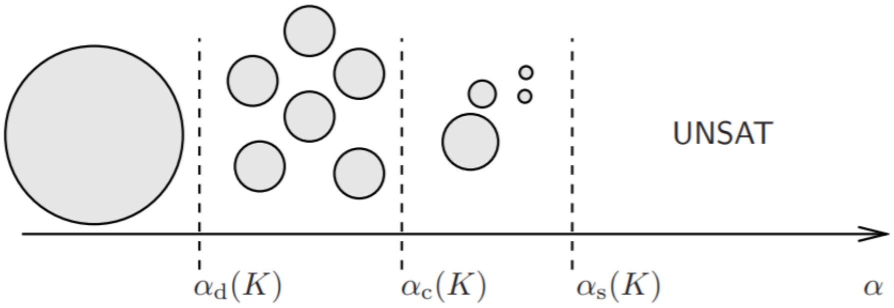


Figure 4.6: Evolution of the set of satisfying assignments (Mezard and Montranari, 2009)

Given a $k-$SAT instance, its associated factor graph and the uniform measure on its satisfying assignment, as the clause/variable density α goes up beyond the clustering or *dynamical* 1RSB threshold α_d, the space of solutions breaks into well separated clusters in Hamming space. In this regime, there exists an exponential number of quasi-solutions to the Bethe equations, each valid in a given cluster.

More precisely, let η be the number of clusters satisfying some "pure state" conditions that we describe in the next chapter, then for any $k \in [\eta]$, there exist a set of fixed point messages $\{\nu^k_{i\to a}, \nu^k_{a\to i}\}_{i\in\mathcal{V}, a\in\mathcal{F}}$ among the quasi-solutions of the Bethe equations:

$$\mu_U(\underline{x}_U) \propto \prod_{a\in F_U} \psi_a(\underline{x}_{\partial a}) \prod_{(i,a)\in} \nu^k_{a\to i}(\underline{x}_i) + err(\underline{x}_U) \quad \text{for all} \quad \underline{x} \in \alpha_k. \tag{4.25}$$

In other words, with each cluster, we can associate a Bethe measure among the BP fixed points.

The 1RSB assumption: Let the free energy given by the k^{th} BP fixed points be denoted $F_k \equiv F(\{\nu^k_{i\to a}, \nu^k_{a\to i}\}) = -\log(\mathcal{Z}^{(k)})/\beta$, where $\mathcal{Z}^{(k)}$ is the normalization constant of said fixed point. The 1RSB assumptions are then threefold:

1. There exists an exponential number of quasi-solutions to the BP equations, the number of which having $F_k \approx Ns$, is up to leading exponential order equal to $e^{N\Sigma(s)}$.

2. The uniform measure on satisfying assignments breaks into a convex combination

of *extremal* (in the sense of being short range correlated) Bethe measures:

$$\mu(\underline{x}) = \sum_{\alpha_k : kin[\eta]} w_k \, \mu^k(\underline{x}), \; w_k \equiv \frac{e^{F_k}}{\sum_{lin[\eta]} e^{F_l}}. \tag{4.26}$$

3. The number of extremal Bethe measures is equal to the number of quasi-solutions to the BP equation, and those whose free energy is approximately equal to Ns is also given by $e^{N\Sigma(s)}$.

Suppose F_k satisfies a large deviation principle: $\mathcal{N}(s) \doteq e^{N\Sigma(s,\alpha)}$, where $\mathcal{N}(s) \equiv |\{k : \; F_k = Ns \pm \delta\}|$ for some small positive δ, following along the same lines as the computation above, we have

$$\mathcal{Z} = \sum_{\alpha_k : k\in[\eta]} e^{F_k} \underset{N\gg 1}{=} \int \mathcal{N}(\underline{s}) \, e^{N\underline{s}} d\underline{s} \doteq \int \mathcal{N}(\underline{s}) \, e^{N\underline{s}} d\underline{s} \doteq e^{N[\Sigma(s^*,\alpha)+s]}.$$

Note that in the case of $k-$SAT, the clause/variable density plays the role of inverse temperature parameter in the $p-$spin case, such that the complexity function depends on both α and the value of the free energy per clause, that we denote by s.

The complexity function provides a neat way to describe the clustering and condensation regimes, here we follow the discussion in (Sun et al., 2014).

Definition 37 (Sun et al., 2014)**.** *Suppose that the number of clusters with approximately F_k free energy: $\mathcal{N}(s) \equiv |\{k : \; F_k = Ns \pm \delta\}|$ for some $\delta > 0$, satisfies $\mathcal{N}(s) \doteq e^{N\Sigma(\alpha,s)}$. The clustering threshold α_d and the condensation threshold α_c are then respectively given by:*

$$\begin{aligned} \alpha_d &\equiv \inf_{\alpha}\{\alpha : \; \exists s \in [0, \log 2], \; s.t. \; \Sigma(\alpha, s) > 0 \; \}, \\ \alpha_c &\equiv \inf_{\alpha}\{\alpha : \; s_1 \neq s_2\}, \end{aligned}$$

where $s_1 \equiv \; argmax_s\big\{\Sigma(\alpha,s) + s : \; s \in [0, \log 2] \; \big\}$ and $s_2 \equiv \; argmax_s\big\{\Sigma(\alpha,s) + s : s \in [0, \log 2]$ and $\Sigma(\alpha,s) \geq 0\big\}$.

The interpretation of these definitions become clearer if we think about the number of exponentially large clusters $\mathcal{N}(s)$ as α goes up. Qualitatively speaking, the onset of the clustering transition corresponds to the value of α at which the complexity function $\Sigma(\alpha, s)$ becomes positive, such that there appears an exponential number of clusters $\{\alpha_k\}$ each carrying $\mathcal{Z}_{\alpha_k} \doteq e^{Ns}$ solutions.

On the other hand, at the onset of the condensation transition, the number of these clusters becomes drops drastically from $\mathcal{O}(e^N)$ to $\mathcal{O}(1)$, i.e. bounded as $N \to \infty$. Again, this correspond to the definition of α_d, where the complexity function becomes zero: $\Sigma(\alpha, s) = \; 0$, and the number of clusters carrying most of the probability mass is then bounded in the thermodynamic limit.

4.4 Message passing beyond the replica symmetric regime

In replica theoretic terms, the clustering or pure state decomposition, is a statement about the distribution of the overlap. Recall the discussion in the third chapter of the multiplicity of the overlap giving clues to the geometry of the support of the Gibbs measure in different temperature regimes.

At the 1RSB level, given $x, y \overset{iid}{\sim} \mu$, the overlap $q_{x,y} \equiv \sum_{i \in [N]} x_i y_i / N$ is concentrated on three values: (a). $q_{x,y} = 1$ for the trivial all equal assignment $x = y$, (b). $q_{x,y} = q_0$ if they belong to the different clusters, and (c). $q_{x,y} = q_1 > q_0$ if x and y belong to the same one.

In other words, for any pair of assignments belonging to the same cluster, there is with high probability Nq_0 variables having the same value, these variables are called *core variables* or *frozen variables.* In each cluster there is a set of such variables that we denote with $\mathcal{C}_k \subset \mathcal{V}$ locked to some cluster-dependent value $\{\underline{z}_j : j \in C_k\}$.

Since these variable are frozen in a given cluster $x_i = \underline{z}_i \; \forall \, x \in \alpha_k$, for a Bethe measure to be valid in said cluster, the marginals of core variables should be identity functions of the frozen core assignments $\{\underline{z}_j : j \in C_k\}$, i.e. $\nu_j^k(x_j) \equiv \prod_{a \in \partial j} \nu_{a \to j}^k(x_j) = \mathbf{1}\{x_j = \underline{z}_j\}, \; \forall \, j \in \mathcal{C}_k$. This defines a natural mapping from valid Bethe measures to some cluster-dependent frozen assignments.

To recapitulate; in the clustering regime, there is an exponential number of BP fixed points, an exponential number of Bethe measures, and an exponential number of clusters. Every Bethe measure is a quasi-solution to the BP equations, while the converse is not generally true, see counterexample in p.432 of (Mezard and Montanari, 2009).

Moreover, every cluster has a set of frozen variables posing strong constraints on the set of valid marginals in the associated Bethe measure, this defines a natural mapping from clusters to Bethe measures, whose bijective nature, while desirable, is unfortunately not true, as we will see in the next section.

Survey propagation (SP) is a message-passing algorithm with the ambitious goal of approximating marginals of a probability distribution defined over the set of Bethe measures, i.e. to compute marginals of $\mathbf{P}[\{\nu_{i \to a}, \nu_{a \to i}\}_{i \in \mathcal{V}, a \in \mathcal{F}} = \{\underline{\nu^k}_{i \to a}, \underline{\nu^k}_{a \to i}\}_{i \in \mathcal{V}, a \in \mathcal{F}}]$.

Recall the extremality condition on correlation decay introduced above. Since, valid Bethe measures are good approximators of the extremal measures $\mu^{\alpha_k}()$ whose convex combination yield the uniform measure on satisfying assignments: $\mu(.) = \sum_{\alpha_k : k \in [\eta]} w_k \, \mu^{\alpha_k}(.)$, and that in the clustering regime, all clusters have equal weight $w_k = 1/\eta, \; \forall k \in [\eta]$, the set of marginals that SP computes, e.g. $\mathbf{P}[\nu_{a \to i} = . \;]$, are statistical averages (or *surveys*) over all pure states $\{\alpha_k, \; k \in [\eta]\}$ of the particular message ($\nu_{a \to i}$ in this case) whose marginal probability we are interested in.

Recall that $\nu_{a \to i}^k(x_i = 1)$ is the marginal probability (over satisfying assignments) that x_i takes the value 1, in the factor graph obtained by deleting the branch rooted at the edge (i, a). Hence, given a $k-$SAT instance with its set of negation specifications $\{J_{ai} : i \in \partial a\}_{a\in[M]}$ where $J_{ai} \equiv \mathbf{1}\{x_i \textit{ is negated in } a\}$, since frozen variable have $0-1$ marginals in the clustering regime, SP messages have a neat interpretation as probability distributions over a set of *warnings* between variable and function nodes, where $\nu^k_{a\to i}(\underline{x}_i = 1 - J_{ai}) = 1$, means that the variable x_i is forced by all clauses $b \in \partial i\backslash\, a$, to satisfy clause a, in every satisfying assignment $\underline{x} \in \alpha_k$.

This defines a probability distribution over possible warnings, each associated with a cluster, over which SP computes messages in the same way that BP does on assignments, let:

i). $Q^S_{i\to a} \equiv \mathbf{P}[\nu^k_{a\to i}(x_i = 1 - J_{ai}) = 1]$ be the uniform probability over clusters that x_i be forced by $b \in \partial i\backslash\, a$ to satisfy clause a.

ii). $Q^U_{i\to a} \equiv \mathbf{P}[\nu^k_{a\to i}(x_i = J_{ai}) = 1]$: the probability that x_i is forced by $b \in \partial i\backslash\, a$ to violate clause a.

Since for a given cluster, free variables need not take specific values, we define $Q^*_{i\to a}$ as the probability that x_i is *not* forced by $b \in \partial i\backslash\, a$ to take any particular value.

The message-passing procedure then follows the same update-rules as BP, since valid Bethe measures are extremal, i.e. are only short range correlated, frozen variables $\mathcal{C}_k$ typically have loops of length $\mathcal{O}(\log N)$, thus allowing us, much like in BP in the replica symmetric case, to write marginals as a product of incoming messages $Q^{S(U)}_{i\to a}$ from $\partial a\backslash\, i$:

$$\hat{Q}_{a\to i} \propto \prod_{j\in\partial a\backslash\, i} Q^U_{j\to a}. \tag{4.27}$$

Now let $\mathcal{W}_{\sim a}(i) \subset \partial_{\sim a} i \equiv \{b \in \partial i\backslash\, a : J_{ai} = J_{bi}\}$, $\mathcal{W}_{\nsim a}(i) \subset \partial_{\nsim a} i \equiv \{b \in \partial i\backslash\, a : J_{ai} \neq J_{bi}\}$, be the set of clauses forcing x_i to either satisfy or violate them, x_i is then forced according to a majority vote, i.e. according to $\max\{|\mathcal{W}_{\sim a}(i)|, |\mathcal{W}_{\nsim a}(i)|\}$, and is not forced in the case of equality, hence the SP(y) equations:

$$\begin{aligned}
Q^U_{i\to a} &\propto \sum_{|\mathcal{W}_{\nsim a}(i)|>|\mathcal{W}_{\sim a}(i)|} e^{-y|\mathcal{W}_{\sim a}(i)|} \prod_{b\in\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} \hat{Q}_{b\to i} \prod_{b\notin\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} (1-\hat{Q}_{b\to i}).\\
Q^S_{i\to a} &\propto \sum_{|\mathcal{W}_{\nsim a}(i)|<|\mathcal{W}_{\sim a}(i)|} e^{-y|\mathcal{W}_{\sim a}(i)|} \prod_{b\in\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} \hat{Q}_{b\to i} \prod_{b\notin\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} (1-\hat{Q}_{b\to i}).\\
Q^*_{i\to a} &\propto \sum_{|\mathcal{W}_{\nsim a}(i)|=|\mathcal{W}_{\sim a}(i)|} e^{-y|\mathcal{W}_{\sim a}(i)|} \prod_{b\in\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} \hat{Q}_{b\to i} \prod_{b\notin\mathcal{W}_{\sim a}(i)\cup\mathcal{W}_{n\sim a}(i)} (1-\hat{Q}_{b\to i}).
\end{aligned}$$

Looking back at the BP update rules,

$$\begin{aligned}
\nu^{t+1}_{j\to a}(x_j) &\propto \prod_{b\in\partial j\backslash a} \nu^t_{a\to j}(x_j)\\
\nu^t_{a\to j}(x_j) &\propto \sum_{\underline{x}_{\partial a\backslash j}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\backslash j} \nu^t_{k\to a}(x_k)
\end{aligned}$$

and reinterpreting messages as actual assignments, e.g. the probability that $\nu_{a\to i}(x_i = \underline{z}) = 1$, as the assignment of $x_i = \underline{z}$, it is easy to see that SP can be seen as computing marginals of the uniform distribution on a more general form of assignments in

$\{0,1,*\}^N$. Note that the uniformity of this target distribution follows from the fact that, in the clustering regime $w_k = 1/\eta$. Some of the details are swept under the rug, since they are not directly relevant to the rest of the discussion, for a more complete discussion see the survey of (Braunstein et al., 2005) and the later chapters (18,19,20) of (Mezard and Montanari, 2009).

In other words, if we replace the uniform distribution on satisfying assignments as a target distribution for BP, by a uniform distribution over a particular type of *generalized assignments* in $\{0,1,*\}^N$, we recover the Survey propagation algorithm.

This equivalence between BP on this special class of generalized assignments and classical SP was discovered in (Braunstein and Zecchina, 2004), and expanded upon in an important paper by (Maneva et al., 2007), where the authors provide a formal description of this particular type of objects in $\{0,1,*\}^N$ which they call true covers. Moreover, they show that Belief propagation with the uniform measure on true covers as a target distribution, is equivalent to Survey Propagation.

Given a $k-$SAT instance, with its set of negation specifications $\{J_{ai} : i \in \partial a\}_{a\in[M]}$, denote set of variables satisfying (resp. violating) clause a as $\partial^+ a \equiv \{i \in \partial a : J_{ai} = 1 - \underline{x}_i\}$, $\partial^- a \equiv \{i \in \partial a : J_{ai} = \underline{x}_i\}$.

Definition 38 (Kroc et al., 2007). *A generalized assignment $\underline{z} \in \{0,1,*\}^N$ is* ***in****valid for clause a if* ***either****:*

i). $\partial^+ a = \emptyset$

ii). ***or*** *that* $\partial a = \partial^- a \bigsqcup \{*\}$, *i.e. that all variables in ∂a do not satisfy clause a except for* ***one*** *variable set to $*$.*

If the above two conditions do not hold for any clause, the assignment is then said to be ***valid****.*

Definition 39 (Kroc et al., 2007). *x_i is* ***supported*** *by clause a, iif x_i is the only satisfying variable for a : $\partial^+ a = i$. Moreover, x_i is called a* ***constrained variable*** *if there exists at least one clause which supports it, otherwise it is unconstrained.*

Definition 40 (Kroc et al., 2007). *A valid generalized assignment where all unconstrained variables are set to $*$ is a* ***cover****. Moreover, a cover $\underline{z} \in \{0,1,*\}^N$ is called a* ***true cover*** *iif there exists at least one satisfying assignment $\underline{x} \in \mathcal{S}$, such that $\forall i \in [N]$, $\underline{z}_i \neq * \implies \underline{z}_i = \underline{x}_i$.*

.

We now motivate the above definitions:

- Condition (i). guarantees that any valid assignment is a satisfying one.
- And (ii). guarantees that there is "slack" in each clause, so that no variable with the $*$ assignment is forced to take to take a specific $0-1$ value.
- As for the second definition, constrained variables are meant to model frozen variables in clusters, and are thus forbidden to be set to $*$, in order to identify them.

- Lastly, in the event that all clauses satisfy condition (ii) rather than (i), a generalized assignment can still be a cover, while not being extendable to a satisfying assignment through fixing its $*$ variables. Such covers are called *fake covers*, they model all the properties of the cluster-representative frozen variables $\mathcal{C}_k$, without actually being associated to a cluster, hence the last condition, which guarantees that there exists at least one satisfying assignment $\underline{x} \in \mathcal{S}$, in which the true cover's constrained variables assignments are frozen in $\underline{x}_k = \underline{z}_k, \forall\ k \in \mathcal{C}_k$.

4.5 Sampling uniformly from the solutions set

Although both of Survey propagation and Belief propagation have the uniform distribution on covers (resp. on satisfying assignments) as target distributions, in practice they do not sample uniformly from the solutions set.

Classically, for general, soft-constrained strictly positive probability distributions, message passing procedures such as Belief-propagation and its loopy variants, start with random initial messages and are terminated after a given maximum running time, then used directly to draw maximum a posteriori samples.

On the other hand, hard constrained distributions, by having zero density in some portions of their support cause numerical underflow in dynamical algorithm, such as message passing procedures, and are therefore not as forgiving. Hence, they are generally cast as a backtrack search problem as in the classical DPLL decimation type procedures, where, we iteratively fix variables according to the approximate marginals computed via message passing, and backtrack as needed, until we find a satisfying assignment, or terminate.

Algorithm 2: Survey Propagation

```
Result: $\mathcal{S}ol \subset \mathcal{S}$
Initialization: $\rho = \underline{\rho}$, $\delta = \underline{\delta}$;
$\mathcal{S}ol \leftarrow \emptyset$;
for $t = 1, ...T_{max}$ do
    $\{\mu_i(0), \mu_i(1), \mu_i(*)\} \leftarrow$ $SP\_marginals$;
    Fix the first $N\rho$ variables with largest marginal biases
      $\max_{i \in [N]}\{|\mu_i(1) - \mu_i(0)|\}$ ;
    Simplify the $k-$SAT formula;
    Do a random walk on the remaining variables;
    if $\underline{x}_t$ is a satisfying assignment then
        Sol$\leftarrow$ $\mathcal{S}ol \cup \{\underline{x}_t\}$end
    end
```

In other words, for each sampled solution we run SP (or BP) once, then use the approximate marginals to fix variables starting with to the most biased ones. Note that if the message-passing procedure converges to the same fixed point, then the same solutions will be sampled over and over. In practice, this is however not a problem. In fact, for SP, we the decimation procedure is done only for a fixed subset of variables, the rationale being that these should coincide with true covers' marginals, the rest of the variables are found via a random walk.

Rigorously speaking, SP has only been shown to compute exact marginals on instances where the associated factor graph is a tree. For these instances, it can be shown that there exists only one true cover; the *trivial* cover $\underline{z} = (*, *, \dots, *)$. In practice SP is one of the only algorithms capable of finding assignments in linear time w.h.p., deep in the clustering regime, in 10^6-sized instances, whose factor graph contains many loops. The surprising success of SP was hypothesized to be a result of SP being able to approximate true cover marginals (Braunstein and Zecchina, 2004).

This hypothesis was however shortly put in question in (Maneva et al., 2007), where the authors provided experimental evidence that the solutions sampled by SP typically do not have any frozen variables, i.e. that random $k-$SAT formulas typically only have the one trivial cover. The authors then proceeded to conclude that covers may not be a successful route to put the 1RSB clustering picture on firm footing, and provided an alternative way, based on a Markov random field with a smoothing parameter which interpolates between BP and SP.

However, this conclusion was shortly disputed by (Kroc et al., 2007), where the authors attribute the failure of SP to an inherent bias in the sampling procedure away from non-trivial true covers, and show through extensive experiments that there exists a significant number of non-trivial true covers in the clustering regime. More precisely, starting from an arbitrary assignment, if we alternate between random walk and Simulated annealing moves, approximately 25% of the sampled solutions lead back, through $*$-propagation, to a non-trivial cover.

Indeed, its a well known fact, that the problem of sampling uniformly from $\mathcal{S}$ is much more challenging than that of just finding satisfying assignments. For instance, random walk procedures have been proven to find solutions in polynomial time for $2-$SAT instances, but they provably biased away with exponentially decaying probability from a portion of the state space. On the other hand, Markov Chain Monte Carlo (MCMC) procedures can be proven to sample uniformly from the support of the target distribution (provided the target is uniform).

In the next chapter, we will introduce general MCMC procedures, their provable guarantees, and relate them back to the discussion about dynamics in the second chapter. Moreover, we will survey their use in the case of the $k-$SAT problem, and the pros and cons of specific variations thereof. Finally we will present an alternative sampling algorithm, which uses the observations about true covers to overcome some of the limitations of previous work.

Chapter 5

Dynamics and sampling

Chapter organization

Given the apparent bias towards coverless-solutions in Survey propagation, we discuss an alternative method for sampling solutions uniformly from the solution set that falls under the realm of Monte Carlo Markov Chain (MCMC) methods (**5.1**). Then, we continue the discussion started in (**2.3.2**) concerning the speed of correlation decay in spin systems and relate it to the k-SAT case (**5.2**). Afterwards, we provide the complete definition of *the pure state decomposition* that we referred to (in a somewhat hand-wavy manner) in (**2.3.3**), in the case of k-SAT (**5.3**), and discuss its implications for MCMC algorithms (**5.3.1**). Subsequently, we begin a comparison between *Simulated annealing* and *random walk sampling*, by stating two results; one taking account the 1RSB prediction (**5.3.3**), and another that does not (**5.3.2**). Then, after introducing all the prerequisite notions, we delve deeper into the problem of sampling *uniformly* from the set of satisfying assignments, with a discussion of: previous work (**5.4.1**), a *speed vs. uniformity* trade-off (**5.4.2**), and a *solution-time vs. coverless-solutions* trade-off (**5.5**). Finally, building on the insight of previous work, we introduce a novel algorithm with the goal of improving the uniformity of sampling, whose main strategy consists of avoiding a relapsing phenomenon that we discuss in (**5.5**) by: *i*) identifying covers, and *ii*) taking a self-avoiding walk in the state space. After presenting the algorithm that we name *SAW-SAT*, we discuss the results of some experiments pointing to the success of strategy *ii*) and the failure of *i*). We close the chapter with a discussion of a particular interesting avenue for future work, that takes into account the symmetries of k-SAT instances.

5.1 An introduction

Facing the intractability of the partition function $\mathcal{Z}$, we can use the RSB picture of vanishing short range correlations to write variable marginals as products of BP fixed points $\mu(x_i) = \prod_{j \in \partial a} \nu^*_{a \to j}(x_i)$ in the replica symmetric case, and analogously write the marginals of the uniform distribution over BP messages as a product of fixed point SP messages, at the onset of the clustering transition, to account for the non-vanishing point set correlations. In this approach, we use the physics predictions regarding the asymptotic correlation structure of μ, which, assuming some conditions, guarantees that BP (or SP) yields good approximations for variable marginals, which is equivalent to computing $\mathcal{Z}$.

Definition 41. *A discrete-time Markov chain is a sequence of random variables: $x_1, x_2, \dots x_T$, representing the state of the Markov chain at each time-step t, where for all $1 < t \leq T$, the probability that x_t is equal to some value $\underline{x}_t$ depends only on the previous state x_{t-1}:*

$$\mathbf{P}[x_t = \underline{x}_t |\ x_1 = \underline{x}_1, x_2 = \underline{x}_2, \dots x_{t-1} = \underline{x}_{t-1}] = \mathbf{P}[x_t = \underline{x}_t |\ x_{t-1} = \underline{x}_{t-1}]. \tag{5.1}$$

We assume that the variables $\{x_t\}_{t\in[T]}$ share the same support Σ that we call the state space of the Markov chain.

In an orthogonal direction, we can altogether bypass the issue of computing $\mathcal{Z}$, by sampling according to some Markov chain, whose state space is Σ, and which is guaranteed to converge in variation distance (which we shall define below) to μ after a number of iterations $t \geq \tau_r$ with τ_r being *the relaxation time* (or time-to-equilibrium in the physics jargon). To do this, we define an energy based sampling procedure, that takes into account the hierarchical landscape of the 1RSB picture, and whose time average is equivalent to taking the expectation w.r.t μ, i.e. where: $1/T^* \sum_{t=1}^{T^*} \mathcal{O}_t = \sum_{\sigma\in\Sigma} \mathcal{O}(\sigma)\mu_\beta(\sigma)$ holds true, for some observable $\mathcal{O} : \Sigma \mapsto \mathbf{R}$.

This leads us to the large class sampling algorithms that come under the name of *Markov Chain Monte Carlo* (MCMC). A nice introductory book on the subject is (Brooks et al., 2011).

Definition 42. *Given an initial assignment $x_0 = \underline{x}$, in each iteration $0 < t < T_{max}$, the MCMC algorithm explores the state space by iterating between these two steps:*

1). *firstly proposing the next state $\underline{x}^{new}$ according to the proposal distribution $\underline{x}^{new} \sim q(. \mid \underline{x}^{curr})$,*

2). *and secondly deciding to accept or reject the move to $\underline{x}^{new}$ according to the acceptance probability $\alpha(\underline{x}^{curr}, \underline{x}^{new})$.*

Together, these two steps form a Markov chain called the kernel of the MCMC procedure:

$$k(\underline{x}^{new} \mid \underline{x}^{curr}) \equiv q(\underline{x}^{new} \mid \underline{x}^{curr}).\ \alpha(\underline{x}^{new},\ \underline{x}^{curr}).$$

Algorithm 3: The Metropolis Hasting algorithm

```
Result: $\underline{x}_T \overset{approx}{\sim} \mu_\beta(.)$
$\underline{x}_0 \leftarrow \tau$;
for $t = 1, ...T_{max}$ do
    $\underline{x}^{new} \sim q(. \mid \underline{x}_{t-1})$;
    $\alpha(\underline{x}_{t-1}, \underline{x}^{new}) \leftarrow \min \left\{ 1, \frac{\mu_\beta(\underline{x}^{new})\, q(\underline{x}_{t-1} \mid \underline{x}^{new})}{\mu_\beta(\underline{x}_{t-1})\, q(\underline{x}^{new} \mid \underline{x}_{t-1})} \right\}$;
    draw $u \sim \mathcal{U}[0,1]$ ;
    if $u \geq \alpha(\underline{x}_{t-1}, \underline{x}^{new})$: then
        $\underline{x}_t \leftarrow \underline{x}^{new}$;
    else
        $\underline{x}_t \leftarrow \underline{x}_{t-1}$;
    end
end
```

Definition 43 (Mezard and Montanari, 2009)**.** *Consider an $N-$particle spin system $(x_1, \dots x_N) \sim \mu_\beta$ with some Hamiltonian $\mathcal{H}(x)$ and whose state space is Σ. Glauber dynamics is a discrete-time Markov chain that is conditioned on an initial state sampled from the target distribution: $x_0 \sim \mu_\beta$, i.e. $\mathbf{P}_{Gb}[x_1 = \underline{y}] = k(x_1 | x_0 = \underline{x}_0)\mu_\beta(\underline{x}_0)$ and $\mathbf{P}[x_t = \underline{x}_t |\ x_1 = \underline{x}_1, x_2 = \underline{x}_2, \dots x_{t-1} = \underline{x}_{t-1}] = \mathbf{P}[x_t = \underline{x}_t |\ x_{t-1} = \underline{x}_{t-1}]$ for $t \geq 2$. Given the current state of the Markov chain $x_t = \underline{x}_t$, the next state is generated as follows:*

1. *Propose the next state uniformly at random from the set of immediate neighbors of the current state: $x_{t+1} \sim \mathcal{U}(\mathcal{N}(\underline{x}_t))$ where $\mathcal{N}(\underline{x}_t) \equiv \{y \in \Sigma :\ d(y, \underline{x}_t) = 1\}$, with $d(,)$ begin the Hamming distance.*

2. *Accept the proposed state according to the probability:*

$$\alpha(\underline{x}_{t+1}, \underline{x}_t) \equiv \min \left\{1, e^{-\beta(\mathcal{H}(\underline{x}_{t+1}) - \mathcal{H}(\underline{x}_t))}\right\}.$$

Definition 44. *An MCMC procedure is said to have detailed balance w.r.t. the target distribution μ_β, if its kernel satisfies*

$$\mu_\beta(\underline{x}^{curr}).\ k(\underline{x}^{new} \mid \underline{x}^{curr}) = \ \mu_\beta(\underline{x}^{new}).\ k(\underline{x}^{curr} \mid \underline{x}^{new}),$$

Definition 45 (Mezard and Montanari, 2009)**.** *The variation distance between two discrete probability distributions ν, μ sharing the same support Σ, is defined as*

$$||\nu - \mu||_v \equiv\ 1/2 \sum_{\underline{x} \in \Sigma} \left|\nu(\underline{x}) - \mu(\underline{x})\right|. \tag{5.2}$$

Theorem 14 (Mezard and Montanari, 2009)**.** *If the probability of reaching x starting from y in a finite number of steps, along the Markov chain defined by the kernel of the MCMC procedure: $k(.| \ .)$, is bounded away from zero, for any two states x, y in Σ the support of the target distribution μ_β, and the kernel satisfies detailed balance w.r.t. μ_β, then the MCMC procedure has μ_β as a stationary distribution. In other* *words, the sampling procedure is guaranteed to converge arbitrarily close in variation distance to the target distribution.*

Since we will mainly discuss Glauber dynamics for the rest of the chapter, we will show, using the above theorem, that is does converge to the target distribution.

Proposition 5. *Given an N-particle spin system $x \sim \mu_\beta$, its Glauber dynamics converges to the target distribution μ_β.*

Proof. Since the transition probabilities are bounded away from zero for all $\beta < \infty$, we need to show the show the detailed balance equation:

i). If $\mathcal{H}(\underline{x}_t) \leq \mathcal{H}(\underline{x}_{t+1})$, then: $\alpha(\underline{x}_{t+1}, \underline{x}_t) \equiv \min\{1, e^{-\beta(\mathcal{H}(\underline{x}_{t+1})-\mathcal{H}(\underline{x}_t))}\} = 1$, and $\alpha(\underline{x}_t, \underline{x}_{t+1}) = \min\{1, e^{-\beta(\mathcal{H}(\underline{x}_t)-\mathcal{H}(\underline{x}_{t+1}))}\} = e^{-\beta(\mathcal{H}(\underline{x}_t)-\mathcal{H}(\underline{x}_{t+1}))}$, such that

$$\begin{aligned}
\mu_\beta(\underline{x}_t).\, k(\underline{x}_{t+1}|\, \underline{x}_t) &= \frac{e^{-\beta\mathcal{H}(\underline{x}_t)}}{\mathcal{Z}}.\, q(\underline{x}_{t+1}|\, \underline{x}_t)\alpha(\underline{x}_{t+1}, \underline{x}_t) \\
&= \frac{e^{-\beta\mathcal{H}(\underline{x}_t)}}{\mathcal{Z}}.\, \frac{\mathbf{1}\{d(\underline{x}_t, \underline{x}_{t+1}) = 1\}}{N}.\, 1 \\
&= \frac{e^{-\beta\mathcal{H}(\underline{x}_{t+1})}}{\mathcal{Z}}.\, \frac{\mathbf{1}\{d(\underline{x}_t, \underline{x}_{t+1}) = 1\}}{N}.\, e^{-\beta(\mathcal{H}(\underline{x}_t)-\mathcal{H}(\underline{x}_{t+1}))} \\
&= \mu_\beta(\underline{x}_{t+1}).\, q(\underline{x}_t|\, \underline{x}_{t+1}).\, \alpha(\underline{x}_t, \underline{x}_{t+1}) \\
&= \mu_\beta(\underline{x}_{t+1}).\, k(\underline{x}_t|\, \underline{x}_{t+1}).
\end{aligned}$$

ii). The second case $\mathcal{H}(\underline{x}_t) > \mathcal{H}(\underline{x}_{t+1})$ follows the exact same argument. ■

5.2 Relaxation time: speed of convergence to the target distribution

5.2.1 Glauber dynamics

For informative purposes, we will recall the meaning of time averages discussed in chapter two. We have the following definition adapted from the comments of ch4 of (Mezard and Montanari, 2009) that we make more explicit. Note that, to distinguish the individual spin indices $k \in [N]$ from the time index t of the Markov chain, for the remainder of this section, we will write $x_k^{(t)}$ to denote the value of the k^{th} spin of the state $x^{(t)} \in \Sigma$.

Definition 46. *Given an N-particle spin system $x \sim \mu_\beta$, and a function $\mathcal{O} : \Sigma \mapsto \mathbf{R}$, we define the time average of $\mathcal{O}(t) \equiv \mathcal{O}(x^{(t)})$ as its expectation w.r.t. Glauber dynamics starting from $x^{(0)} \sim \mu_\beta$, that we denote by brackets:*

$$\langle \mathcal{O}(t) \rangle_{\underline{x}^{(0)}} = \mathbf{E}_{Glauber}\big[\mathcal{O}(x^{(t)}) | x^{(0)} = \underline{x}^{(0)}\big] = \sum_{\underline{x}^{(t)} \in \Sigma} \mathcal{O}(\underline{x}^{(t)})\, \mathbf{P}_{Glauber}\big[x^{(t)} = \underline{x}^{(t)} | x^{(0)} = \underline{x}^{(0)}\big]. \tag{5.3}$$

To make matters more explicit on a simple example, consider the time average of the value of the k^{th} spin after t Glauber steps starting from the realization of $x^{(0)} \sim \mu_\beta$:

$$\begin{aligned}
\langle x_k(t)\rangle_{\underline{x}^{(0)}} &= \sum_{\underline{x}^{(t)}\in\Sigma} \underline{x}_k^{(t)}\ \mathbf{P}_{Glauber}\big[x^{(t)} = \underline{x}^{(t)} | x^{(0)} = \underline{x}^{(0)}\big] \\
&= \sum_{\underline{x}^{(t)}\in\Sigma} \underline{x}_k^{(t)}\ \frac{\mathbf{P}_{Glauber}\big[x^{(t)} = \underline{x}^{(t)}, x^{(0)} = \underline{x}^{(0)}\big]}{\mu_\beta(x^{(0)} = \underline{x}^{(0)})} \\
&= \frac{1}{\mu_\beta(x^{(0)} = \underline{x}^{(0)})} \sum_{\underline{x}^{(1)}\ldots\underline{x}^{(t)}\in\Sigma} \underline{x}_k^{(t)} \left[\prod_{s=2}^{t} p_{Gb}\big[\underline{x}^{(s)}|\underline{x}^{(s-1)}\big]\right] p_{Gb}\big[\underline{x}^{(1)}|\underline{x}^{(0)}\big],
\end{aligned}$$

where the kernel of Glauber dynamic is as previously discussed: $p_{Gb}(\underline{x}^{(t)}|\underline{x}^{(t-1)}) \equiv k(x^{(t)} = \underline{x}^{(t)}|x^{(t-1)} = \underline{x}^{(t-1)}) = \big(\mathbf{1}\{d(\underline{x}^{(t)}, \underline{x}^{(t-1)}) = 1\}/N\big).\ \min\{1, e^{\beta(\mathcal{H}(\underline{x}^{(t)})-\mathcal{H}(\underline{x}^{(t-1)}))}\}$.

One can object to the condition of starting from an equilibrium state $\underline{\tau} \sim \mu_\beta$, however it is quite standard in MCMC sampling to run the algorithm for a *burned in* period, i.e. some prescribed number of iterations (see section **[1.11]** in (Brooks et al., 2011)), to get to a low energy initial state such that it is approximately sampled according to μ_β.

The treatment of dynamics of constraint satisfaction problems is very analogous to that of spin systems with the main difference being, is that in the case of hard constraints, only the states which satisfy a number of constraints carry probability mass, which typically leaves a large portion of the state space with zero probability, effectively separating it into highly disconnected regions delimited by infinite energy barriers, intuitively this will cause a dramatic slowdown of relaxation time.

The goal of the current section is to describe the dependence of the relaxation time on the relative density of these disconnected regions w.r.t. their immediate boundary in the general case, and in the next section we will relate this dependence to the 1RSB predictions and state some bounds on the relaxation time in the replica-symmetric and clustering regimes.

Consider the temperature parameterized distribution $\mu_\beta(\underline{x}) \propto e^{-\beta E(\underline{x})}$, where assignments satisfying a larger number of clauses are associated with lower energies; $E(\underline{x}) \equiv |\{a \in [M] : \underline{x}_{\partial a} \text{ doesn't satisfy } a\}|$. In the case where a k-SAT instance is satisfiable, the uniform measure measure on its satisfying assignments can thus be seen as the low temperature limit of μ_β. In this sense, the typically rugged energy landscape of glassy system such as the REM or the p-spin model, is a smoothed out version of that of hard constrained systems such as $k-$SAT.

Still, the analysis of the dynamics on rugged landscapes provides valuable insight into the algorithmic consequences of the 1RSB predictions, that also hold in the $p-$spin model for $p \geq 3$, which as we will explain below, provided some conditions on the number of interactions that a given spin is allowed to take part into, encompasses the k-SAT and a multitude of random CSPs as special cases.

5.2.2 Speed of correlation decay

Consider an N-particle system evolving according to Glauber dynamics, and suppose that we want to study how fast the value of some observable $\mathcal{O}(x^{(t)}), \mathcal{O} : \Sigma \mapsto \mathbf{R}$, becomes (nearly) uncorrelated with its initial value $\underline{x}^{(0)}$. To get an idea of the timescale over which correlations decay for *arbitrary* observables, it makes sense to look at the observable with *slowest correlation decay* as to get a worst case upper bound for said timescale, that we call the *relaxation time* τ_r.

Definition 47 (Mezard and Montanari, 2009)**.** *Consider an N-particle spin system $x \sim \mu_\beta$ whose support is Σ, and an arbitrary function $\mathcal{O} : \Sigma \mapsto \mathbf{R}$, and let $\mathcal{O}(t) \equiv \mathcal{O}(x^{(t)})$, where $x^{(t)}$ is the state of the Glauber dynamic Markov chain defined above. The exponential auto-correlation time (also known as the relaxation time) τ_r is defined as:*

$$\tau_r \;\equiv\; \sup_{\mathcal{O}}\{\tau_{\mathcal{O},r}\} \quad \textit{where} \quad \tau_r \;\equiv\; -\lim_{t\to\infty} \frac{1}{t} \log C_{\mathcal{O}(t)}, \tag{5.4}$$

where $C_{\mathcal{O}}(t) \equiv \langle \mathcal{O}(0)\mathcal{O}(t)\rangle - \langle \mathcal{O}(0)\rangle\langle \mathcal{O}(t)\rangle$.

Theorem 15 (Mezard and Montanari, 2009)**.** *For any Markov Chain on Σ, that satisfies detailed balance and whose transition probabilities $\{w_{x\to y}\}$ are strictly positive, for any two disjoint subsets $\mathcal{A}, \mathcal{B} \subset \Sigma$, let $\mathcal{W}_{\mathcal{A}\to\mathcal{B}} \equiv \sum_{x\in\mathcal{A}, y\in\mathcal{B}} \mu(x)\, w_{x\to y}$, we have*

$$\tau_r \;\geq\; \frac{\mu(x \in \mathcal{A})\; \mu(x \notin \mathcal{A})}{\mathcal{W}_{\mathcal{A}\to\Sigma\setminus\mathcal{A}}}. \tag{5.5}$$

We will not prove this theorem since it is ubiquitous in statistics and computer science, but we will instead illustrate its utility in the following informative example.

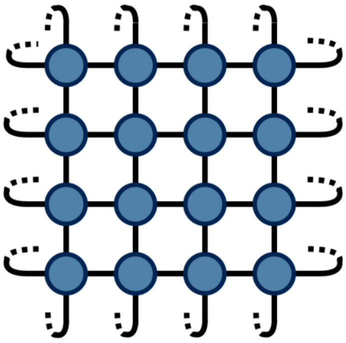


Figure 5.1: Periodic boundary conditions in a $2D$ Ising model with $L = 4$ (TensorNetwork, n.d.)

Consider the $2D$ Ising model, where spins rest on a two dimensional $L \times L$ grid $\mathcal{G} \equiv (\mathcal{V}_{\mathcal{G}}, \mathcal{E}_{\mathcal{G}})$ where $\mathcal{V} = [L^2]$, and $\mathcal{E}_{\mathcal{G}} \equiv \{(i,j) : j \in \mathcal{N}(i),\ \forall i \in \mathcal{V}\}$ with periodic boundaries, such as the one in the figure above. Only nearest neighbor spins are allowed to interact, hence $\mathcal{J} \equiv \{J_{ij} = 0 \ : \ \forall\ (i,j) \notin \mathcal{E}_{\mathcal{G}})\}$. The low temperature behaviour of the $2D$ Ising model is different from that of mean field ones such as the SK model, in the system may be in a ferromagnetic phase, such that its magnetization $M(\sigma) \equiv \sum_{i\in[N]} \sigma_i$ (which random as σ is $\sim \mu$) is concentrated around two symmetric values $\pm N M_+(\beta)$ with $N = L^2$ the total number of particles.

Let Σ, as usual, denote the state space of the $N-$particle system, and let $\mathcal{A} \equiv \{\sigma \in \Sigma : M(\sigma) \geq 1\}$, and suppose that L is odd such that the magnetization cannot be zero, hence $\overline{\mathcal{A}} = \Sigma \backslash \mathcal{A}$ and $\mu_\beta(\mathcal{A}) = \mu_\beta(\Sigma \backslash \mathcal{A}) = 1/2$. By the above theorem we then have:

$$\tau_r \geq \frac{1/2.\ 1/2}{\sum_{\underline{x}:M(\underline{x})=1} \sum_{\underline{y} \in \overline{A}:\, d(\underline{x},\underline{y})=1} \mu_\beta(\underline{x})\, w_{\underline{x} \to \underline{y}}}. \tag{5.6}$$

Note that in the denominator of the lower bound of τ_r we always sum over the boundary of $\mathcal{A}$, which we can define as $\partial\mathcal{A} \equiv \{x \in \mathcal{A} \ : \ \min_{y \in \Sigma \backslash \mathcal{A}} d(y,x) = 1\}$ i.e. the portion of $\mathcal{A}$ from which we can escape $\mathcal{A}$ in one step, provided we follow single-flip dynamics i.e. the transition probabilities $\{w_{x \to y}\}$ are non-zero only for neighboring states.

Hence, the relaxation time grows larger as the probability mass on the boundary of $\mathcal{A}$ grow smaller, and since the above theorem holds for any subset $\mathcal{A} \subset \Sigma$, it suffices that there exists one portion of the state whose boundary carries a small enough probability mass that the lower bound of τ_r becomes quite large, rendering the relaxation time much slower.

After some work we get a size dependent lower bound $\tau_r \geq e^{2\beta\sigma(\beta)L + o(L)}$, we refer the reader to chapter 13 of (Mezard and Montanari, 2009) for details.

5.3 A further characterization of pure state decomposition in the clustering regime

The description of the support of the uniform measure on satisfying assignments μ beyond the clustering threshold has been rather informal, to relate its 1RSB picture with the performance of MCMC algorithms, it is useful to go into further detail. This subsection will be reminiscent of the treatment of pure state decomposition in chapter two, and although it is self contained, it can be helpful to have read the second chapter beforehand, as it gives some intuition from the dynamical point of view.

Recall that in the discussion of intermediate ergodicity regimes, ergodicity is really only broken in the large N limit, however, in the low temperature regime as $\beta \gg 1$, the growth of energy barriers $\Delta E_{\sigma \to \tau} \equiv \mathcal{H}(\tau) - \mathcal{H}(\sigma)$, results in exponentially vanishing escaping probabilities $\mathcal{W}_{\mathcal{A} \to \overline{\mathcal{A}}}$, making the relaxation time prohibitively large and thus the convergence of standard MCMC sampling to be approximate to μ_β too slow to be practically useful.

Since the breaking of ergodicity is size-dependent, the *pure state decomposition* (or clustering) at the onset of the clustering phase can be characterized asymptotically. Since we have situated the clustering in the 1RSB picture and discussed its qualitative properties (see the second-last section of ch.4), here we give the complete definition.

Definition 48 (Mezard and Montanari, 2009)**.** *Consider a sequence of finite factor graphs* $\mathbf{F}_N$*, each associated with a* $k-SAT$ *instance and its set of negation spec-*

ifications $\{J_{ai} : i \in \partial a\}_{a\in[M]}$ that uniquely determine its set of satisfying assignments $\mathcal{S}_N$. And let μ_N be the uniform measure on $\mathcal{S}_N$ with support on the entire state space $\Sigma_N = \mathcal{X}^N$ (but positive only on $\mathcal{S}_N$). The pure state decomposition can be formalized by considering the partitioning of the state space into separate blocks, $\Sigma_N = \bigsqcup_{k\in[\eta_N]} \alpha_{k,N}$, where each containing a cluster. Moreover, clusters are separated by bottlenecks $\partial_\epsilon \alpha_{k,N} \equiv \{x \in \Sigma_N : 1 \leq \min_{\underline{y}\in\alpha_k} d(x,\underline{y}) \leq N\epsilon\}$. We say that μ_N is in a pure state decomposition if the it satisfies the three following conditions:

1. $\max_{k\in[\eta_N]} \Big\{\mu_N(\alpha_{k,N})\Big\} \leq 1-\delta$ *for some* $\delta > 0$.

2. $\lim_{N\to\infty} \max_{k\in[\eta_N]} \dfrac{\mu_N(\partial_\epsilon \alpha_{k,N})}{\mu_N(\alpha_{k,N})} = 0$ *for some* $\epsilon > 0$.

3. *The pure state measures* $\mu_N^{\alpha_{k,N}}$ *admit no further pure state decomposition of the type above.*

As we noted in chapter 2, the reformulation of μ_N into a convex combination of measures with non-overlapping support, follows from conditioning it on clusters:

$$\mu_N^{\alpha_k}(\underline{x}) \equiv \mathbf{P}_{\sim\mu_N}[\underline{x} \in \mathcal{S}_N \mid \underline{x} \in \alpha_{k,N}] = \frac{\mathbf{P}_{\sim\mu_N}[\{\underline{x} : \underline{x} \in \mathcal{S}_N\} \cap \{\underline{x} : \underline{x} \in \alpha_{k,N}]\}]}{\mathbf{P}_{\sim\mu_N}[\{\underline{x} : \underline{x} \in \alpha_{k,N}]\}]}$$
$$= \frac{\mu_N(\underline{x})\ \mathbf{1}\{\underline{x} \in \alpha_{k,N}\}}{\mu_N(\alpha_{k,N})}.$$

Let the Gibbs weight of the k^{th} cluster be $w_k \equiv \mu_N(\alpha_{k,N})$, the above equality then yields: $\mu_N(\underline{x})\ \mathbf{1}\{\underline{x} \in \alpha_{k,N}\} = w_k\ \mu_N^{\alpha_k}(\underline{x})$, such that

$$\mu_N(\underline{x}) = \sum_{k\in[\eta_N]} \mu_N(\underline{x})\ \mathbf{1}\{\underline{x} \in \alpha_{k,N}\} = \sum_{k\in[\eta_N]} w_k\ \mu_N^{\alpha_k}(\underline{x}). \tag{5.7}$$

Note that the clusters do not contain the entirety of the set of solutions, nor does $\mu_N^{\alpha_k}(z) = 1/|\mathcal{Z}^{\alpha_k}|$, $\forall z \in \alpha_{k,N}$. Since the 2^{nd} condition holds for arbitrarily $\epsilon-$small bottlenecks, the clusters are well delimited. This holds thermodynamically for all clusters $l \in [\eta_N]$, since

$$\lim_{N\to\infty} \mu_N(\partial_\epsilon \alpha_{l,N})/\mu_N(\alpha_{l,N}) \leq \lim_{N\to\infty} \max_{k\in[\eta_N]} \mu_N(\partial_\epsilon \alpha_{k,N})/\mu_N(\alpha_{k,N}) = 0.$$

5.3.1 Algorithmic implications of pure state decomposition

Throughout, this subsection we will refer to regions of Σ with rapid mixing under local Markov dynamics (e.g. Glauber) as being ergodic, and broken ergodicity is taken to mean exponential energy barriers.

Recall that the satisfiability problem consists in answering whether the solution set $\mathcal{S}$ of a given $k-$SAT instance is non-empty and if so, to generate samples from it. Ideally we would like to sample $\sim \mu$, that is, uniformly from the solution set, the uniformity criterion is however very costly; for instance, a result of (Papadimitriou, 1991) shows that biased random walk strategies yield solutions of $2-$SAT instances in polynomial time $\mathcal{O}(N^2)$, the issue however is that such biased random walks tend to oversample from a portion of the solution space such that the sampling procedure is highly non-uniform.

For target product measures such as $\mu(\underline{x}) \propto \prod_{a\in[M]} e^{-E(\underline{x}_{\partial a})}$, a common choice for the acceptance probabilities is to take $\alpha(x,y) \equiv \max\{1, -\beta\Delta E_{x\to y}\}$, and $q(.|x) \equiv \mathcal{U}(\mathcal{N}(y))$ as a proposal distribution. In the $k-$SAT case, $E(\underline{x}_{\partial a}) = -\log(\psi_a(\underline{x}_{\partial a}))$.

However, since the convergence of such sampling procedures requires that the transition probabilities must be bounded away from zero, i.e. $w_{x\to y} > 0,\ \forall\, x, y \in \Sigma$, such that the system is able to explore all states with positive probability under the target probability distribution μ, and that the uniform measure on satisfying assignment is zero for large portion of the state space, effectively separating it into disconnected clusters as $\alpha \geq \alpha_{cl}$, we need to introduce a smoothed out version of μ_β which for low enough temperature produces samples approximately $\sim \mu$.

To this end, we can define a Gibbs measure $\mu_\beta(\underline{x}) \propto e^{-\beta E(\underline{x})}$, where each state's energy is given by the number of clauses it violates. Recalling the notation above, given a $k-$SAT formula, let $J_{a,i} \equiv \mathbf{1}\{x_i \text{ is negated in a}\}$, such that $E(\underline{x}) = \sum_{a\in[M]} \prod_{i\in\partial a} \delta_{\underline{x}_i, J_{a,i}}$. Notice that the set of solution all have the same probability under μ_β, we can therefore produce uniform samples from $\mathcal{S}$ by using a Markov chain whose target distribution is mu_β, and discard all samples $\notin \mathcal{S}$ using rejection sampling.

Note that there is a trade-off in lowering β, for small β, states with $E(\underline{x}) \geq 1$ will carry a substantial part of the probability mass of μ_β. such that the MCMC procedure will produce a large number of rejected samples, however, when β is too high, we can get the target distribution μ_β to be arbitrarily close to μ in variation distance $(||\mu_\beta - \mu||_v \equiv 1/2 \sum_{\underline{x}\in\Sigma} \mu_\beta(\underline{x}) - \mu(\underline{x}) \leq \epsilon)$, but what we gain in reducing the number of rejected samples is lost in the time spent stuck in local energy minima due to the exponential vanishing of escaping probablities, more precisely; since in SA, the moves out of subvalleys of energy are accepted with probability $w_{x\to y} \equiv \min\{1, e^{-\beta\Delta E_{x\to y}}\}$, for $\beta \gg 1$, we have $w_{x\to y} \approx 0\ , \forall\, y \in \{z \in \Sigma :\ E(z) \geq 1\}$.

5.3.2 A 1RSB-independent result: Simulated annealing is uniform but very slow

Before discussing the slowdown of the relaxation time as a consequence of the 1RSB picture in the case of the $k-$SAT for $k \geq 3$, it is perhaps helpful to characterize the *solution time*, i.e. the time it takes to find the first solution, in the simpler case of $k = 2$. Consider a $2-$SAT instance of the form

$$\Phi(\mathrm{x}) \ = \ (a \vee c_1) \ \wedge \ (a \vee c_2) \ \wedge \ \ldots \ (a \vee c_n) \ \wedge \ (a \vee b) \ \wedge \ (\neg a \vee \neg b), \tag{5.8}$$

and suppose that the sampling procedure has the uniform distribution over N variables as proposal distribution: $q(x_t \mid x_{t-1}) \equiv \mathcal{U}([N])$, and $\alpha(x_t, x_{t-1}) \equiv \min\{1, e^{-\beta\Delta E_{x_{t-1},x_t}}\}$ as accepting probability, such that for $\beta < \infty$, the transition probabilities are bounded away from zero and the transition kernel $k(x_t \mid x_{t-1}) = q(x_t \mid x_{t-1})\, \alpha(x_t, x_{t-1})$ satisfies detailed balance, and the resulting Markov chain is therefore guaranteed to converge to the target distribution $\mu_\beta(\underline{x}) \propto e^{-\beta E(\underline{x})}$, with E counting the number of unsatisfied clauses by assigning $\underline{x}$. For high enough β, $\mu_\beta \simeq \mu$ in variation distance.

Theorem 16 (Wei et al., 2004)**.** *Fixed temperature Simulated annealing finds a solution to the above formula in $\mathcal{O}(e^N)$ time with high probability.*

The complete proof (Papadimitriou, 1991) is involved and rather long, we will however present the gist of the argument, as discussed in (Wei et al., 2004), for intuition's sake. Starting from an a uniformly chosen assignment, we would like to estimate the number of iteration before finding the first satisfying assignment. Given $\underline{x}_0 \sim \mathcal{U}(\Sigma)$, there are two options:

a. $\underline{a}_0 = 1$: let $\mathcal{V}$ denote the set of variables, it is easy to see that flipping any variable $x \in \mathcal{V} \setminus \{a\}$, will not increase the number of unsatisfied clauses, such that, the change in energy induced by flipping x is greater or equal to zero, and therefore $e^{-\beta \Delta E_{a,\neg a}} \geq 1$, and will hence be flipped with probability one, since the acceptance probability is $\alpha \equiv \{1, e^{-\beta \Delta E}\} = 1$.

   Let $N \equiv |\mathcal{V}|$, since the sampling procedure flips any variable uniformly at random, if $|\mathcal{V} \setminus \{a, b\}|$ is large enough, it can be approximated by an unbiased random walk on the $(N-2)-$dimensional hypercube defined by the variables in $\{c_k\}$. The probability of reaching a region of the entire $N-$dimensional hypercube where only $\mathcal{O}(\sqrt{N})$ of the variables in $\{c_k\}$ are set to zero (or false).

   Suppose we are in a region of the state space where $|\{k \,:\, c_k = 0\}| = \mathcal{O}\sqrt{N}$ and $a = 0$, it is easy to see that the last two clauses enforce the constraint that $a = 1$ in any solution. Hence to reach a satisfying assignment, we need to flip the variable a, but since $|\{k \,:\, c_k = 0\}| = \mathcal{O}\sqrt{N}$ this will result in a change of energy of $\Delta E_a = \sqrt{N}$, and therefore, the acceptance probability of flipping a, will be $\alpha = \min\{1, e^{-\beta\sqrt{N}}\} = \mathcal{O}(e^{-N})$.

b. $\underline{a}_0 = 0$: a will be flipped in polynomial time with high probability, and then we're back in case (a).

On the other hand, one should note that the relaxation time on a full hyper cube is $\mathcal{O}(N \log(N))$ (Wei et al., 2004), which would explain the experimentally uniform sampling among solutions to the same cluster, however, the 1RSB picture does not *guarantee* that clusters contain solutions exclusively, but rather that they are much denser than their surrounding bottlenecks (see the inequality right down below).

5.3.3 The relaxation time in replica-symmetric vs. clustering regimes

As previously discussed, in the clustering regime

$$\frac{\mu_N(\partial_\epsilon \Omega_{r,N})}{\mu_N(\Omega_{r,N})} \leq e^{-N^q}, \tag{5.9}$$

where q depends on the degree profile of the random graph ensemble from which the $k-$SAT instance is generated (Mezard and Montanari, 2009).

Since we are considering single-flip Glauber dynamics, the set of moves out of a cluster $\mathcal{A} \equiv \alpha_{k,N}$, with nonzero probability are those with Hamming distance equal to one, i.e. those whose end state is located at the immediate boundary of the cluster. It is easy to see that the immediate boundary of the cluster $\alpha_{k,N}$ is given by its bottleneck of radius $\epsilon = 1/N$, such that $\partial_{1/N}\alpha_{k,N} = \{x \in \Sigma_N : \min_{\underline{y}\in\alpha_{k,N}} d(x,\underline{y}) = 1\}$. By the inequality above, we have $\mu_N(\partial_\epsilon \alpha_{k,N})\, e^{N^q} \leq \mu_N(\alpha_{k,N})$, and recalling the above theorem regarding relaxation times, for $\mathcal{A} \equiv \alpha_{k,N}$, we get

$$\begin{aligned}\tau_r \;&\geq\; \frac{\mu_N(x\in\mathcal{A})\,\mu_N(x\notin\mathcal{A})}{\mathcal{W}_{\mathcal{A}\to\Sigma\backslash\mathcal{A}}} \;=\; \frac{\mu(\alpha_{k,N})\,\mu(\partial_\epsilon\alpha_{k,N})}{\sum_{\underline{x}\in\alpha_{k,N}}\sum_{\underline{y}\in\,\partial_{1/N}\alpha_{k,N}}\mu_N(\underline{x})\,w_{\underline{x}\to\underline{y}}}\\ &\geq\; e^{N^q}\,\underbrace{\left(\frac{\mu^2(\partial_\epsilon\alpha_{k,N})}{\sum_{\underline{x}\in\alpha_{k,N}}\sum_{\underline{y}\in\,\partial_{1/N}\alpha_{k,N}}\mu_N(\underline{x})\,w_{\underline{x}\to\underline{y}}}\right)}_{\equiv\,\gamma(N)},\end{aligned}$$

with some work, we could upper bound $\gamma(N)$ by an $N-$independent constant. As pointed out in the last chapter in (Mezard and Montanari, 2009), in many factor graph ensembles, the constant q is equal to one, such that the relaxation time beyond the clustering transition is of order $\tau_r \geq\ cst_1\, e^N$. As we will show below, $\tau_r \leq\ cst_2\, e^N$, such that the relaxation time is exactly exponential in size; $\tau_r =\ e^{\Theta(N)}$.

Recall the non-reconstruction criterion of vanishing long range correlation, without which it is possible to reconstruct the assignment from a limited number of known variable assignments. A very analogous problem to $k-$SAT is $q-$coloring, since it can be translated as a satisfiability problem and is believed to be in the same discontinuous 1RSB universality class described above.

Definition 49. *We say that* $\mathbf{F} \equiv (\mathcal{V},\mathcal{F},\mathcal{E})$, *is an* $l-$*regular factor graph, if all of its variable nodes have the same degree:* $|\partial i| = l,\ \forall i \in \mathcal{V}$.

Definition 50. *Consider a d-regular graph* $\mathcal{G} \equiv (\mathcal{V},\mathcal{E})$, *and let* $\mathcal{K} \equiv [k]$ *be a set of* k *colors and* $N \equiv |\mathcal{V}|$. *A k-coloring of* $\mathcal{G}$ *is an assignment* $\underline{\sigma} \in \mathcal{K}^N$ *where* $\forall\,(i,j)\in\mathcal{E},\ \underline{\sigma}_i \neq \underline{\sigma}_j$. *We denote by* $\mathcal{S}_\mathcal{G}$ *the set of all possible* $k-$*colorings of* $\mathcal{G}$.

Theorem 17 (Zhang, 2017)**.** *Consider the problem of sampling from the uniform measure of* $k-$*coloring of* $d-$*regular trees* $\mu(\sigma) \equiv \mathbf{1}\{\sigma\in\mathcal{S}_\mathcal{G}\}/\mathcal{Z}$, *where* $\mathcal{Z} \equiv |\mathcal{S}_\mathcal{G}|$ *and* $d \leq d_{rec}$ *the reconstructability threshold. Then, there exists a constant* k_0 *such that for* $k \geq k_0, l\beta < 1$ *and* $d \leq\ k.[\log k + \log(\log k) + \beta]$, *the mixing time of Glauber dynamics is* $\mathcal{O}(N\log N)$.

.

A more complete description of mixing time below the non-reconstuctable regime is detailed in chapter five of the PhD thesis of Yumeng Zhang (Zhang, 2017). A key condition of rapid mixing is that correlation length be small, in fact (Montanari and Semerijan, 2008) showed that the Glauber dynamics of $p-$spin glass models on random regular graphs satisfies $c_1 l \leq\ \tau_r \leq\ \exp\{c_2 l^d\}$ where l is the correlation length.

More generally, consider an $p-$spin Ising model and its associated factor graph $\mathbf{F} \equiv (\mathcal{V}, \mathcal{F}, \mathcal{E})$, and let $N \equiv |\mathcal{V}|$, $M \equiv |\mathcal{F}|$, and suppose that the $p-$spin model is represented by an $l-$regular factor graph, such that all variable nodes have degree l, and (by definition of the p-spin model) all factor nodes have degree p, such that $Nl = Mp$. It is easy to see that the regular random $k-$SAT and regular k-coloring are special cases of this model. Moreover, we recall that the energy of each state in the p-spin case is given by the Hamiltonian $\mathcal{H}(\sigma) \equiv -\sum_{a=1}^{M} J_a \prod_{i\in\partial a} \sigma_i$, and it's probability by $\mu_\beta(\sigma) \equiv e^{-\beta\mathcal{H}(\sigma)}/\mathcal{Z}$.

When $p = 2$, $l \geq 3$, the system undergoes the statical phase transition (condensation) without the dynamical one (clustering), and when $p = l = 2$, the system experiences no phase transition in finite temperature. The situation becomes qualitatively different when $p, l \geq 3$, as the system experiences both. The set of N-particle spin system that fall into this case make up what is called the discontinuous 1RSB universality class.

To be more specific about the notion of correlation length, much like the slowest observable to equilibrium, we can define the *most correlated observables* to measure the point-set correlation of a given variable x_i with the set of variables at a distance greater or equal than r that we denote by $x_{\sim i,r}$:

Definition 51. *Consider a factor graph* $\mathbf{F} \equiv (\mathcal{V}, \mathcal{F}, \mathcal{E})$, *we define the distance between two variables* $d(i,j)$ *for* $x_i, x_j \in \mathcal{V}$, *as the number of function nodes along the shortest path from* x_i *to* x_j.

To avoid confusion, we note that while the Hamming distance takes as input two states: $d_h : \Sigma \times \Sigma \mapsto \{1 \ldots N\}$, the distance between two variables takes as input two variables indices $d : \{1 \ldots N\} \times \{1 \ldots N\} \mapsto \{1 \ldots M\}$, where $M \equiv |\mathcal{F}|$.

Definition 52 (Montanari and Semerijan, 2006)**.** *Let* $\mathcal{F}$ *be the space of bounded functions of one variable:* $f : \mathcal{X} \mapsto [-1, 1]$. *We define the point-set correlation of* x_i *with* $x_{\sim i,r} \equiv \{x_j : d(i,j) \geq r\}$ *as:*

$$G_i(t) \equiv \sup_{f,F\in\mathcal{F}} \Big| \langle f(x_i).F(x_{\sim i,r})\rangle - \langle f(x_i)\rangle.\langle F(x_{\sim i,r})\rangle \Big|.$$

Definition 53 (Montanari and Semerijan, 2006)**.** *Given a variable* x_i, *we define its correlation length as the radius (in terms of* $d(i,j)$*) beyond which, the sup point set correlation between* x_i *and all variables outside of the radius are arbitrarily small, more precisely:*

$$l_i(\epsilon) \equiv \min \Big\{ l \leq 0 \ : \ G_i(r) \leq \epsilon, \ \forall\, r \geq l \Big\}.$$

Note that, the correlation length provides a worst case guarantee, since we are considering the decay of the *supremum* point set correlation between x_i and all variables outside of the radius. As previously discussed, we can define the relaxation time via a positive monotonically decreasing *worst case time correlation function* of a given variable x_i:

Definition 54 (Montanari and Semerijan, 2006). *Given a variable x_i, we define the worst-case time-correlation function as:*

$$C_i(t) \equiv \sup_{f \in \mathcal{F}} \Big| \langle f(x_i(0)).f(x_i(t)) \rangle - \langle f(x_i(0)) \rangle . \langle f(x_i(t)) \rangle \Big|.$$

The relaxation time is then given by

$$\tau_i(\epsilon) \equiv \inf \Big\{ \tau \geq 0 : C_i(t) \leq \epsilon \, , \forall t \geq \tau \Big\}.$$

Let the set of variables at a distance (number of function nodes along the shortest path) less or equal to k, be denoted by $\mathcal{B}_i(k)$ the ball of radius k centered around x_i, and let the Markov chain of the sampling procedure be such that, transition probabilities from $(\underline{x}_1, \dots, \underline{x}_i, \dots)$ to $(\underline{x}_1, \dots, \underline{x}_i^{new}, \dots)$, be denoted by $\kappa_i^{\underline{x}}(\underline{x}_i^{new})$ and satisfy $\kappa_i^{\underline{x}} \geq 0$,$\sum_{y \in \mathcal{X}} \kappa_i^{\underline{x}}(y) = 1$, and let the Markov chain probabilities be bounded away from zero by $0 < \kappa_0 \leq \kappa_i^{\underline{x}}(y)$, for all $y \in \mathcal{X}$, $\underline{x} \in \Sigma$. Along Glauber dynamics $\kappa_i^{\underline{x}} \equiv e^{-\beta \Delta_i E_{\underline{x}}}$. We have the following result.

Theorem 18 (Montanari and Semerijan, 2006).

$$C_1 \, l_i(|\mathcal{X}|\sqrt{2\epsilon}) \leq \tau_i(\epsilon) \leq 1 + A \exp\Big\{ C_2 \, \big|\mathcal{B}_i(l_i(\epsilon/2))\big| \Big\},$$

for $A \equiv \log(4/\epsilon)$, $C_1 \equiv 1/2e\Delta^2$, $C_2 \equiv -\log(\kappa_0(1 - e^{-1}))$, *where* $\Delta \equiv \max\{ |\partial i|, |\partial a| : \forall i \in [N], a \in [M]\}$, *with the lower bound being true if* $l_i(|\mathcal{X}|\sqrt{2\epsilon}) > \log_2(2/\epsilon)$.

.

Let $T_c < T_d$ ($\beta_d < \beta_c$) be the critical (inverse) temperatures for the static (c) and dynamic (d) phase transitions. In the high temperature regime, $T > T_d$, the system is rapidly mixing and the relaxation time is independent of size: $\tau_i = \mathcal{O}(1)$, and the correlation lengths l_i are therefore finite.

For $T < T_d$, heuristic arguments based on energy barriers, using the quenched potential method (Franz and Parisi, 1995) and (Franz, 2006), reveal an exponential relaxation time: $\tau_i = \mathcal{O}(e^N)$, correlation lengths are therefore necessarily divergent in N. The method is however non-rigorous, (Montanari and Semerijan, 2006) provide a rigorous result in the proposition below.

Note that, since the ball of radius r of $l-$regular factor graph contains at most $\gamma \equiv |\mathcal{B}_i(r)| \leq l(p-1)^r(l-1)^r \leq l[(p-1)(l-1)]^{l_i(\epsilon/2)}$ variables, we have $\log_{(p-1)(l-1)}(\gamma/l) \leq l_i(\epsilon/2)$, the local tree structure provides lower bound: $cst_1. \log(N) \leq \log_{(p-1)(l-1)}(\gamma/l) \leq l_i(\epsilon/2)$. Moreover, the correlation length $l_i(\epsilon)$ cannot possibly be larger than the diameter the entire factor graph which is of order $\mathcal{O}(\log(N))$, hence

$$cst_1. \log(N) \leq \log_{(p-1)(l-1)}(\gamma/l) \leq l_i(\epsilon/2) \leq cst_2. \log(N).$$

In other words, when $T < T_d$, the above theorem implies that the correlation length is exactly linear in size, $l = \Theta(\log N)$ and $l = \mathcal{O}(1)$ for $T > T_d$. Moreover, since $|\mathcal{B}_i(r)| \leq N$, we have $\tau \leq e^{cst_3 N}$.

Consider a $p-$spin model on an $l-$regular factor graph, with $p, l \geq 3$ and suppose the state of the system is evolving along Glauber dynamics, i.e. with transition probability $\kappa_i^{\underline{\sigma}} \equiv \Delta E_i(\underline{\sigma})$ it is easy to see that the change in energy induced by flipping the i^{th} spin is given by: $\Delta E_i(\underline{\sigma}) = (1 + \underline{\sigma}_i \tanh \beta \psi_i(\underline{\sigma}))/2$ with $\psi_i(\sigma) \equiv \sum_{a \in \partial i} J_{a,i} \prod_{j \in \partial a \setminus i} \sigma_j$.

Rather than characterizing the relaxation time within epsilon distance of the critical points, which are computed by the cavity method, in (Montanari and Semerijan, 2008), the authors defined lower and upper bounds beyond which the correlation time of the i^{th} spin τ_i becomes of a different order. Let T_{ann} be an upper bound on T_c, and T_{barr}, T_{fast} lower and upper bounds, respectively, on T_d, such that $T_{barr} < T_d < T_{fast} < T_c < T_{ann}$.

Proposition 6 (Montanari and Semerijan, 2006)**.** *Let* $T_{p,l}^{fast} \equiv 1/arc\tanh(1/l(p-1))$, *and* $T_{p,l}^{barr}, T_{p,l}^{ann}$ *be as defined in appendix D of (Montanari and Semerijan, 2008), then*

- *If* $T > T_{p,l}^{fast}$, *we have*

$$\tau_i(\epsilon) \leq \log(1/\kappa\epsilon)/\kappa \quad \textit{for} \quad \kappa \equiv 1 - l(p-1)\tanh\beta.$$

- *If* $T_{p,l}^{ann} < T < T_{p,l}^{barr}$ *then there exists* q_* *and* $\Upsilon > 0$ *such that,* $\forall\, 0 < \delta < 1/4$,

$$\tau_i(\epsilon) \geq e^{N(\upsilon-\delta)} \quad \textit{holds for at least } N(q_* - \delta - \epsilon) \quad \textit{spins } \sigma_i, \textit{ w.h.p.}$$

5.4 Uniform exploration of the solution space

5.4.1 Previous work and challenges

For the rest of this chapter, we will write $x^{(i)}$ to denote the assignment resulting from flipping the value of the the i^{th} variable x_i (from 1 to 0 or vice versa), and leaving all others unchanged. Moreover, we will abuse notation to write ΔE when referring to the energy change in an MCMC procedure, from the current state x to the proposed state y, that we denote by $\Delta E_{x \to y}$. Finally, we will refer to the set of assignment with Hamming distance 1 from the current assignment as $\mathcal{N}(x) \equiv \{y \in \Sigma : d_h(x, y) = 1\}$, as we did in previous sections.

As a compromise between the speed of random walk approaches (such as Walk-SAT), and the uniformity guarantee of MCMC sampling, (Wei et al., 2004) proposed a Hybrid approach named *Sample-SAT*, which alternates between random walk moves and fixed temperature SA moves with probability p_{rw}.

The algorithm thus depends on two parameters; p_{rw} and β the fixed inverse temperature of the SA moves. Extensive experiments showed optimal performance for $p_{rw}^* = 0.5$, $\beta^* = 100$ (Wei et al, 2004). Now, let $\underline{x}^{(i)}$ denote the assignment obtained by flipping the i^{th} variable of $\underline{x}$, and let $\mathcal{S}$ be the set of satisfying assignments and

$\mathcal{C}_u(\underline{x}_t)$ the set of clauses unsatisfied by $\underline{x}_t$, and recall that $\psi_a(\underline{x}_{\partial a}) = 1 - \prod_{j \in \partial a} \delta_{\underline{x}_i, J_{a,i}}$, the proposed algorithm is then as follows:

Algorithm 4: The Sample-SAT algorithm

```
Result: $\mathcal{S}ol \subset \mathcal{S}$
Initialization: $p^*_{rw} = 0.5,\ \beta^* = 100$;
$\mathcal{S}ol \leftarrow \emptyset$;
$\underline{x}_0 \sim \mathcal{U}(\Sigma)$;
for $t = 1, ...T_{max}$ do
    draw $u \sim \mathcal{U}[0,1]$ ;
    if $u \geq p_{rw}$: then
        $\mathcal{C}_u(\underline{x}_t) \leftarrow \emptyset$;
        while $\mathcal{C}_u(\underline{x}_t) \neq \emptyset$ do
            $a \sim \mathcal{U}(\mathcal{C}_u(\underline{x}_t))$;
            $i \sim \mathcal{U}(\partial a)$;
            $\underline{x}_t \leftarrow \underline{x}_t^{(i)}$;
            if $\psi_a(\underline{x}_{\partial a,t}) = 1$ then
                $\mathcal{C}_u(\underline{x}_t) \leftarrow \mathcal{C}_u(\underline{x}_t) \setminus \{a\}$;
            end
        end
        $\mathcal{S}ol \leftarrow \mathcal{S}ol \sqcup \{\underline{x}_t\}$;
    else
        $\mathcal{S}ol \leftarrow \mathcal{S}ol \sqcup SA(\beta)\_samples$;
    end
end
```

The uniformity of zero temperature SA on clusters was shown in (Wei et al., 2004), by interchanging the detailed balance condition with stationarity and irreducibility of the Markov chain w.r.t to μ. However, a careful consideration points to the fact that the authors implicitly assume that clusters are connected components of solutions in Hamming space, i.e. that they contain no unsatisfying assignments, such that excluding states in the immediate boundary of the cluster, moves to any neighbor are accepted with probability one.

This is however not very important, since for $N \gg 1$, pure state decomposition implies that $\max_{k \in [\nu]} \mu(\partial \alpha_k)/\mu_{\alpha_k} = 0$, such that the relative density of clusters to their boundary is large enough, that for any $\underline{x} \in \alpha_k$, $\sum_{j \in [N]} \mu(\underline{x}^{(i)})/N \approx 1 - o(1)$. In other words, when starting inside a cluster, the proposed end state when chosen uniformly $i \sim \mathcal{U}([N])$, is with high probability a satisfying assignment.

Even though the random walk component throws out any convergence guarantee to sampling $\overset{approx}{\sim}$ μ, interweaving SA moves with a random walk, significantly improves uniformity of sampling over the solution space compared to Walk-SAT. This can be seen in the reduction of the difference in the frequency of each of the generated solutions, from Walk-SAT to Sample-SAT in the left-side plots below. The uniformity of SA is also reflected in the relative frequency between solutions belonging to the same cluster, as can be observed from the plots in the right-side, showing generated solutions in a low dimensional projection of the N dimensional Hamming space (Wei et al., 2004).

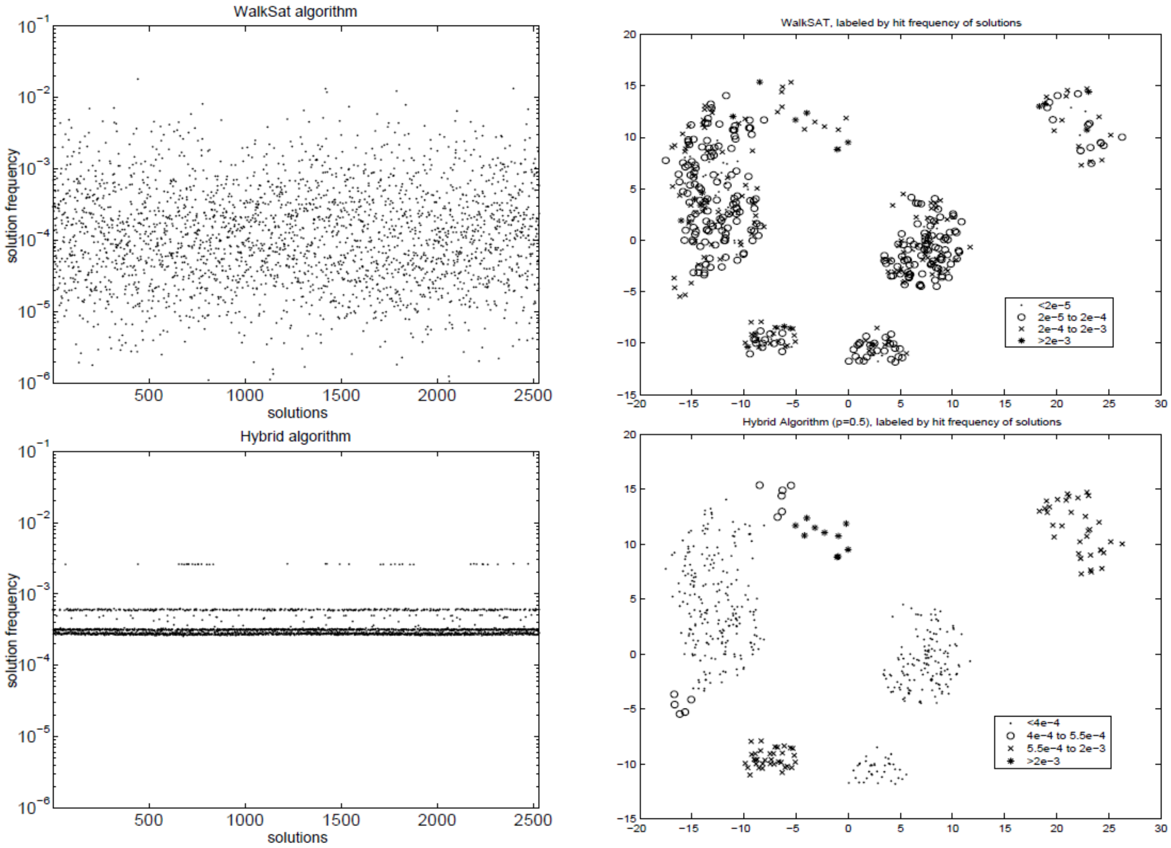


Figure 5.2: *Right:* Solution frequency in Hamming-space for Walk-SAT (top) vs. Sample-SAT (bottom). *Left*: Variance of solutions frequencies (high variance signals non-uniformity) (Wei et al., 2004).

In the clustering regime, the experiments show that the algorithm alternates between two phases. Indeed, given an arbitrary initial assignment, the system moves according to a random walk until it finds a cluster, once it does, it is locked within said cluster sampling uniformly from it until "equilibriation", then reverting back to a random walk, and so on.

This two-phase behaviour is not so surprising: since solutions are grouped into clusters, all proposed single-flip end-states, inside a cluster (say α_k), have zero unsatisfied clauses, and hence $\Delta E_i(\underline{x}_t) \equiv E(\underline{x}_t) - E(\underline{x}_t^{(i)}) = 0$ for all $(\underline{x}_t, \underline{x}_t^{(i)}) \in \alpha_k$, such that the acceptance probability $\alpha(\underline{x}_t, \underline{x}_t^{(i)}) \equiv \min\{1, e^{-\beta \Delta E_i(\underline{x}_t)}\}$ is equal to one for any such move, and the algorithm finds itself effectively locked into the SA phase as soon as it hits a cluster, until it reaches the sparsely populated (in terms of number of solutions) boundary of the cluster, where most proposed states are not solutions, and have therefore higher energy, such that $\alpha(\underline{x}_t, \underline{x}_t^{(i)}) \equiv e^{-\beta \Delta E_i(\underline{x}_t)} \approx 0$, for large enough β, and are thus rejected, until the algorithm generates a long enough sequence of $u > p_{rw}$ that the random walk gets far away from the cluster and the algorithm reverts back to performing a random walk in search of a new solution.

5.4.2 Uneven visiting-frequency among clusters

As is clear from the right-side plots, while the sampling is uniform inside a given cluster, some clusters are more frequently visited than others. More precisely, let $\mathbf{P}_{S-SAT}$ be the probability of reaching an assignment in $\mathcal{S}$ using Sample-SAT, then for any satisfying assignment $\underline{y}$, we have

$$\begin{aligned}\mathbf{P}_{S-SAT}[x_t = \underline{y}] &= \sum_{\alpha_k : k \in [\eta]} \mathbf{P}_{S-SAT}[x_t = \underline{y}]\, \mathbf{1}\{\underline{y} \in \alpha_{k,N}\} \\ &= \sum_{\alpha_k : k \in [\eta]} \mathbf{P}_{S-SAT}[x_t = \underline{y} \mid \underline{y} \in \alpha_k].\, \mathbf{P}_{S-SAT}[\underline{y} \in \alpha_k].\end{aligned}$$

Let $\mathcal{Z}_{\alpha_k}$ be the set of solution inside the cluster α_k, since SA samples uniformly from clusters, we have: $\mathbf{P}_{S-SAT}[x_t = \underline{y} \mid \underline{y} \in \alpha_k] = 1/\mathcal{Z}_{\alpha_k}$, such that,

$$\mathbf{P}_{S-SAT}[x_t = \underline{y}] = \sum_{\alpha_k : k \in [\eta]} \frac{\mathbf{1}\{\underline{y} \in \alpha_k\}}{\mathcal{Z}_{\alpha_k}}.\, \mathbf{P}_{S-SAT}[\underline{y} \in \alpha_k].$$

Recall that in the clustering regime, the Gibbs weight of the k^{th} cluster is $w_k \equiv \mu_N(\alpha_k) \gg \mu_N(\partial_\epsilon \alpha_k)$, the number of assignments outside clusters carry exponentially little probability mass, such that $w_k \approx \frac{\mathcal{Z}_{\alpha_k}}{\sum_{k \in [\eta_N]} \mathcal{Z}_{\alpha_k}}$, and the pure state decomposition would then imply:

$$\mu_N(\underline{y}) = \sum_{k \in [\eta_N]} w_k\, \mu_N^{\alpha_k}(\underline{y}) \approx \sum_{k \in [\eta_N]} \frac{\mathbf{1}\{\underline{y} \in \alpha_k\}}{\mathcal{Z}_{\alpha_k}} \cdot \frac{\mathcal{Z}_{\alpha_k}}{\sum_{k \in [\eta_N]} \mathcal{Z}_{\alpha_k}}.$$

For the overall procedure to be uniform, the probability of the random walk to reach an assignment in a given cluster needs to be: $\mathbf{P}_{S-SAT}[x_t \in \alpha_k] \approx w_k$, but without detailed balance w.r.t. μ, we have no guarantee of this being true. In fact, in (Wei et al., 2004) the authors show that certain assignments in $2-$SAT formulas have exponentially low probability to be found via a pure ($\mathbf{Pr}_{rw}[x_t = x_{t-1}^{(i)}] = 1/N,\ \forall i \in [N]$) random walk.

Moreover, the difference in the frequency of visiting certain clusters can be explained by the fact that, just after escaping a given cluster, that is, early on when reverting to the random walk phase, if the current state is not too far from the cluster, the biased random walk is very likely to lead right back to the cluster it just escaped from, and get locked once again in the SA phase, revisiting each state a second time.

5.5 Greed, energy plateaus and bias towards coreless assignments

5.5.1 Walk-SAT as a stochastic local search method

As previously discussed, it can be very useful to interweave energy-based moves in a random walk (RW) approach. In the Sample-SAT case, since Simulated annealing

(SA) accepts moves with probability $e^{-\beta\Delta E_{x\to y}}$, for low enough temperature, only $\Delta E \leq 0$ moves are accepted, in the clustering regime this leads to chaining SA moves, and the algorithm is then locked intermittently into SA/RW phases.

SA is a very general optimization method, often used as a blackbox regardless of the objective cost function, this generality however comes at a price. Firstly, there is no control on the maximum descent in energy $|\Delta E|$, moreover, any convergence guarantee to the target distribution is lost when switching between RW and SA moves.

Furthermore, any given solution sampled by Sample-SAT is either found completely via a RW, or completely via SA, and the p_{rw} parameter merely acts to control the uniformity/speed of solution-time trade-off in a *set* of samples, but does not control the selection of intermediate states towards finding a single solution.

More precisely, we make the following distinction in the selection criteria: the transition kernel of SA selects end states *globally*, i.e. from all immediately adjacent states, according to $\mathcal{U}(\{x_t : d(x_{t-1}, x_t) = 1\})$, while the RW part chooses the flipped variable uniformly from $\{\partial a : \ a \in \mathcal{C}_u(\underline{x}_t)\}$, where $\mathcal{C}_u(\underline{x}_t)$ is the set of unsatisfied clauses under the current assignment.

In this sense, the problem of reaching a satisfying assignment can be cast as a knapsack problem with the number of satisfied clauses $E(\underline{x}_t)$ as a cost function, and the variables involved in a randomly selected clause set of currently unsatisfied clauses as the candidate set: $\mathcal{K}_t = \partial a,\ a \sim \mathcal{U}(\mathcal{C}_u(\underline{x}_t))$, then the notion of *greedy* search becomes more intuitive. It is well known that greedy algorithms are sub-optimal, a classical example is the case of traveling salesman problem (Gutin et al., 2002).

Greed and *circumspection* in energy-descent, i.e. control of $\max|\Delta E|$, have been thoroughly studied in the theory of *stochastic local search* (SLS), while the combination of SA/RW of Sample-SAT does not lend itself to theoretical analysis. It is therefore advantageous to consider a modified version of Sample-SAT, cast as a SLS algorithm.

Algorithm 5: A Stochastic (local) search variant

```
Result: $\mathcal{S}ol \subset \mathcal{S}$
Initialization: $q \leftarrow p_{greed},\ T_{max} \leftarrow \max\_samples$;
$\mathcal{S}ol \leftarrow \emptyset$;
for $t = 1, ...T_{max}$ do
    $\underline{x}_0 \sim \mathcal{U}(\Sigma)$;
    $\mathcal{C}_u(\underline{x}_t) \leftarrow \emptyset$;
    while $\mathcal{C}_u(\underline{x}_t) \neq \emptyset$ do
        $a \sim \mathcal{U}(\mathcal{C}_u(\underline{x}_t))$;
        draw $q \sim \mathcal{U}[0,1]$ ;
        if $u \leq q$: then
            $j \leftarrow argmin_{i \in \partial a}\ f(\underline{x}_t, i) \equiv\ argmin_{i \in \partial a} \sum_{b \in \mathcal{F}} \left|\psi_b(\underline{x}^{(i)}_{t,a}) - \psi_b(\underline{x}_{t,a})\right|$;
            $\underline{x}_{t+1} \leftarrow \underline{x}^{(j)}_t$;
        else
            $i \sim \mathcal{U}(\partial a)$;
            $\underline{x}_t \leftarrow \underline{x}^{(i)}_t$;
        end
        $s_t \leftarrow \{a \in \mathcal{F} \backslash \mathcal{C}_u(t) :\ \psi_a(\underline{x}_t) = 1\ \}$;
        $\mathcal{C}_u(t) \leftarrow \mathcal{C}_u(t) \backslash\ s_t$;
    end
    $\mathcal{S}ol \leftarrow \mathcal{S}ol \bigsqcup \{\underline{x}_t\}$;
end
```

Note that, in the global search algorithms such as GSAT, we select the variable to be flipped from the entire set of indices $\{1 \dots N\}$ and not just from ∂a where a is uniformly selected from the set of unsatisfied clauses as we did in the algorithm above. In previous discussions, the cost function was counting the number of unsatisfied clauses under $\underline{x}_t$, $f(\underline{x}_t, i) \equiv \Delta E_{\underline{x}_t \to \underline{x}^{(i)}_t}$, but there are a number of other options. In Random-Walk-SAT, the precursor of Walk-SAT, the algorithm chooses from $\mathcal{K}_t = \partial a$, $a \sim \mathcal{U}(\mathcal{C}_u(\underline{x}_t))$, the variable which annuls the least number of previously satisfied clauses, the cost function is then given by $f(\underline{x}_t, i) \equiv \sum_{b \in \mathcal{F} \backslash\ \mathcal{C}_u(\underline{x}_t)} \left|\psi_b(\underline{x}^{(i)}_{t,a}) - \psi_b(\underline{x}_{t,a})\right|$.

Another popular choice is the *Focused Metropolis search* criterion which maximizes energy-descent $f(\underline{x}_t, i) \equiv \sum_{b \in \mathcal{C}_u(\underline{x}_t)} \left|\psi_b(\underline{x}^{(i)}_{t,a}) - \psi_b(\underline{x}_{t,a})\right| = \Delta E_{\underline{x}_t \to \underline{x}^{(i)}_t}$, for the rest of the discussion we will focus on this variant.

The greed parameter q allows us to interpolate from pure random walk at $q = 0$ to pure FMS at $q = 1$. In (Barthel et al., 2003), the authors show the existence of a phase transition in solution-times of $3-$SAT for a pure random walk search, that occurs at a critical point below the clustering threshold $\alpha_{exp} \approx 2.7 \leq \alpha_d = 3.92$, i.e. still in the replica symmetric regime, where the solution-time goes from linear to exponential in N.

Moreover, they show experimentally, as well as analytically, though non-rigorously (since they assume that the probability of annulling a satisfied clause by flipping a variable is variable-independent and time-independently equal to $p = 1/(2^k - 1)$), that below $\alpha_{exp} \approx 2.7$, a pure random walk finds a solution in linear time, and that above this threshold, the number of unsatisfied clauses $|\mathcal{C}_u(\underline{x}_t)|$ quickly goes down to reach a non-zero plateau around which fluctuations are exponentially rare, in this regime, $|\mathcal{C}_u(\underline{x}_t)|$ keeps fluctuating around an instance dependent (function of $\{J_{ai} : i \in \partial a\}_{a\in[M]}$) value, until an exponentially rare fluctuation is large enough that $|\mathcal{C}_u(\underline{x}_t)|$ becomes equal to zero, and the search terminates.

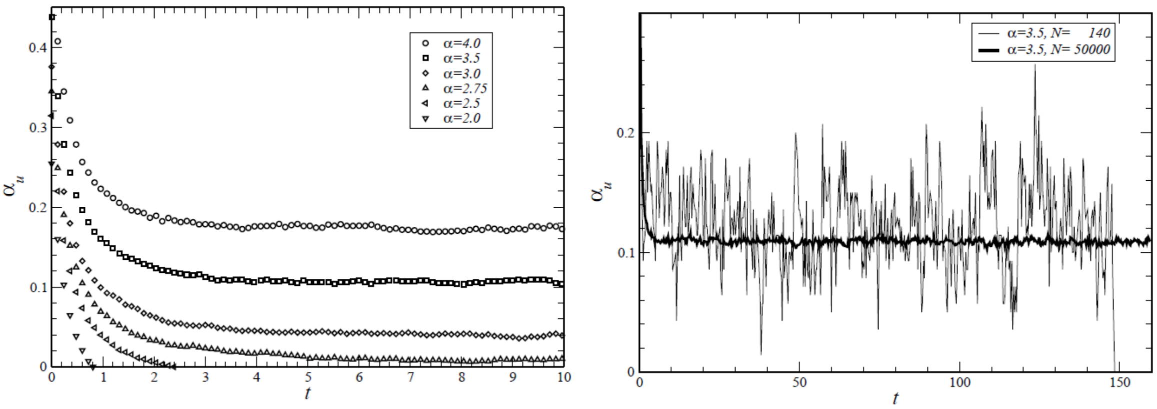


Figure 5.3: *(Left):* time-evolution of the normalized number of unsatisfied clauses $\alpha_u(t) \equiv \mathcal{C}_u(\underline{x}_t)/N$, and *(right)* concentration effect around for large N, from (Barthel et al., 2003).

For small sized systems, like the example in the left graph above: $N = 140$, a large enough fluctuation results in a solution after just 148 sweeps i.e. $\approx 2.10^4$ iterations, while for larger systems $N \approx 5.10^4$, such macroscopic fluctuations are exponentially rare, and a simple pure RW approach is thus not feasible for 3−SAT instance with $\alpha > \alpha_{exp}$.

However, if we allow greedy steps, i.e. set $q > 0$, then the authors observe that: (i). the energy plateau, around which $|\mathcal{C}_u(\underline{x}_t)|$ fluctuates, drops down to a lower non zero value with the same qualitative behaviour (rare fluctuation around the mean), and (ii). the experimental threshold α_{exp} is slightly pushed back from 2.7 to 2.8.

Interestingly, in (Alava et al., 2008), the authors show that if we take $q = 1$, i.e. follow a pure FMS search, then the search leads to all cover-less solutions, i.e. solutions whose only cover is the all $*$ trivial cover, for $k = 4, 5, 6$ and 7. These experiments seems to suggest, as is the case with the Survey propagation decimation algorithm, that overly greedy search is strongly biased towards cover-less solutions.

Moreover, the authors show that using circumspection, i.e. bounding the maximum allowed energy descent, in a pure FMS, results in $\mathcal{O}(N)$ solution-time almost surely, futhermore, this was shown to hold well beyond the clustering and condensation thresholds.

Contrasting these results with Sample-SAT who is shown to be able to find a sig-

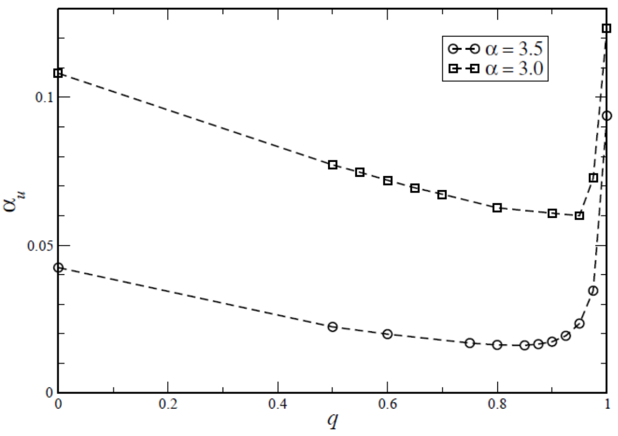


Figure 5.4: The effect of greed on lowering the energy plateau (Barthel et al., 2003).

nificant number of solutions with non-trivial true-covers simply through pure RW, leads us to another trade-off; by reducing greed, we trade efficiency (linear solution-time beyond the clustering and condensation thresholds) for a higher likelihood of finding solutions with non-trivial covers.

In an orthogonal direction, in (Schoning, 2002) the author proves that a pure random walk with restarts every $3N$ iterations when no solutions are found has a worst-case solution time of $\Omega(\frac{4}{3}^N)$. In other words, by just restarting $\underline{x}_0 \sim \mathcal{U}(\Sigma)$, we gain an exponential increase in the speed of finding solutions, moreover, we do not have to worry about cover-less solutions since $q = 0$.

5.6 Large moves in the state space

To explore combinatorial state spaces of general spin systems efficiently, (Hamze et al., 2013) developed an energy based self-avoiding random walk approach, which allows large moves in the state space, while satisfying detailed balance w.r.t. the target distribution μ_β (in their paper the authors considered a 2d Ising model).

As discussed in the case of k-SAT, the exponentially vanishing probability of escaping clusters $\mathcal{W}_{\alpha_k \to \partial_\epsilon \alpha_k}$ is such that the relaxation time in the clustering regime is exponential in N. Practically speaking, in the $2d$ Ising case, single-flip Glauber dynamics at low temperature take 10^{10} trials to leave a metastable state, i.e a deep local minimum (which would be the $\beta-$smoothed equivalent of a cluster in the $p-$spin case), this amounts to 10^{15} minutes of running time (Hamze et al., 2013).

In their paper, the authors propose an energy based proposal distribution which allows moves between states with Hamming distance larger than one, then show that the induced Markov chain is reversible and satisfies a stronger version of detailed balance which implies the standard case, hence proving convergence. Moreover, the energy based aspect of the proposal distribution's main use consists in landing in typical (i.e. low energy) states. However, as noted in their article, this approach becomes problematic when the solution space is highly disconnected (as is the case in k-SAT).

Indeed, by defining the proposal distribution as: $q(x_{t+1} = \underline{y} \mid \underline{x}_t) \propto \; e^{-\beta E(\underline{y})}$ (where E is the Hamiltonian of the target distribution μ_β), with a support consisting of all states with Hamming distance equal to l (the step length parameter) from the current state: $d_h(\underline{y}, \underline{x}_t) = l$, if we start from a state located in a low probability region of the state space, the proposal distribution will select the states with the lowest energy *relative to all those with Hamming distance equal to* l, even if they actually result in a low acceptance probability.

In the case of k-SAT beyond the clustering threshold, if we start with an assignment located in one of the bottlenecks separating clusters, this choice of proposal distribution might cause the algorithm to get stuck in local minima, by yielding states that have a high probability under $q(.|\underline{x}_t)$ but low acceptance probability $\alpha(., \underline{x}_t) \equiv \min\{1, e^{-\beta(E(.)-E(\underline{x}_t))}\}$.

5.6.1 Avoiding covers

Recall that the main problem with Sample-SAT, is the disparity in the number of visits among clusters. A heuristic way to solve this issue, is to keep track of the visited solutions during the run of the SA phase, and drop their values from the support of the random walk, this is however not practical for two main reasons:

a). Keeping track of all sampled solutions is memory-wise and computationally costly.

b). Even if we do drop the assignments visited in the first run of SA, clusters are very dense in solutions, such that the algorithm may still get locked into a second SA phase in the same cluster, visiting different solutions. And while the overall sampling will still be uniform, since the 1^{st} round solutions are dropped and all solutions are therefore visited once, the probability of reaching different clusters is however not equal and hence the exploration of the state space remains non-uniform.

We propose an algorithm that circumvents these problems, by alternating single-flip moves with much larger jumps in the state space. In addition, we use the 1RSB prediction of the existence of a bijection between clusters and true covers (Ding et al., 2015), to encode all solutions within a given cluster by their representative core variables assignment, and eliminate those from the space on which the random walk moves along.

5.6.2 The SAW-SAT algorithm

To recapitulate, our goal is uniform exploration of the set of satisfying assignments in the clustering regime. To this end, we surveyed a number of state of the art methods for finding solutions, and found that Survey propagation as well as greedy search (s.a. FMS) were found to be highly biased towards cover-less solutions (i.e. solutions whose cover is the trivial all-$*$ generalized assignment).

To our knowledge, the most uniform sampling method beyond α_d to date, seems to be Sample-SAT (Kroc et al., 2007). However, it has also been established that Sample-SAT seems to oversample from some clusters compared to others. Indeed, the experiments in (Wei et al., 2004) suggest that the reason for this is that right after terminating an SA phase, the random walk seems to lead right back to previously visited cluster to sample from it again.

To prevent this very issue, we enforce faster exploration of the state space by proposing a multi-flip Self-Avoiding-Walk (SAW), which allows us to avoid relapsing into clusters upon escape. Moreover, we experiment with the option of avoiding visited clusters by rejecting any end state whose true cover is that of their representative generalized assignments (as defined towards the end of ch.4).

Starting from an arbitrary initial assignment, SAW-SAT alternates between single-flip steps and larger self-avoiding moves, until it lands inside a cluster, where it gets locked into the SA phase. To detect this "phase change", we keep track of the ratio of (single-flip samples)/(SAW samples), above a threshold parameter that we denote by ρ, we suppose that the system has effectively entered the SA phase and proceed to extract the true cover that is representative of said cluster. Then, only after terminating the SA phase, if the extracted true cover is non-trivial (i.e. not the all-$*$ assignment), we drop it from the state space by rejecting any move with said true cover as an end state.

In order to extract true-covers we use *the peeling method* proposed in (Maneva et al., 2007), where we start by a satisfying assignment and iteratively set all of its unconstrained variables to $*$, recall that a variable x_i is constrained if there exists a clause where all variables but x_i are unsatisfying, or there are two variables in said clause assigned to $*$.

Suppose, we have a function $Constr(\Phi, \tau)$, which takes as input the $k-$SAT instance and a generalized assignment $\tau \in \{0, 1, *\}$ and returns the set of constrained variables, we can then obtain true covers from an arbitrary satisfying assignment as follows:

```
Algorithm 6: Peeling
  Input: The instance $\Phi$ and a satisfying assignment $\underline{x} \in \mathcal{S}$ ;
  Result: $\tau$, the true cover of the input solution.
  while $\exists i \in [N]$, $\tau_i \neq *$ such that $x_i$ is not a constrained variable do
      $\mathcal{C}str \leftarrow Constr(\Phi, \tau)$;
      $i \sim \mathcal{U}(\mathcal{C}str)$;
      $\tau_i \leftarrow *$;
  end
```

As discussed, the main idea of SAW-SAT is to alternate between single-flip steps and larger self-avoiding moves. We call a *SAW-path* of length p, the sequence of intermediate states obtained by iteratively flipping the set of indices $(i_1, \ldots, i_p) \in [N]^p$, where $i_k \neq i_l \; \forall i, k \in [p]^2$, e.g. the SAW-path of length three given by $(1, 3, 4)$ and starting from $\underline{x}_t = (0, 1, 1, 0)$ has $\underline{x}_{t+1} = (1, 1, 0, 1)$ as the end state.

Just like in the single flip case, we can either use an energy based approach, in which we would first propose the SAW-path by sampling without replacement p indices from $[N]$, then accept the move with probability $\min\{1, e^{-\beta\Delta E}\}$, or we could take the random walk (RW) approach where all SAW-moves are accepted with probability one.

Note that the latter case can be recovered by setting $\beta = 0$, such that we can cast the procedure in the more general energy based approach and use β_{saw}, that we differentiate from the single flip inverse temperature β_{sf}, as a greed parameter.

Algorithm 7: $SAW\_move(.)$

```
Input: $\underline{x}_{t-1}$, $\beta_{saw}$, $p$;
Result: $\underline{x}_t$;
draw $(i_1, \ldots, i_p)$ without replacement from $[N]$;
$\alpha \leftarrow \min\{1, e^{-\beta_{saw}\Delta E_{x_t \to z}}\}$ ;
draw $u \sim \mathcal{U}([0, 1])$ ;
if $u \leq \alpha$: then
    $\underline{x}_t \leftarrow z$;
else
    $\underline{x}_t \leftarrow \underline{x}_{t-1}$;
end
```

Furthermore, following (Schonning, 2002), when the exploration fails to find a satisfying assignment after $3N$ iterations, we restart from an arbitrary initial assignment, selected uniformly from the entire state space; $\underline{x}_0 \sim \mathcal{U}(\Sigma)$. The SAW-SAT algorithm is then as follows:

Algorithm 8: The SAW-SAT algorithm (with peeling and restarts)

```
Result: $\mathcal{S}ol \subset \mathcal{S}$
Input: $p_{rw}$, $\beta_{sf}$, $\beta_{saw}$, $max\_samples$;
$\mathcal{S}ol \leftarrow \emptyset$;
while $|\mathcal{S}ol| \leq max\_samples$ do
    $t \leftarrow 0$;
    $\rho \leftarrow 0$;
    $\Sigma_0 \leftarrow \Sigma$;
    $\underline{x}_0 \sim \mathcal{U}(\Sigma_0)$;
    while $t < 3N$ do
        draw $q \sim \mathcal{U}[0,1]$ ;
        if $q \leq p_{rw}$: then
            $\underline{x}_{t+1} \leftarrow Glauber\_step(\underline{x}_t, \beta_{sf})$;
            $g \leftarrow True$;
        else
            $\underline{x}_{t+1} \leftarrow SAW\_move(\underline{x}_t, \beta_{saw}, p)$;
            $g \leftarrow False$;
        end
        $t \leftarrow t+1$;
        if $E(\underline{x}_{t+1}) = 0$ then
            $\mathcal{S}ol \leftarrow \mathcal{S}ol \sqcup \{\underline{x}_{t+1}\}$;
            if $g == True$ then
                $\rho \leftarrow \rho + 1$;
            end
            if $\rho > \rho^*$ then
                $\tau \leftarrow Peeling(\Phi, \underline{x}_{t+1})$;
                $\Sigma_{t+1} \leftarrow \Sigma_t \backslash \{y \in \Sigma : y_i = \tau_i \; \forall \; i \in \mathcal{C}str(\Phi, \tau)\}$;
                $\rho \leftarrow 0$;
            end
            break;
        end
    end
end
```

5.6.3 Experiments and discussion

To get some insight on the role of the different pieces in the above algorithm, we have experimented with a number of variants, all of which use Schoning's restarts:

a). $p = 1,\ \beta_{saw} = \beta_{sf} = 0:$ i.e. a single-flip pure random walk search.

b). $p = 1,\ \beta_{saw} = \beta_{sf} = 100,$ which reduces to a pure Focused Metropolis search

c). $p = \lceil N/6 \rceil,\ \beta_{saw} = \beta_{sf} = 0:$ a pure single-flip random walk interweaved with $\lceil N/6 \rceil-$length SAW steps accepted with probability one, where $\lceil . \rceil$ is the ceiling function.

d). $p = \lceil N/6 \rceil, \ \beta_{saw} = \beta_{sf} = 100:$ a fully energy based local search approach which alternates between single-flip and $\lceil N/6 \rceil-$length SAW steps, all accepted with probability $\min\{1, e^{-100\Delta E}\}$.

e). $p = \lceil N/6 \rceil, \ \beta_{saw} = 0, \ \beta_{sf} = 100:$ a pure self-avoiding random walk interweaved with single-flip Glauber steps.

f). $p = \lceil N/6 \rceil, \ \beta_{saw} = 60, \ \beta_{sf} = 100:$ a less greedy version of *d*) in the acceptance of SAW moves.

Generating satisfiable $3-$SAT instances for $\alpha > 3$ is notoriously hard (Achlioptas et al., 2000), in fact just finding a satisfiable $3-$SAT instance with $N = 25, M = 95$ took us about 38 minutes in CPU time. Fortunately there is an extensive library of benchmark instances referenced in (Hoos H., 2006), from which we retrieved an instance (under the name of uf125-538) containing 125 variables and 538 clauses, i.e. at clause density of $\alpha = 4.3$, and then we subsequently sampled 492 clauses from the original 538, to get an instance just above the clustering threshold $\alpha \approx 3.93 > \alpha_d = 3.92$.

We start with the question of relative efficiency between the different variants. In this part, we are more interested in the speed of finding solutions than assessing the uniformity of the sampling procedure, and have therefore opted to do without peeling/avoiding covers since these result in a considerable slowdown of solution-times.

In fact, a single lookup of constrained variables using $Constr(\Phi, x)$, requires checking each of the αN clauses ($a \in [M]$) to count the number of unsatisfying variables in ∂a, which amounts $NMk = \mathcal{O}(N^2)$, experimentally, for the instance considered above, a sample of ten solutions found by pure FMS (*b*) (with peeling) yielded a median solution time of 43 seconds in CPU time, which is more than twice as slow as without peeling.

Some preliminary experiments confirmed the observation made in (Alava et al., 2008), concerning the divergence of solution-times (in the sense of not being concentrated around some mean value) at smaller sizes, which led us to consider the *median* (rather than mean) solution-time as a measure of efficiency to compare the different variants, and we found the following results:

Variants	Median solution-times (in sec. CPU time)
$p = 1, \ \beta_{saw} = \beta_{sf} = 0$	no convergence
$p = 1, \ \beta_{saw} = \beta_{sf} = 100$	16.7
$p = \lceil N/6 \rceil, \ \beta_{saw} = \beta_{sf} = 0$	no convergence
$p = \lceil N/6 \rceil, \ \beta_{saw} = \beta_{sf} = 100$	74.4
$p = \lceil N/6 \rceil, \ \beta_{saw} = 0, \ \beta_{sf} = 100$	496.1
$p = \lceil N/6 \rceil, \ \beta_{saw} = 60, \ \beta_{sf} = 100$	28.7

Note that; for a given variant, if the algorithm does not converge to a solution after ten minutes we terminate, this was the case for both non-energy based variants (*a*) and (*c*).

Moreover, when looking at the evolution of the number of unsatisfied assignments $|\mathcal{C}_u(t)|$ for variants (*b*), (*d*) and (*f*). we observed the same behaviour noted in (Barthel et al., 2003), where $|\mathcal{C}_u(t)|$ descends quickly until it reaches an energy plateau, which it leaves upon a Schoning restart or an unlikely deviation.

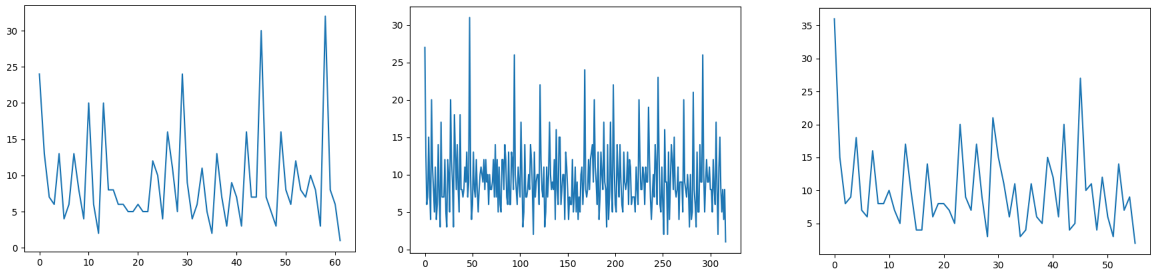


Figure 5.5: The evolution of the number of unsatisfied clauses every 500 iterations for b), d), f). (with restarts) from left to right.

In order to examine the role of the SAW moves in escaping the energy plateau, we also recorded the evolution of $|\mathcal{C}_u(t)|$ *without restarts* for pure FMS (b) vs the energy based variant (f) of SAW-SAT, and found that interweaving energy based SAW moves in a pure FMS has the effect of slowing down the arrival to an energy plateau, and producing the unlikely deviation discussed in (Barthel et al., 2003) which results in finding a solution, and which does not happen in the case of pure FMS search which stays on the low energy plateau even after 600 iterations, as can be seen in the graph below.

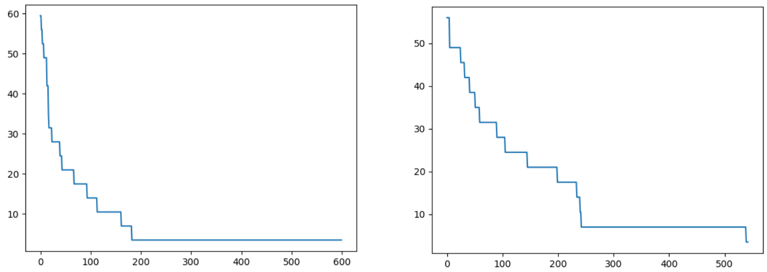


Figure 5.6: The evolution of the number of unsatisfied clauses every iteration for pure FMS (left) and energy based SAW e).(right), both without restarts.

Lastly, to compare the uniformity of the different sampling procedures, for each of $(b), (d)$ and (e), we try 2 versions; one which peels above-threshold solutions (i.e. those found when $\rho > \rho^*$) to their true covers and drops them from the state space Σ_t, and another version which skip this part altogether. We sample 200 solutions and count the number of duplicate solutions and find the following

Furthermore, keeping track of the number of non-trivial covers when sampling (without peeling), we observed that all solutions found via pure FMS (b) have zero non-trivial covers, which confirms the findings in (Alava et al., 2008), and explains why the number of duplicates does not improve when switching from sampling *with* peeling to without. The number of duplicates actually slightly increases slightly in this case, but this is just due to the randomness in the FMS.

Variants	Duplicates (with peeling)	without
$p = 1,\ \beta_{saw} = \beta_{sf} = 0$	23	25
$p = \lceil N/6 \rceil,\ \beta_{saw} = 0,\ \beta_{sf} = 100$	3	2
$p = \lceil N/6 \rceil,\ \beta_{saw} = 60,\ \beta_{sf} = 100$	0	0

Table 5.1: Number of duplicate solutions with peeling vs. without.

As for variants (e) and (f), we found that 34 and 22 solutions, respectively, have a non-trivial cover, which makes about 17% and 11% resp. of the total sampled solution, which is a bit less than expected considering the experiments in (Kroc et al., 2007) where the authors find that about 26% of the sampled solutions have non-trivial covers. However, a closer look at these covers reveals that while the underlying solutions are non identical most of their covers are (that we compute through peeling). In fact for both these samples we have only found 3 non-trivial covers, which may explain why the uniformity of sampling did not improve by avoiding true covers.

Conclusion and future work

Conclusion

To conclude, we find that SAW steps with $\beta_{saw} = 60$ significantly improve upon pure single flip FMS in the uniformity of sampling, while still keeping a very competitive solution time (28 sec vs. 16 for FMS) when contrasted with single flip random walk approaches which do not seem to converge in less than ten minutes. Moreover, we observed the same behaviour w.r.t. energy plateaus in (Barthel et al., 2003) and concluded that energy based SAW steps bring about the unlikely deviation which leads to finding a solution, which is not the case in pure FMS search without restarts.

Furthermore, we note that while SAW steps significantly decrease the number of duplicate solutions when compared to pure FMS, we unfortunately find that avoiding covers does not result in any further improvement, mostly due to the very small number of true covers in the sampled solutions, the decrease in the number of duplicates in SAW-SAT may be explained by the fact that larger moves in the state space get rid of the cluster-relapse phenomenon observed in Sample-SAT.

Interesting avenue for future work: Large iso-energetic moves in the state space

Another promising avenue for overcoming high energy barriers and faster exploration of the state space in general spin systems, is the use of symmetries in the energy function. More specifically, the search for mappings $\mathcal{M} : \Sigma \mapsto \Sigma$ under which the energy function stays invariant: $E \circ \mathcal{M}(\tau) = E(\tau)$.

The cluster algorithm (Houdayer, 2001), was originally designed for the $2d$ Ising model, and makes use of this simple idea to propose an algorithm that is able to make large *iso-energetic* moves in the state space, and hence improve significantly on previous algorithms in the speed of exploration of the state space. Let $\mathcal{J}$ consist of nearest neighbors interactions on the 2 dimensional grid, and suppose we sample $\sigma^1, \sigma^2 \overset{iid}{\sim} \mu_\beta(.) \propto e^{-\beta \mathcal{H}(.)}$, where $\mathcal{H}(\sigma) \equiv -\sum_{\mathcal{J}} J_{ij}\sigma_i\sigma_j - \sum_{i\in[N]} h_i\sigma_i$.

Moreover, let $\mathcal{R} \equiv (\sigma^1, \sigma^2)$ be the replicated system, and $q_i(\mathcal{R}) \equiv \sigma_i^1\sigma_i^2$ be its *local overlap*, then, given an assignment $\underline{\mathcal{R}}$, the $[N]$ spin sites are split into two disjoint subsets; ones where spins have the same value and others where they differ: $\mathcal{C}^{\sim} \equiv \{i : \underline{\sigma}_i^1\underline{\sigma}_i^2 = 1\}$, $\mathcal{C}^{\nsim} \equiv \{i : \underline{\sigma}_i^1\underline{\sigma}_i^2 = -1\}$, where $\mathcal{C}^{\sim} \bigsqcup \mathcal{C}^{\nsim} = [N]$.

Clusters (in the sense of Houdayer) are connected component in Hamming space in either $\mathcal{C}^{\sim}$ or $\mathcal{C}^{\nsim}$. More precisely, given an assignment $\underline{\mathcal{R}}$ and the resulting partition of spin sites between $\mathcal{C}^{\sim}, \mathcal{C}^{\nsim}$, the algorithm goes as follows:

i). we sample $k \sim \mathcal{U}(\mathcal{C}^{\nsim})$,

ii). we flip the cluster containing the $\underline{k}^{th}$ in both systems σ^1 and σ^2, i.e. the flip all variables connected to $\sigma_k^{1(2)}$ through a chain of interactions $J_{jl} \rightarrow J_{lp} \rightarrow J_{pk}$ between spins $(l, p, \ldots, k) \in \mathcal{C}^{\nsim}$.

It is then easy to verify that the probability law of the replicated system $\mathbf{P}[\sigma^1, \sigma^2] \propto e^{\beta[\mathcal{H}(\sigma^1)+\mathcal{H}(\sigma^2)]}$ is invariant to a cluster move (ii), and therefore that the two-step algorithm proposed above satisfies detailed balance w.r.t. μ_β. Note, that since cluster moves are large (> 1) the resulting algorithm is not ergodic, in the sense that not all states in Σ are accessible with positive probability, hence the need to alternate cluster moves with single-flip Glauber updates.

For even faster mixing, cluster moves (along with single-flip Glauber moves) are interweaved with *Exchange Monte Carlo* updates more commonly known as *Parallel tempering* which originated from (Swendsen and Wang, 1986), where we run the algorithm in parallel for m independent systems or *replicas* each at a different temperature $\{\beta_l : l \in [m]\}$, and exchange states between $l-$neighboring replicas with probability $\alpha_{\sigma^{l-1} \leftrightarrow \sigma^{l-1}} \equiv \min\left\{1, e^{(\beta_l - \beta_{l-1})[\mathcal{H}(\sigma^l) - \mathcal{H}(\sigma^{l-1})]}\right\}$.

By the time the Houdayer's algorithm was proposed, Parallel tempering (PT) was a staple ingredient for faster mixing in energy landscapes with high energy barriers, and while the acceptance probability of PT is such that the exchanged pair' states are close in energy, their Hamming distance is typically $\Theta(N)$, and therefore, when clusters become too large, cluster moves reduce to PT updates and thus become redundant.

The extent to which cluster moves improve upon PT (alternated with Glauber updates), then depends heavily upon the typical size of clusters, more precisely, if the bond-percolation threshold of the factor graph associated with the model's Hamiltonian is $p_{perc} < 0.5$, which as Houdayer explains, makes the cluster algorithm unusable for the Ising model with $d \geq 3$.

Going back to random CSPs, in (Alon et al., 2004) the authors prove that random $d-$regular graphs display a phase transition at a critical bond-percolation threshold given by $p_{perc} = 1/(d-1)$ where the size of the giant component goes from $\mathcal{O}(\log N)$ to $\Theta(N)$. Cluster moves are thus system-sized for CSPs with $d \geq 3$.

Nonetheless, some distinctions between $k-$SAT problems and the Ising model, need to be made. While in the latter case, a replicated system's probability $\mathbf{P}[\mathcal{R}]$ is invariant to a cluster moves, this is not true in the $k-$SAT case, more precisely, consider two satisfying assignments x, y of the same $k-$SAT formula, if there exists a variable $i \in \mathcal{C}^{\nsim}$ such that either x_i (or y_i) is a constrained variable (i.e. the only satisfying assignment in some clause a), then the cluster move will yield a unsatisfying assignment for the system with the constrained variable.

However, one could suggest considering the subset of solutions whose negation $-\underline{x}$ are also in $\mathcal{S}$. The problem of finding such solutions is more commonly known as the *not all equal* k-SAT problem or $k-$NAESAT, and requires, for an assignment to be considred a solution, that:

$i)$. Each clause $a \in [M]$ contains at least one variable $i \in \partial a$ such that $\underline{x}_i = 1 - J_{ai}$ as in the $k-$SAT case.

$ii)$. And comes with the further requirement that: (ii). each clause has not all equal truth values, i.e. that $\forall a \in [M],\ i, j \in \partial a$ such that $J_{ai}\underline{x}_i \neq J_{aj}\underline{x}_j$.

Then by the second condition, any $k-$NAESAT solution remains in $\mathcal{S}$ under negation. Unfortunately, for $\alpha > 2^{k-1} \log 2$, the set of NAE-solutions is empty w.h.p. (Coja-Oghlan and Panagiotou, 2012), for 3$-$SAT, not-all-equal solutions vanish even below the clustering threshold at $\alpha_{NAE} \approx 1.2 < \alpha_d = 3.92$. Which leads us to consider other mappings than $\mathcal{M}(x) \equiv -x$ under which $E \circ \mathcal{M}(\tau) = E(\tau)$.

One interesting avenue for such an endeavor is the use of basic group theory to find not so obvious symmetries specific to a given $k-$SAT instance. In (Aloul, 2010), the author surveys algebraic approaches to the $k-$SAT problems, and the way in which symmetries induce equivalence classes between assignments related by different mappings, which permit to extend a sample of solutions to a larger set.

This approach is in a way orthogonal to the 1RSB picture, but it would be interesting to see what we can gain by relaxing the condition that all variables must be related in such symmetries, and look for mappings of *subsets of variables*, such that $E(\mathcal{M}(x_{i_1}, \ldots, x_{i_k}), x_{l \in [N] \setminus \{i_1, \ldots, i_k\}}) = E(x)$.

References

Simons B. 1997. Lecture notes ”Phase Transitions and Collective Phenomena”.

Castellani T.; Cavagna A. 2005. Spin-Glass Theory for Pedestrians, cond-mat/0505032

Ben Arous G.; Cerny J. 2005. Dynamics of trap models, Les Houches Summer School “Mathematical statistical physics”.

Mezard M.; Parisi G.; Virasoro M.A. 1987. Spin glass theory and Beyond, World Scientific.

Fischer K. H.; Hertz J. 1991. Spin Glasses, Cambridge University Press.

Mezard M.; Montanari A. 2009. Information, Physics and Computation, Oxford University Press.

Binder K.; Young A.P. 1986. Spin glasses: Experimental facts, theoretical concepts, and open questions, Rev. Mod. Phys. 58, 801.

Claudius G. 2017. Lecture notes ”Thermodynamik Statistische Mechanik”, URL: https://itp.uni-frankfurt.de/ gros/Vorlesungen/TD/5_Thermodynamic_potentials.pdf

Binder K.; Billoire A.; Hartmann A.; Henkel M.; Wolfhard J. 2008. Rugged Free Energy Landscapes, Lecture Notes in Physics.

Fischer K. H.; Hertz J. 1991. Spin Glasses, Cambridge University Press.

Mezard M.; Montanari A. 2009. Information, Physics and Computation, Oxford University Press.

Castellani T.; Cavagna A. 2005. Spin-Glass Theory for Pedestrians, cond-mat/0505032

Mezard M.; Parisi G.; Virasoro M.A. 1987. Spin glass theory and Beyond, World Scientific.

Zia R.K.P.; Redish E.F.; McKay S.R. 2009. Making sense of the Legendre transform, American Association of Physics Teachers.

Edwards S.F.; Anderson P.W. 1975. Theory of spin glasses, J. Phys. F: Met. Phys. 5 965.

Fischer K. H.; Hertz J. 1991. Spin Glasses, Cambridge University Press.

Binder K.; Billoire A.; Hartmann A.; Henkel M.; Wolfhard J. 2008. Rugged Free Energy Landscapes, Lecture Notes in Physics.

Mezard M.; Parisi G.; Virasoro M.A. 1987. Spin glass theory and Beyond, World Scientific.

Hartmann A.K.; Weigt M. 2005. Phase Transitions in Combinatorial Optimization Problems: Basics, Algorithms and Statistical Mechanics, Wiley-VCH.

Mezard M.; Montanari A. 2009. Information, Physics and Computation, Oxford University Press.

Sherrington D.; Kirkpatrick S. 1975. Solvable Model of a Spin-Glass, Phys. Rev. Lett. 35, 1792.

Derrida B. 1980. The Random Energy Model, Review Section of Physics Letters 67, No. 1.

Touchette H. 2009. The large deviation approach to statistical mechanics, Phys. Rep. 478, 1-69, 2009.

Talagrand M. 2011. Mean Field Models for Spin Glasses Volume I: Basic Examples, Springer.

Bollobás B. 2001. Random graphs, Cambridge University Press.

Mezard M.; Montanari A. 2009. Information, Physics and Computation, Oxford University Press.

Bender E.; Canfield E.R. 1978. The asymptotic number of labeled connected graphs with a given degree sequence, J. Comb. Theory.

Aldous D.; Steele M.J. 2003. The Objective Method: Probabilistic Combinatorial Optimization and Local Weak Convergence, Probability on Discrete Structures.

Flajolet P.; Sedgewick R. 2008. Analytic Combinatorics, Cambridge University Press.

Dembo A.; Montanari A. 2010. Ising models on locally tree-like graphs, Annals of Applied Probability 2010, Vol. 20, No. 2, 565-592.

Montanari A.; Semerjian G. 2006. Rigorous Inequalities between Length and Time Scales in Glassy Systems, J. Stat. Phys. 125, 23.

Jordan M.I.; Wainwright M.J., 2008. Graphical Models, Exponential Families, and Variational Inference, Foundations and Trends in Machine Learning.

Ding J.; Nike S.; Sly A. 2015. Proof of the Satisfiability Conjecture for Large k, Proceedings of the forty-seventh annual ACM symposium on Theory of Computing.

Braunstein A.; Mezard M.; Zecchina R. 2005. Survey propagation: an algorithm for satisfiability, Random Structures and Algorithms 27, 201-226.

Maneva E.; Mossel E.; Wainwright M.J. 2007. A New Look at Survey Propagation and its Generalizations, Journal of the ACM.

L. Kroc L.; Sabharwal A.; Selman B. 2007. Survey Propagation Revisited, UAI.

Mezard M.; Montanari A. 2009. Information, Physics and Computation, Oxford University Press.

Montanari A.; Semerjian G. 2006. Rigorous Inequalities between Length and Time Scales in Glassy Systems, J. Stat. Phys. 125, 23

Braunstein A.; Zecchina R. 2004. Survey propagation as local equilibrium equations, J. Stat. Mech., P06007.

Braunstein A.; Mezard M.; Zecchina R. 2005. Survey propagation: an algorithm for satisfiability, Random Structures and Algorithms 27, 201-226.

Ding J.; Nike S.; Sly A. 2015. Proof of the Satisfiability Conjecture for Large k, Proceedings of the forty-seventh annual ACM symposium on Theory of Computing.

Maneva E.; Mossel E.; Wainwright M.J. 2007. A New Look at Survey Propagation and its Generalizations, Journal of the ACM.

L. Kroc L.; Sabharwal A.; Selman B. 2007. Survey Propagation Revisited, UAI.

Brooks S.; Gelman A.; Jones G.; Meng X. 2011. Handbook of Markov Chain Monte Carlo, Chapman Hall/CRC Handbooks of Modern Statistical Methods, 1st Edition.

Wei W.; Erenrich J.; Selman B. 2004. Towards Efficient Sampling: Exploiting Random Walk Strategies, AAAI.

Papadimitriou C.H. 1991. On Selecting a Satisfying Truth Assignment, Proceedings of the Conference on the Foundations of Computer Science, pages 163-169.

Zhang Y. 2017. Phase Transitions of Random Constraints Satisfaction Problem, PhD thesis, University of California, Berkeley.
Franz S.; G Parisi G. 1995. Recipes for metastable states in spin glasses, Journal de Physique I 5 (11), 1401-1415.

Silvio Franz S. 2006. Metastable states, relaxation times and free-energy barriers in finite-dimensional glassy systems, EPL (Europhysics Letters) 73 (4), 492.

Hamze F.; Wang Z.; Freitas N.D. 2013. Self-Avoiding Random Dynamics on Integer Complex Systems, ACM Transactions on Modeling and Computer Simulation, Vol. 23, No. 1, Article 9.

Gutin G.; Yeo A.; Zverovich A. 2002. Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP, Discrete Applied Mathematics Volume 117, Issues 1–3, 15 March 2002, Pages 81-86.

Barthel W.; Hartmann A.K.; Weigt M. 2003. Solving satisfiability problems by fluctuations: The dynamics of stochastic local search algorithms, Phys. Rev. E 67, 066104.

Schoning T. 2002. A probabilistic algorithm for k-SAT and constraint satisfaction problems, IEEE.

Alava M.; Ardelius J.; Aurell E.; Kaski P.; Krishnamurthy S.; Orponen P.; Seitz S. 2008. Circumspect descent prevails in solving random constraint satisfaction problems, PNAS.

Houdayer J. 2001. A cluster Monte Carlo algorithm for 2-dimensional spin glasses, Eur. Phys. J. B 22, 479–484.

Alon N.; Benjamini I.; Stacey A. 2004. Percolation on finite graphs and isoperimetric inequalities Ann. Probab. Volume 32, Number 3, 1727-1745.

Swendsen R.H.; Wang J. 1986. Replica Monte Carlo Simulation of Spin-Glasses, Phys. Rev. Lett. 57, 2607.

Aloul F.A. 2010. Symmetry in Boolean Satisfiability, Symmetry.

Achlioptas D.; Gomes C.; Kautz H.; Selman B. 2000. Generating Satisfiable Problem Instances, AAAI.

Hoos H. 2004. Uniform Random-3-SAT, url: https://www.cs.ubc.ca/ hoos/SATLIB/ Benchmarks/SAT/RND3SAT/descr.html