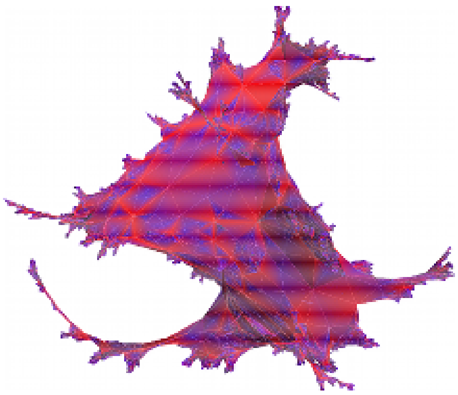
Lecture Notes, Statistical Physics for Optimization & Learning

Course held by Florent Krzakala, Lenka Zdeborová

Giancola Simone Maria[1†]

[†]Bocconi University, Milan

October 22, 2022



[1]simonegiancola09@gmail.com

Abstract

Statistical Physics can be seen as a set of theoretical results and methods to describe and tackle the computational hurdles of large inference problems. Building on the great contributions from the thermodynamics and statistical mechanics worlds, one can show that the same limiting properties apply for models spurring out of the two fields into computer science, physics, and machine learning. Such a formalism allows to draw similarities of solutions across different questions.
The document is a redaction of lecture notes from the homonymous PhD course offered at *École Polytechnique Fédérale de Lausanne* by Professors Krzakala Florent and Zdeborová Lenka [KZ21a]. While it mostly follows the videos and the lecture notes [KZ21b], it gives a different (less experienced, but self developed) structure, which is the result of autonomous understanding of the concepts explained.

Disclaimer 1 The document is subject to major updates. There are 50 **TODO** sections with potential expansions or missing proofs!

Disclaimer 2 Chapters are in the order of teaching. Images are entirely taken from the lectures, and come from different sources, please refer to the main source [KZ21a], to find the exact origins. Hopefully, it will be fixed in future versions.

Contents

List of Figures

List of Theorems

List of Symbols

$\mathbf{s}$	spins
$\mathscr{H}$	hamiltonian
$\mathbb{P}$	probability
β	inverse temperature
N	system size
$Z_N(\beta)$	partition sum
$H(\cdot)$	binary entropy
$\Phi_N(\beta, h)$	free entropy density
$I(m)$	large deviation rate
$\delta(x)$	dirac delta
$R_{s_j}^{j\to a}, V_{s_i}^{a\to i}$	auxiliary partition functions
$\chi_{s_j}^{j\to a}, \psi_{s_i}^{a\to i}$	messages
$s(\cdot)$	entropy
$\widehat{x}(\mathbf{y})$	estimator
$\mathcal{L}(\cdot, \cdot)$	error function
$\mathcal{R}^{av}(\cdot)$	averaged risk
$\mathcal{R}^{bayes}(\widehat{x})$	Bayes risk
$H(\cdot)$	statistical entropy
$D_{KL}(\cdot\|\|\cdot)$	Kullback-Leibler divergence
$I(\cdot, \cdot)$	mutual information
λ	SNR
A_N	Wigner matrix
$S_A(\cdot)$	Stieltjes transform

Chapter 1

Introduction

The best theory is inspired by practice. The best practice is inspired by theory.

Donald Knuth, 1974 Turing Award

The following lecture notes are a learning path through the course *"Statistical Physics for Optimization and Learning"* offered at EPFL. As a disclaimer, they will **not** cover:

- practical guides to machine learning or optimization
- applications of machine learning in a scientific landscape
- deep learning

Instead, the focus will be on:

- learning a threotical statistical physics-inspired approach to machine learning and optimization problems.
- probabilistic arguments and derivations to support claims
- exercises

References: most of the content follows the Lecture Notes for the course [KZ21b], but also the book by M. Mezard & A. Montanari [MM09] and the manuscript by the instructors [ZK16] are worth exploring.

In this chapter, the three main topics will be presented. Given the nature of the introductory lecture, it will be mostly **informal**, and built through instructive examples.

1.1 Graph Coloring

Consider the problem of building a map, where each country is assigned a color. Obviously, it is beneficial to color adjacent regions differently. We can be treat it with a well established mathematical framework.

Definition 1.1 (Planar Graph)**.** A planar graph is a graph that can be embedded in $\mathbb{R}^2$ without crossing edges.

Definition 1.2 (Graph Coloring Problem)**.** Consider a graph with N nodes indexed by $i = 1, \ldots, N$. If i is connected to j then we write $i \leftrightarrow j$. In the Adjacency matrix $A \in \mathbb{R}^N \times \mathbb{R}^N$ trivially set $A_{ij} = A_{ji} = 1 \iff i \leftrightarrow j \ \forall i, j$. On this graph, the aim is to assign a color for each vertex denoted as $s_i \in \{1, \ldots, q\}$, where q is the number of colors.

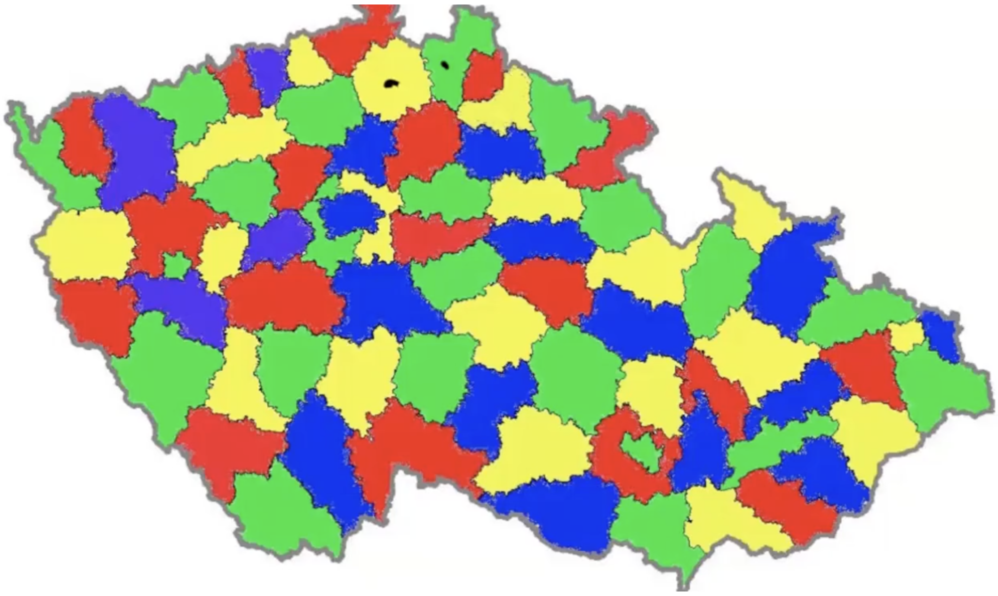


Figure 1.1: Map coloring

Given an adjacency matrix representing the graph, the latter is colorable if and only if:

$$\exists s^* : \mathscr{H}_A(s^*) = 0 \quad where \quad \mathscr{H}_A(s) = \sum_{i<j} A_{ij}\delta_{s_i,s_j} \; s = \{s_i\}_{i=1}^n \tag{1.1}$$

$$\delta_{s_i,s_j} = \begin{cases} 1 \; s_i = s_j \\ 0 \; otherwise \end{cases} \tag{1.2}$$

A very famous result, conjectured by Sir Francis Guthrie in 1852 and eventually proved by [AH76], is informally stated below.

Theorem 1.3 (Graph Coloring Theorem)**.** *For a planar map,* 4 *colors are sufficient to make all regions distinguishable.*
Equivalently, 4 colors are sufficient to solve the graph coloring problem for a planar graph. Figures 1.1 and 1.2 graphically show the equivalence of the two statements.

In this context, instead of considering planar graphs, which are well understood, we will discuss random graphs, which come with interesting properties in the large size limit as $N \to \infty$.

Definition 1.4 (Random Graph)**.** Given a set of vertices, build a random adjacency matrix, inducing a random set of connections as:

$$\forall i < j \; A_{ij} = \begin{cases} 1 & w.p. \dfrac{c}{N} \\ 0 & otherwise \end{cases} \tag{1.3}$$

In such a setting, a Statistical Physics perspective can be implemented observing that:

- $N \to \infty$ is a thermodynamic limit
- $\{s_i\}_{i=1}^N$ are Potts Spins
- $\mathscr{H}_A(s)$ is the energy configuration of the system for an assignment (the Hamiltonian)
- It is possible to build a Boltzmann measure to the various assignments as:

$$\mathbb{P}(s|A) = \frac{1}{Z_A(\beta)} e^{-\beta \mathscr{H}_A(s)} \tag{1.4}$$

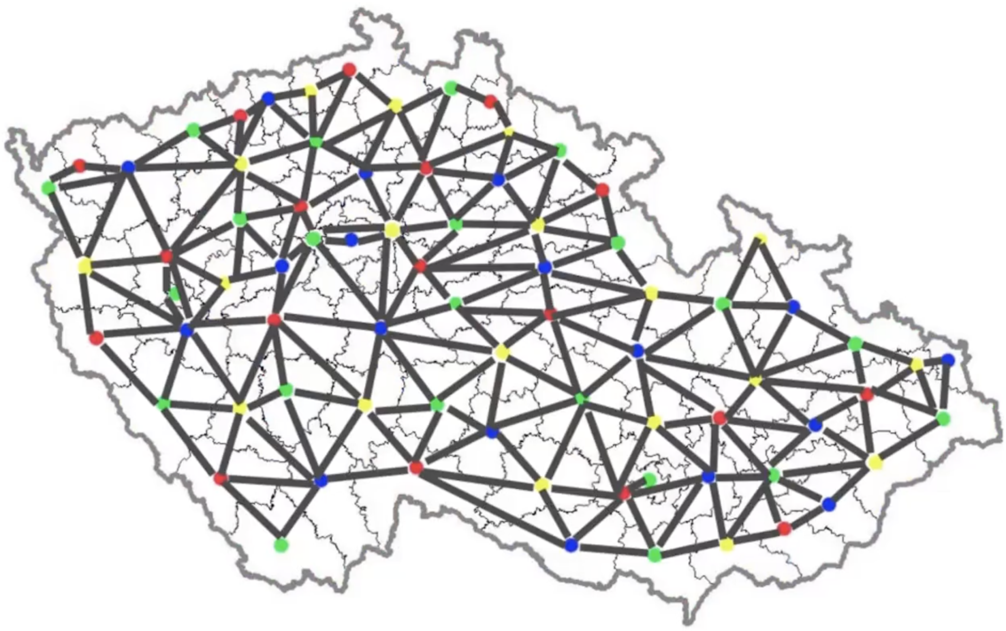


Figure 1.2: Graph coloring

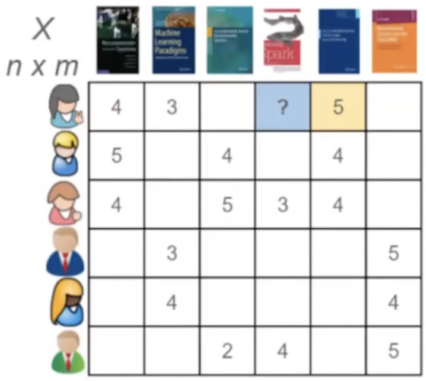


Figure 1.3: Recommendation matrix X

Where $\beta = \dfrac{1}{T}$ is the inverse temperature.

1.2 Recommendation Systems

As a motivating example, there is a list of n users and m books that combined form a matrix X of ratings. Since it is not necessarily true that everyone reads all the books it can have empty entries (see Figure 1.3 for an example).

In such a setting, consider a function that takes a matrix $U \in \mathbb{R}^n \times \mathbb{R}^k$, a matrix $V \in \mathbb{R}^k \times \mathbb{R}^m$ where k is small, and a vector ξ that introduces a randomness component in their multiplication. Assume that the rating matrix can be estimated through f as:

$$X_{ai} = f\left(\sum_{r=1}^{k} u^*_{ak}\nu^*_{ki} + \xi_{ai}\right) \tag{1.5}$$

A low rank matrix estimation task aims to find candidates of u^* and ν^* to resemble X as closely as possible, where the rank of the UV matrix is k.

Under certain randomness assumptions, the limiting setup such that: $m, n \to \infty, \frac{m}{n} \in O(1)$ with k small has a best achievable error bound.

The methodology analyzed will be approximate message passing (AMP), which is optimal for polynomial time computations. Also cases where it does not achieve the optimal error will be considered.

In substance, all of the above will be encapsulated in a physics framework, considering the importance and interpretation of phase transitions between states of matter adapted to the problem.

1.3 Generalized Linear Regression

In an image classification task of cats and dogs, a computer reads an image F_μ as a set of bits representing pixels' colors across the image. Thus, a learning process can be interpreted as finding a function f_w such that:

$$f_w(F_\mu) = y_\mu \quad y_\mu = \pm 1 \tag{1.6}$$

A regression task can come with many layers of complexity. Among the most explicative ones we find:

$f_w(F_\mu) = w \cdot F_\mu$	easy linear regression
$f_w(F_\mu) = \varphi(w \cdot F_\mu)$	generalized linear regression
$f_w(F_\mu) = \varphi^{(L)}\left(W^{(L)}\varphi^{(L-1)}(W^{(L-1)}\dots(W^{(2)}\varphi^{(1)}(W^{(1)}F_\mu))\right)$	L layer neural net

Where φ are in general activation functions.
[HSW89; Cyb89; Bar93; Bar94], but the second case is useful to ligthen up the analysis.

Under randomness, two networks interacting will be analyzed:

- **Teacher Network:** generates a dataset F from $i.i.d$ input vectors, and a weights w^*. With these two, a set of labels y is generated.
- **Student Network:** only observes $F \in \mathbb{R}^n \times \mathbb{R}^p, y$, without knowing w^*.

The aim of the analysis will be determining a closed formula for the best achievable generalization error in the high dimensional limit as $p, n \to \infty, \frac{n}{p} \in O(1)$. Also a comparison against empirical risk minimization will be carried out.

Chapter 2

A primer on Statistical Mechanics

The purpose of this chapter is to present a quick and effective overview of the statistical mechanics concepts needed to face the course content.

2.1 The Problems

Two types of problems will be discussed:

- Sampling problems
- Minimization Problems

A **minimization problem** can be interpreted as the task of finding an optimal configuration for a given function. Depending on the context, this could be a cost $\mathscr{C}$, a loss $\mathscr{L}$, or a Hamiltonian $\mathscr{H}$. In the end, they are all the same. The approach is that of assigning a Boltzmann measure that comes with nice probabilistic properties, with a Physical fashion. For this reason, the Hamiltonian $\mathscr{H}$ is preferred. To state a reference system, some jargon needs to be introduced.

Definition 2.1 (Boltzmann Distribution notation)**.** Given a space of configurations of the system $\{\mathbf{s}\}$, where $\mathbf{s} = \vec{s}$ in some sources, and a Hamiltonian function $\mathscr{H} : \{\mathbf{s}\} \to \mathbb{R}$, we denote as Boltzmann distribution:

$$\mathbb{P}(\mathbf{s}) = \frac{1}{Z_N(\beta)} e^{-\beta \mathscr{H}(\mathbf{s})} \tag{2.1}$$

Where $Z_N(\beta)$ is a normalization factor, often named **Partition sum**

$$Z_N(\beta) = \sum_{\{\mathbf{s}\}} e^{-\beta \mathscr{H}(\mathbf{s})} \tag{2.2}$$

And $\beta = \frac{1}{T}$ is the inverse temperature.
When computing averages with respect to this measure, they will be denoted as:

$$\langle \cdot \rangle_\beta = \frac{1}{Z_N(\beta)} \sum_{\{\mathbf{s}\}} (\cdot) e^{-\beta \mathscr{H}(\mathbf{s})} \tag{2.3}$$

With these simple objects in hand, it is already possible to notice that this custom measure has a peculiar property at the limit.

Proposition 2.2 (Low temperature minimum energy configuration convergence)**.** *As* $\beta \to \infty \iff T \to 0$ *it holds that:*

$$\lim_{\beta \to \infty} \Bigg(- \frac{\partial log(Z_n(\beta))}{\partial \beta} \Bigg) = \lim_{\beta \to \infty} \Bigg(\langle \mathscr{H}(\boldsymbol{s}) \rangle_\beta \Bigg) = \min_{\{\boldsymbol{s}\}} \Bigg\{ \mathscr{H}(\boldsymbol{s}) \Bigg\} \tag{2.4}$$

Proof. First of all, it is easy to notice that:

$$\begin{aligned}
\frac{-\partial log(Z_n(\beta))}{\partial\beta} &= \frac{1}{Z_n(\beta)}\frac{\partial Z_n(\beta)}{\partial\beta} && \text{basic derivation}\\
&= -\frac{1}{Z_n(\beta)}\frac{\partial(\sum_{\{\mathbf{s}\}} e^{-\beta\mathscr{H}(\mathbf{s})})}{\partial\beta} && \text{by Definition}\\
&= -\frac{1}{Z_n(\beta)}\sum_{\{\mathbf{s}\}} -\mathscr{H}(\mathbf{s})e^{-\beta\mathscr{H}(\mathbf{s})} && \text{basic derivation}\\
&= -\langle -\mathscr{H}(\mathbf{s})\rangle_\beta && \text{By Equation 2.3}\\
&= \langle \mathscr{H}(\mathbf{s})\rangle_\beta
\end{aligned}$$

Thus, taking the limits they are the same, and we can say that:

$$\lim_{\beta\to\infty}\Big(\langle\mathscr{H}(\mathbf{s})\rangle_\beta\Big) = \lim_{\beta\to\infty}\Big(\frac{1}{Z_n(\beta)}\sum_{\{\mathbf{s}\}}\mathscr{H}(\mathbf{s})e^{-\beta\mathscr{H}(\mathbf{s})}\Big)$$

A tedious derivation shows that the Boltzmann probability distribution concentrates around the minimum for $\beta\to\infty$:

$$\begin{aligned}
\lim_{\beta\to\infty}\frac{1}{Z_n(\beta)}e^{-\beta\mathscr{H}(\mathbf{s})} &= \lim_{\beta\to\infty}\frac{e^{-\beta\mathscr{H}(\mathbf{s})}}{\sum_{\{\mathbf{q}\}} e^{-\beta\mathscr{H}(\mathbf{q})}}\\
&= \lim_{\beta\to\infty}\frac{e^{-\beta\mathscr{H}(\mathbf{s})}}{\sum_{\{\mathbf{q},\mathbf{q}\neq\mathbf{s}^*\}} e^{-\beta\mathscr{H}(\mathbf{q})} + e^{-\beta\mathscr{H}(\mathbf{s}^*)}} && \text{where } \mathbf{s}^* = \min_{\{\mathbf{s}\}}\mathscr{H}(\mathbf{s})\\
&= \lim_{\beta\to\infty}\frac{\frac{e^{-\beta\mathscr{H}(\mathbf{s})}}{e^{-\beta\mathscr{H}(\mathbf{s}^*)}}}{\frac{e^{-\beta\mathscr{H}(\mathbf{s}^*)}}{e^{-\beta\mathscr{H}(\mathbf{s}^*)}} + \frac{\sum_{\{\mathbf{q},\mathbf{q}\neq\mathbf{s}^*\}} e^{-\beta\mathscr{H}(\mathbf{q})}}{e^{-\beta\mathscr{H}(\mathbf{s}^*)}}}\\
&= \lim_{\beta\to\infty}\frac{e^{-\beta(\mathscr{H}(\mathbf{s})-\mathscr{H}(\mathbf{s}^*))}}{1+\sum_{\{\mathbf{q},\mathbf{q}\neq\mathbf{s}^*\}} e^{-\beta(\mathscr{H}(\mathbf{q})-\mathscr{H}(\mathbf{s}^*))}}\\
&= \begin{cases} 1 & if\ \mathbf{s}=\mathbf{s}^*\\ 0 & otherwise\end{cases}
\end{aligned}$$

Where all of the above steps are just algebra tricks.
Thus, the expected value with respect to the Boltzmann distribution of the Hamiltonian concentrates at its minimum value in the $\beta\to\infty$ limit.

$$\lim_{\beta\to\infty}\Big(\langle\mathscr{H}(\mathbf{s})\rangle_\beta\Big) = \min_{\{\mathbf{s}\}}\Big\{\mathscr{H}(\mathbf{s})\Big\} \tag{2.5}$$

□

For this reason, taking low temperature limits *concentrates* the distribution of energy configurations at the minimum possible. Clearly this is of pivotal importance for an optimization problem, but the concentration properties have not been completely uncovered yet. In later sections, more will be added.

A **Sampling problem** appears whenever a generative model $\mathbb{P}(X)$ is to be inferred. The easiest setting is simple bayes estimation. Assume that an unknown variable X generates a variable Y. Having access to Y, it is possible to estimate the posterior of X as:

$$\mathbb{P}(X \mid Y) = \frac{\mathbb{P}(Y\mid X)\mathbb{P}(X)}{Z} = \frac{e^{log(\mathbb{P}(Y|X)\mathbb{P}(X))}}{Z} \tag{2.6}$$

Where it is possible to set $\mathscr{H}(Y = y) = log(\mathbb{P}(Y \mid X)\mathbb{P}(X))$ and apply the statistical mechanics framework with a Boltzmann measure. Having a posterior, estimating a good candidate $\widehat{X}$ for X is feasible. A practical example is proposed below.

Example 2.3 (Gaussian Bayesian Posterior)**.** Let $X \sim \mathcal{N}(0,1)$, $Y_i = X + Z_i$ where $Z_i \sim \mathcal{N}(0,1)$. A well known result is that for a sequence of observations $\{y\}_{i=1}^n$:

$$\mathbb{P}(x \mid \{y\}) = \frac{\mathbb{P}(\{y\} \mid x)\mathbb{P}(x)}{Z} \propto \mathbb{P}(\{y\} \mid x)\mathbb{P}(x) \qquad \text{as } Z \perp\!\!\!\perp x \tag{2.7}$$

$$\propto \prod_{i=1}^{n} e^{\frac{-(y_i-x)^2}{2}} e^{\frac{-x^2}{2}} \qquad \text{ignoring constants and by } y_i \mid x \; i.i.d \tag{2.8}$$

$$\propto e^{-\frac{1}{2}(x^2+\sum_i (y_i-x)^2)} \qquad \text{reordering} \tag{2.9}$$

$$\propto e^{-\frac{1}{2}(x^2+\sum_i y^2 - 2x\sum_i y_i + nx^2)} \tag{2.10}$$

$$\propto e^{-\frac{1}{2}[(n+1)x^2 - 2x\sum_i y_i]} \qquad \text{ignoring constants} \tag{2.11}$$

Which is the kernel of a normal distribution with mean and variance:

$$\mathcal{N}(\mu, \sigma^2) \quad : \quad \mu = \frac{\sum_i y_i}{n+1} \quad \sigma^2 = \frac{1}{n+1} \tag{2.12}$$

2.2 Curie-Weiss Model

In this section a rather easy ferromagnetic model will be analyzed. Though very restricted, it is enough to give an overview of thermodynamic macroscopic functions, with *exact* computations.

We consider a set of binary spins that can be aligned or antialigned, where each *microscopic* particle interacts with all the others. Since the target is magnetization (alignment), the cost of antialignment is included in the energy formula (the Hamiltonian). For simplicity[1], it is preferrable to have an external magnetic field $h \in \mathbb{R}$. As before, the Boltzmann distribution is embedded in the space. The framework naturally suggests the following mathematical objects:

$$\mathbf{S} = \{S_1, \dots, S_N\} \in \{-1,1\}^N \tag{2.13}$$

$$\mathscr{H}_N(\mathbf{s}) = -\frac{1}{2N}\sum_{ij} S_i S_j - h\sum_i S_i \tag{2.14}$$

$$\mathbb{P}_{N,\beta,h}(\mathbf{S} = \mathbf{s}) = \frac{e^{-\beta\mathscr{H}_N(\mathbf{s})}}{Z_N(\beta,h)} \tag{2.15}$$

$$Z_N(\beta,h) = \sum_{\mathbf{S}\in\{-1,1\}^N} e^{-\beta\mathscr{H}_N(\mathbf{s})} = \sum_{\mathbf{S}\in\{-1,1\}^N} e^{\frac{\beta}{2N}\sum_{ij} S_i S_j + \beta h \sum_i S_i} \tag{2.16}$$

Where often the sum over $\mathbf{S} \in \{-1,1\}^N$ will be shortcut to $\{\mathbf{S}\}$.

Since the aim is alignment, a self explained indicator follows.

Definition 2.4 (Magnetization per spin $\overline{\mathbf{S}}$)**.** Consider $\mathbf{S}$, then:

$$\overline{\mathbf{S}} = \frac{1}{N}\sum_i S_i \tag{2.17}$$

[1] Actually, to avoid having the zero solution

Which can substitute the spins in the hamiltonian as:

$$\begin{aligned}
\mathscr{H}_N(\mathbf{s}) &= -\frac{1}{2N}\sum_{ij} S_i S_j - h\sum_i S_i \\
&= -\frac{1}{2}\sum_i S_i \sum_j \frac{S_j}{N} - h\sum_i S_i \\
&= -\frac{1}{2}\sum_i S_i \overline{\mathbf{S}} - hN\frac{\sum_i S_i}{N} \\
&= -\frac{1}{2}\overline{\mathbf{S}}N\frac{\sum_i S_i}{N} - hN\overline{S} \\
&= -N\left(\frac{1}{2}\overline{\mathbf{S}}^2 + h\overline{\mathbf{S}}\right) \\
&= \mathscr{H}_N(\overline{\mathbf{S}})
\end{aligned}$$

Making the energy function determined by the magnetization per spin only.

By the fact that it is a random variable directly linked to the Hamiltonian, the Probability in Equation 2.15 is reexpressed as the probability that the average spin is a value $m \in \mathcal{S}_N = \{-1, 1, skip = \frac{2}{N}\}$, where[2] $\mathcal{S}_N \subset \{\mathbf{S}\}$:

$$\mathbb{P}(\overline{\mathbf{S}} = m) = \frac{\#configs : m}{Z_N(\beta)} e^{\mathscr{H}(\overline{\mathbf{S}})} = \frac{\Omega(m, N)}{Z_N(\beta)} e^{\beta N(\frac{1}{2}m^2 + hm)} \tag{2.18}$$

Where $\Omega(m, N)$ counts the number of configurations granting magnetization m for an N-particle system. Though apparently difficult to evaluate, some considerations that will prove to be useful in the thermodynamic limit are made below:

Definition 2.5 (Binary Entropy $H(\cdot)$)**.** The binary entropy is a function of the magnetization:

$$H(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right) \tag{2.19}$$

Contrary to the information theoretic notation, the logarithm is natural and not in base 2. Nevertheless, they are equal up to a multiplicative constant.

Theorem 2.6 (Ω Closed form and Properties)**.** *For all m, N valid it holds that:*

1.

$$\Omega(m, N) = \frac{N!}{(N\frac{1-m}{2})!(N\frac{1+m}{2})!} \tag{2.20}$$

2.

$$\frac{e^{NH(m)}}{N+1} \leq \Omega(m, N) \leq e^{NH(m)} \tag{2.21}$$

Proof. **(Claim** 1**)** Observe that:

$$S_i \in \{-1, 1\} \forall i \implies m = \frac{1}{N}\sum_i S_i \in [-1, 1] \iff Nm = \sum_i S_i \in [-N, N]$$

Thanks to this formulation, it is rather easy to claim by induction on N that:

$$N + Nm = N + \sum_i S_i \; even \; \forall m, N$$

More importantly, starting from either of the two extreme configurations, it can be argued that the attainable values of m are at a $\frac{1}{N}$ distance, since to change m from the all ups

[2] These are not all the possible values the average magnetization can take!

$+1$ or all downs -1 (and from all of the intermediate ones), we need to change an **odd** number of spins summing to either -1 or 1, which impacts the overall magnetization by a factor of $\frac{1}{N}$. Thus:

$$m \in \left\{ -1, 1, \ skip = \frac{1}{N} \right\} \iff Nm \in \left\{ -N, N, \ skip = 1 \right\}$$

Using a parallel argument, instead of considering a symmetric view of the problem, observe that for $2N$ particles where N are fixed to be $S_i = +1$ in an arbitrary position:

$$N + Nm \in \left\{ 0, 2N, \ skip = 1 \right\} \quad \forall N, m$$

For a given value of magnetization, if $m > 0$ then it will become the problem of choosing how many zeros and how many up spins to set, while for a negative magnetization, how many zeros and how many -1s we set. This is an equivalent formulation, as the zeros in this case mean an **even** number of ± 1 pairs in the original problem. We recognize that such a view suggests that we are evaluating the number of ways of setting N elements into $N + Nm$ boxes of $1s$ where the first are the fixed ones to make the problem positive, while the second ones are the *all up spin* particles required to attain $Nm = \sum_i S_i$. To redirect the problem to the original formulation, it is necessary to divide the number of boxes by 2, as to make the index go from 0 to N with skip $\frac{1}{2}$. These two facts support our proof that for a given $N, m = \frac{1}{N} \sum_i S_i$ we have:

$$\Omega(m, N) = \binom{N}{\frac{N+Nm}{2}} = \frac{N!}{\left(N - \frac{N+Nm}{2}\right)! \left(\frac{N+Nm}{2}\right)!} = \frac{N!}{(N\frac{1-m}{2})!(N\frac{1+m}{2})!}$$

(Claim 2) *Bounds on the binomial*, Exercise 1.1 Chapter 1 [KZ21b]. ☐

Lemma 2.7 (Ω bounds implication on Boltzmann measure)**.** *Consider* $\phi(m, \beta, h) = H(m) + \frac{1}{2}\beta m^2 + \beta h m$*, shortwritten as* $\phi(m)$*, then by Equation 2.21:*

$$\frac{1}{N+1} \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \leq \mathbb{P}(\overline{\boldsymbol{S}} = m) \leq \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \tag{2.22}$$

Where we have just substituted the bounds on $\Omega(m, N)$

Lemma 2.8 (Ω bounds implication on Partition Sum)**.** *Consider the same* $\phi(m)$*, then:*

$$1 \leq \sum_m \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \leq (N+1) \frac{e^{N\phi(m^*)}}{Z_N(\beta, h)} \tag{2.23}$$

where $m^* = \underset{m \in [-1,1]}{argmax} \{\phi(m, \beta, h)\}$ *is the maximum possible value of* ϕ

Proof. We sum over m the probability and the RHS of Equation 2.22 and get that by the previous Lemma:

$$\sum_m \mathbb{P}(\overline{\mathbf{S}} = m) = 1 \leq \sum_m \frac{e^{N\phi(m)}}{Z_N(\beta, h)} = \frac{1}{Z_N(\beta, h)} \sum_m e^{N\phi(m)} \tag{2.24}$$

Where the possible values of m are $\{1, 1 - \frac{2}{N}, 1 - \frac{4}{N}, \dots, -1\}$. There are N elements in $\mathcal{S}_N$ indicating that:

$$1 \leq \frac{1}{Z_N(\beta, h)} \sum_m e^{N\phi(m)} \leq (N+1) \frac{e^{N\phi(m^*)}}{Z_N(\beta, h)} \tag{2.25}$$

☐

The result of Lemma 2.8 highlights important aspects from a thermodynamic perspective.

Definition 2.9 (Gibbs Free Entropy and Free Entropy density $\Phi_N(\beta, h)$)**.** For a system with N particles we define as free entropy $\log Z_N(\beta, h)$ with density:

$$\Phi_N(\beta, h) = \frac{\log Z_N(\beta, h)}{N} \tag{2.26}$$

Which in the limit as $N \to \infty$ loses the N pedix.

This quantity is important and presents nice properties in the thermodynamic limit. Taking the logarithm on the last inequality of Lemma 2.8 one gets:

$$\log 1 = 0 \leq \log \left[(N+1) \frac{e^{N\phi(m^*)}}{Z_N(\beta, h)} \right] \tag{2.27}$$

$$\implies \log(N+1) + N\phi(m^*) - \log(Z_N(\beta, h)) \geq 0 \tag{2.28}$$

$$\implies \frac{\log(Z_N(\beta, h))}{N} = \Phi_N(\beta, h) \leq \phi(m^*) + \frac{\log(N+1)}{N} \tag{2.29}$$

Which is an upper bound on the free entropy density. To get a lower bound, the LHS and central element of Lemma 2.7 are considered, and applying the logarithm:

$$\frac{1}{N+1} \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \leq \mathbb{P}(\overline{\mathbf{S}} = m) \leq 1 \tag{2.30}$$

$$\implies \log \left[\frac{1}{N+1} \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \right] \leq \log(1) = 0 \tag{2.31}$$

$$\implies -\log(N+1) + N\phi(m) - \log(Z_N(\beta, h)) \leq 0 \tag{2.32}$$

$$\implies \frac{\log Z_N(\beta, h)}{N} = \Phi_N(\beta, h) \geq -\frac{\log(N+1)}{N} + \phi(m) \tag{2.33}$$

This result is true subject to $m \in \mathcal{S}_N$. It is then true for $m^{max} = \underset{m \in \mathcal{S}_N}{argmax}\{\phi(m)\}$. A complete maximization over the set of possible values $[-1, 1]$. The next Lemma helps understanding why this can be ignored, especially in Thermodynamics. Moreover, it allows to formulate the lower bound in terms of $\phi(m^*)$

Lemma 2.10 ($\mathcal{S}_N$ is enough in the limit)**.** *As $N \to \infty$ it holds that:*

$$\phi(m^{max}) \to \phi(m^*) \tag{2.34}$$

$$\phi(m^*) - \frac{\log(N(N+1))}{N} \leq \Phi_N(\beta, h) \tag{2.35}$$

Proof. Apply the mean value theorem for m^*, m^{max}. Then:

$$\exists c \in [m^*, m^{max}] \mid \phi(m^*) = \phi(m^{max}) + \phi'(c)(m^* - m^{max}) \tag{2.36}$$

Where $|m^* - m^{max}| \leq \frac{2}{N}$ as the skips are of width $\frac{2}{N}$.
Informal: therefore, we can argue that:

$$\phi(m^{max}) > \phi(m^*) - \frac{\log(N)}{N} \tag{2.37}$$

Which gets closer and closer to $\phi(m^*)$ as $N \to \infty$. Morevoer, substituting into the result of Lemma 2.8 one gets:

$$-\frac{\log(N+1)}{N} + \phi(m^*) - \frac{\log(N)}{N} < \phi(m^{max}) - \frac{\log(N+1)}{N} \leq \Phi_N(\beta, h)$$

$$\implies \phi(m^*) - \frac{\log(N(N+1))}{N} < \Phi_N(\beta, h)$$

∎

All of the above results suggest the following conclusion, outlined as a Theorem.

Theorem 2.11 (Free Entropy Density thermodynamic limit)**.** *Let $N \to \infty$, then:*

$$\lim_{N\to\infty} \Phi_N(\beta, h) = \Phi(\beta, h) = \phi(m^*) \tag{2.38}$$

Proof. From Lemmas 2.7, 2.8 and 2.10 it is possible to conclude that the free entropy density is bounded in the interval[3]:

$$\phi(m^*) - \frac{\log(N(N+1))}{N} < \Phi_N(\beta, h) \leq \phi(m^*) + \frac{\log(N+1)}{N} \tag{2.39}$$

Taking the limits on all inequalities:

$$\lim_{N\to\infty}\left[\phi(m^*) - \frac{\log(N(N+1))}{N}\right] < \lim_{N\to\infty} \Phi_N(\beta, h) \quad \leq \lim_{N\to\infty}\left[\phi(m^*) + \frac{\log(N+1)}{N}\right] \tag{2.40}$$

$$\phi(m^*)^- < \Phi(\beta, h) \qquad \leq \phi(m^*)^+ \tag{2.41}$$

Then, it is the case that:

$$\Phi(\beta, h) = \phi(m^*) = \underset{m\in[-1,1]}{max}\{\phi(m)\} \tag{2.42}$$

□

Mathematicians call this a large deviation behavior. For the purpose of this book, the starting result will be outlined as a theorem and exploited throughout. The idea is that in the thermodynamic limit the probability is determined by a large deviation rate dependent on the argument that makes it exponentially small as it *goes away* from it.

Theorem 2.12 (Large Deviation Behavior of the Boltzmann Measure)**.** *Given a system of N particles:*

$$\lim_{N\to\infty} \frac{\log(\mathbb{P}(\overline{\boldsymbol{S}} = m))}{N} = \phi(m) - \phi(m^*) \tag{2.43}$$

Where $\phi(m) - \phi(m^*)$ is the rate of the large deviation behavior.

Proof. It is sufficient to notice that:

$$as\ N \to \infty \quad \mathbb{P}(\overline{\mathbf{S}} = m) \propto \frac{e^{N\phi(m)}}{Z_N(\beta, h)} = \frac{e^{N\phi(m)}}{e^{N\Phi_N(\beta,h)}} \underset{N\to\infty}{\asymp} e^{N(\phi(m)-\phi(m^*))} \tag{2.44}$$

$$\implies \lim_{N\to\infty} \frac{log(\mathbb{P}(\overline{S} = m))}{N} = \lim_{N\to\infty} \frac{N(\phi(m) - \phi(m^*))}{N} \tag{2.45}$$

$$= \phi(m) - \phi(m^*) \tag{2.46}$$

Where the $\underset{N\to\infty}{\asymp}$ reads "asymptotically equal in the limit $N \to \infty$". □

This result is of pivotal importance. From a combinatorially exploding sum the problem is reduced to a one dimensional function to optimize ϕ. We see that the probability decays depending on the realized value of ϕ. This is a concentration phenomena, which guarantees that physical quantities are deterministic! For clarity of notation, we define the object of interest below.

[3]Notice that the lower bound is strict!

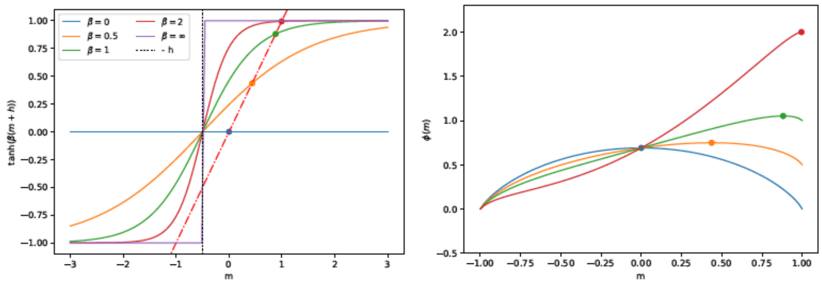


Figure 2.1: Left: Equation 2.50, $h \neq 0$ for different β values. Right: relation in terms of $\phi(m)$, with solutions corresponding to its global maximum.

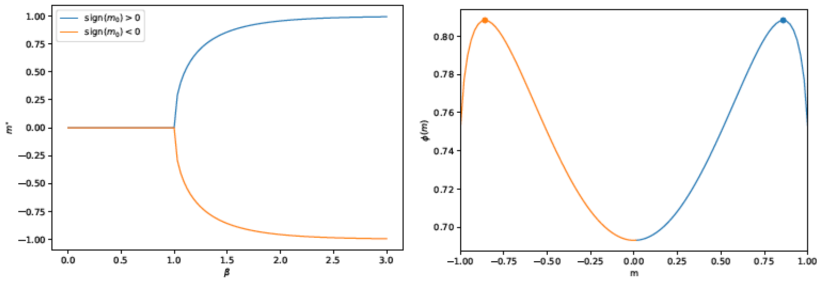


Figure 2.2: For $h = 0$, left plot is m^* as a function of β, right plot shows $\phi(m)$ as a function of β.

Definition 2.13 (Large Deviation Rate $I(m)$)**.** For a realized magnetization value define the large deviation rate as:

$$I(m) = \phi(m^*) - \phi(m) \quad where \quad \lim_{N\to\infty} \frac{log(\mathbb{P}(\overline{S} = m))}{N} = -I(m) \tag{2.47}$$

To find the attained value at thermodynamic limit, we seek a magnetization m^* : $\phi'(m^*) = 0$. The equation is of the form $\phi(m) = H(m) + \frac{1}{2}\beta m^2 + \beta hm$. Then:

$$\phi'(m) = H'(m) + \beta m + \beta h = 0 \tag{2.48}$$

After some elementary calculations on the derivative of $H(m)$ it becomes:

$$\frac{1}{2}\log\left(\frac{1+m}{1-m}\right) = \beta(m+h) \tag{2.49}$$

$$\implies m = \tanh\beta(h+m) \qquad \text{by } \tanh^{-1}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right) \text{ where } x = m \tag{2.50}$$

We refer to this as a **mean field equation** or **saddle point equation** for the Curie Weiss model.

Depending on the values β, h, there can be up to three solutions. In particular, for $\beta > 1 \wedge h = 0$, there will be two coexistent maxima. This implies that with probability $\frac{1}{2}$ either of the two will be take place in the thermodynamic limit. This phenomena in Physics is called *phase coexistance*. Concerning this, more details will come later. If instead $\beta > 1$ but $h \neq 0$, either of the two will be chosen, depending on the sign of h itself. Some plots from the original lecture notes [KZ21b] can be found below.

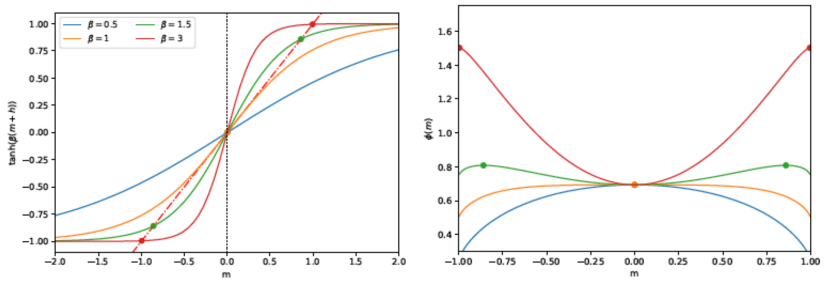


Figure 2.3: Same setting of Figure 2.1 but with $h = 0$. For $\beta < 1$ we have one degenerate maximum of $\phi(m)$ at $m^* = 0$, while for $\beta > 1$ there are two symmetric *coexistent* maximums.

2.3 On the enoughness of Φ

As it will turn out, by just having access to the free entropy density, many conclusions can be carried out about a model. For historical notational reasons, different reasearchers tend to use different (equivalent) objects. Here, free energy density Φ will be used.

For N finite, we can evaluate:

$$\frac{1}{\beta}\frac{\partial \Phi_N(\beta,h)}{\partial h} = \frac{1}{\beta}\frac{\partial}{\partial h}\frac{\log Z_N(\beta,h)}{N} \tag{2.51}$$

$$= \frac{1}{\beta N}\frac{1}{Z_N(\beta,h)}\frac{\partial Z_N(\beta,h)}{\partial h} \tag{2.52}$$

$$= \frac{1}{\beta N}\frac{1}{Z_N(\beta,h)}\frac{\partial}{\partial h}\left(\sum_{\{\mathbf{S}\}} e^{-\beta\mathscr{H}(\mathbf{S})}\right) \quad \text{where } \mathscr{H}(\mathbf{S}) = -N\left(\frac{1}{2}\overline{\mathbf{S}}^2 + h\overline{\mathbf{S}}\right) \tag{2.53}$$

$$= \frac{1}{\beta N}\frac{1}{Z_N(\beta,h)}\sum_{\{\mathbf{S}\}} \beta N\overline{\mathbf{S}}e^{-\beta\mathscr{H}(\mathbf{S})} \quad \text{exchanging derivative and sum} \tag{2.54}$$

$$= \frac{\sum_{\{\mathbf{S}\}}\overline{\mathbf{S}}e^{-\beta\mathscr{H}(\mathbf{S})}}{Z_N(\beta,h)} \quad \text{which is a Boltzmann weighted expected value of } \overline{\mathbf{S}} \tag{2.55}$$

$$= \langle\overline{\mathbf{S}}\rangle_{N,\beta,h} \tag{2.56}$$

Observation 2.14 (About Equation 2.56)**.** *In a large deviation setting, the most probable is also the most likely (as it must concentrate on it). Unfortunately, the passage of Equation 2.54 is not directly doable in the limit. Below, the required tools will be outlined.*

Laplace method, Exercise 1.2, Chapter 1 [KZ21b].

The Laplace method on Z We recognize that in presence of a magnetic field $h \in \mathbb{R}$ one can write:

$$\mathscr{H}_N(\mathbf{S}) = -\frac{N}{2}\overline{\mathbf{S}} - h\overline{\mathbf{S}}N = \mathscr{H}_N^0(\mathbf{S}) - h\overline{\mathbf{S}}N \tag{2.57}$$

Where $\mathscr{H}_N^0(\overline{\mathbf{S}})$ is the Hamiltonian in absence of a magnetic field. Going further, one could approximate the sum over magnetizations by an integral, which can be determined with the Laplace method if there is a single maximum m^*, and we are away from phase transitions. Indeed, denoting as $I_0^*(m,\beta) = \phi(m,\beta,0) - \phi(m^*,\beta,0)$ the true deviation rate when $h=0$. If we assume it exists, then:

$$Z_N(\beta,h) = \sum_m \left(\sum_{\{\mathbf{S}\}} e^{-\beta\mathscr{H}_N^0(\mathbf{S})}\mathbf{1}(\overline{\mathbf{S}}=m)\right)e^{N\beta hm} \asymp \int_{-1}^{1} dm\ e^{N(\phi(\beta,0)-I_0^*(m,\beta)+\beta hm)} \tag{2.58}$$

$$= \sum_m \Omega(m,N)e^{-N(H(m)+\beta\mathscr{H}_N^0(m)+\beta hm)} \asymp \int_{-1}^{1} dm\ e^{-N(H(m)+\beta\mathscr{H}_N^0(m)+\beta hm)} \tag{2.59}$$

Where the Laplace integral at the limit gives a concentration on the maximum of the argument function:

$$Z_N(\beta,h) \underset{N\to\infty}{\overset{Laplace}{\approx}} e^{N\Phi(\beta,h)} \qquad \text{lookup Equation 2.44 for the idea} \tag{2.60}$$

$$\underset{N\to\infty}{\overset{Laplace}{\approx}} e^{\max_m\{N(\phi(\beta,0)-I_0^*(m,\beta)+\beta hm)\}} \tag{2.61}$$

$$\implies \Phi(\beta,h) = \max_m\{\beta hm - I_0^*(m) + \phi(\beta,0)\} \tag{2.62}$$

$$= \max_m\{\beta hm - I_0^*(m)\} + \Phi(\beta,0) \qquad \text{Eqn. 2.42} \tag{2.63}$$

$$\implies \Phi(\beta,h) - \Phi(\beta,0) = \max_m\{\beta hm - I_0^*(m)\} \tag{2.64}$$

Or similarly, using Equation 2.59:

$$\Phi(\beta,h) = \max_m\left\{H(m) - \beta\mathscr{H}_N^0(m) + \beta hm\right\} \tag{2.65}$$

$$= H(m^*) - \beta\mathscr{H}_N^0(m^*) + \beta hm^* \tag{2.66}$$

$$= -I^*(m) + \beta hm^* = \phi(m^*) \tag{2.67}$$

Deriving this with respect to h as before in the finite size case:

$$\left.\frac{\partial\Phi}{\partial h}\right|_{h=0^+} = \beta m^* - \left.\frac{\partial I^*}{\partial m}\frac{\partial m}{\partial h}\right|_{h=0^+} \tag{2.68}$$

$$= \beta m^* \qquad \text{As } \left.\frac{\partial I^*}{\partial m}\right|_{h=0^+} = 0 \text{ when maximum w.r.t. } m \tag{2.69}$$

By the large deviation principle, we can obtain the most likely m^* by just computing the derivative of the free entropy density. In this case, we checked $m^*(h=0^+)$, we could do for a different $h \neq 0$. For $h=0$ solutions are multiple and the Laplace method is more convoluted.

Legendre Transform Written in either of the two ways, we are doing a Legendre transform of the variable as:

$$\Phi(\beta,h) - \Phi(\beta,0) = \max_m\{\beta hm - I_0^*(m)\} \tag{2.70}$$

$$\Phi(\beta,h) = \max_m\{\beta hm - I^*(m)\} \tag{2.71}$$

Further in this direction, by applying the inverse Legendre transform of $\Phi(\beta,h)$ we can define:

$$I_0(m) = \underset{h}{max}\{\beta hm - \Phi(\beta,h)\} + \Phi(\beta,0) \tag{2.72}$$

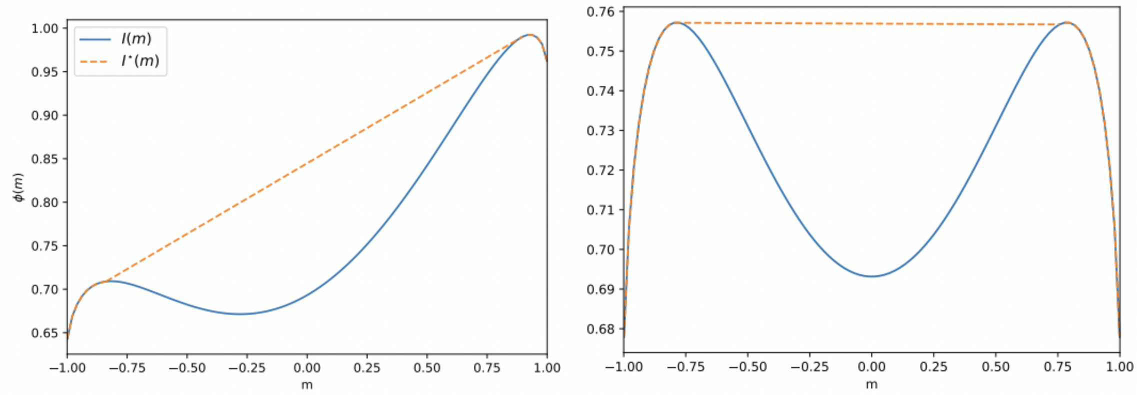


Figure 2.4: Two convex envelopes (orange) of large deviation rates (blue)

Where $I_0(m)$ is the convex envelope of $I_0^*(m)$. Thus, if we have a way to compute the free energy $\Phi(\beta, h)$, we can recover the large deviation rate with a Legendre transform. In the event that the deviation rate is not convex, we would only recover *exposed* points, and have an upper bound for unexposed points. An example is shown in Figure 2.4.

To summarize, we found that:

- as $N \to \infty$ we can recover the large deviation rate of concentration for an observable such as m, the average magnetization, except in very special cases [4] with probability 1.
- m^* is found via the derivative of the free entropy density $\Phi_N = \dfrac{\log Z_N}{N}$
- the Legendre transform of the free entropy density gives a convex envelope of the large deviation rate.

2.4 Computation Toolbox

A question worth asking is how to compute the free energy. In this section, we will analyze some of the most important techniques in Machine Learning: the cavity and replica method. To familiarize with the process of putting them in practice, they will be applied to the Curie Weiss Model.

Without long theoretical derivations, they show up as heuristics to compute Φ_N when $N \to \infty$.

2.4.1 Intro to the Cavity Method

The key observation is that we attempt to compare two systems: one with N particles, and another with $N + 1$, aiming to see what changes.

[4] e.g. $\beta > 1, h = 0$

Writing the Hamiltonian of the increased size system, we denote the added spin as S_0. It is possible to write such an object as a function of $\mathscr{H}_N$:

$$-\beta\mathscr{H}_{N+1}(\mathbf{S}) = \beta\frac{1}{2}(N+1)\left(\frac{\sum_{i=1}^{N} S_i + S_0}{N+1}\right)^2 + \beta h(S_0 + \sum_{i=1}^{N} S_i) \tag{2.73}$$

$$= \beta\frac{1}{2}(N+1)\frac{S_0^2}{(N+1)^2} + \beta\frac{1}{2}(N+1)\left(\frac{\sum_i S_i}{(N+1)}\right)^2\frac{N^2}{N^2} \tag{2.74}$$

$$+ \beta\frac{1}{2}(N+1)2\left(\frac{S_0\sum_i S_i}{(N+1)^2}\right)\frac{N}{N} + \beta h S_0 + \beta h\sum_i S_i \qquad \text{expanding the square} \tag{2.75}$$

$$= \beta\frac{1}{2(N+1)} + \beta\frac{1}{2}\left(\frac{\sum_i S_i}{N}\right)^2\frac{N^2}{(N+1)} \tag{2.76}$$

$$+ \beta\frac{N}{(N+1)}\frac{\sum_i S_i}{N}S_0 + \beta h S_0 + \beta h\sum_i S_i \qquad \text{by } S_0^2 = 1 \text{ and reordering} \tag{2.77}$$

With a change of variables $\beta' = \beta\dfrac{N}{N+1}$ and $h' = h\dfrac{N}{N+1}$:

$$-\beta'\mathscr{H}_{N+1}(\mathbf{S}, h') = \beta\frac{1}{2(N+1)} + \beta\frac{1}{2}N\left(\frac{\sum_i S_i}{N}\right)^2 \tag{2.78}$$

$$+ \beta S_0\left(\frac{\sum_i S_i}{N}\right) + \beta h\sum_i S_i + \beta h S_0 \qquad \text{where the first term is constant} \tag{2.79}$$

$$= o(1) - \beta\mathscr{H}_N(\mathbf{S}) + \beta S_0\left(\frac{\sum_i S_i}{N}\right) + \beta h S_0 \tag{2.80}$$

And the new system is equal to the old one plus a small perturbation. Thanks to this relation, it is possible to write the expectation over $N+1$ particles as:

$$\langle S_0\rangle_{N+1,\beta',h'} = \frac{\sum_{\mathbf{S},S_0} S_0 e^{-\beta'\mathscr{H}_{N+1}}}{\sum_{\mathbf{S},S_0} e^{-\beta'\mathscr{H}_{N+1}}} \qquad \text{where } S_0 = \pm 1 \tag{2.81}$$

$$= \frac{\sum_{\mathbf{S}}\sum_{S_0} S_0 e^{-\beta\mathscr{H}_N} e^{\beta S_0\sum\frac{S_i}{N} + \beta h S_0}}{\sum_{\mathbf{S}}\sum_{S_0} e^{-\beta\mathscr{H}_N} e^{\beta S_0\sum\frac{S_i}{N} + \beta h S_0}} \qquad \text{up to } o(1) \text{ approx} \tag{2.82}$$

$$= \frac{\sum_{\mathbf{S}}\sum_{S_0}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} S_0 e^{\beta S_0\overline{\mathbf{S}} + \beta h S_0}}{\sum_{\mathbf{S}}\sum_{S_0}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} e^{\beta S_0\overline{\mathbf{S}} + \beta h S_0}} \qquad \text{adjusting} \tag{2.83}$$

$$= \frac{\sum_{\mathbf{S}}\sum_{S_0}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} S_0 e^{\beta S_0(\overline{\mathbf{S}}h)}}{\sum_{\mathbf{S}}\sum_{S_0}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} e^{\beta S_0(\overline{\mathbf{S}}+h)}} \qquad \text{reordering} \tag{2.84}$$

$$= \frac{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} e^{\beta(\overline{\mathbf{S}}+h)} - e^{-\beta(\overline{\mathbf{S}}+h)}}{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N} e^{\beta(\overline{\mathbf{S}}+h)} + e^{-\beta(\overline{\mathbf{S}}+h)}} \qquad \text{expanding } \sum_{S_0} \tag{2.85}$$

$$= \frac{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N}\frac{1}{2}(e^{\beta(\overline{\mathbf{S}}+h)} - e^{-\beta(\overline{\mathbf{S}}h)})}{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N}\frac{1}{2}(e^{\beta(\overline{\mathbf{S}}+h)} + e^{-\beta(\overline{\mathbf{S}}+h)})} \qquad \text{add } \frac{1}{2} \tag{2.86}$$

$$= \frac{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N}\sinh\beta(\overline{\mathbf{S}}+h)}{\sum_{\mathbf{S}}\frac{e^{-\beta\mathscr{H}_N}}{Z_N}\cosh\beta(\overline{\mathbf{S}}+h)} \qquad \text{hyperbolic identities} \tag{2.87}$$

$$= \frac{\langle\sinh\beta(\overline{\mathbf{S}}+h)\rangle_{N,\beta,h}}{\langle\cosh\beta(\overline{\mathbf{S}}+h)\rangle_{N,\beta,h}} \tag{2.88}$$

Further assuming that, out of the double maxima (phase coexistence) case, $\overline{\mathbf{S}}$ is concentrated around m^* by large deviation we would have that:

$$\overline{\mathbf{S}} = m^* = \langle \mathbf{S} \rangle \tag{2.89}$$

And we can state that:

$$\langle S_0 \rangle_{N+1,\beta',h'} = \frac{\langle \sinh \beta(\overline{\mathbf{S}} + h) \rangle_{N,\beta,h}}{\langle \cosh \beta(\overline{\mathbf{S}} + h) \rangle_{N,\beta,h}} \approx \frac{\sinh \beta(m^* + h)}{\cosh \beta(m^* + h)} = \tanh \beta(m^* + h) \tag{2.90}$$

Moreover, as $N \to \infty$ we have that $\beta' \to \beta$ and then:

$$m^* = \tanh \beta(m^* + h) \tag{2.91}$$

For a proof of such claim, we reroute the reader to the original lecture notes [KZ21b].

2.4.2 Intro to the Replica Method

To conclude the lecture we will overview a widely used trick in the specific context of the Curie Weiss Model. In particular, we will exploit the Dirac-Fourier method, which is the starting point of replica computations. In order to do this, we first introduce some basic notions.

Definition 2.15 (Delta Dirac distribution $\delta(x)$)**.** The dirac δ function is such that:

$$\delta(x) = \begin{cases} +\infty & x = 0 \\ 0 & x \neq 0 \end{cases} \quad s.t. \quad \int_{\mathbb{R}} dx \delta(x) = 1 \tag{2.92}$$

Constrained to being a distribution.

Theorem 2.16 (Dirac Delta Property)**.**

$$\int dm f(m) \delta(m - x) = f(x) \tag{2.93}$$

$$\equiv \int dm f(m) \delta(Nm - x) = \frac{1}{N} f(x) \tag{2.94}$$

Proof. The Dirac delta is concentrated at 0 in its Definition. For the case $\delta(m - x)$ we instead have that it is concentrated at:

$$m - x - 0 \implies m = x \tag{2.95}$$

Integrating over $\mathbb{R}$ the expected value is that of a degenerate distribution concentrated at x and:

$$\int dm f(m) \delta(m - x) = f(x) \tag{2.96}$$

The equivalence with the second claim is carried out as follows. We recognize that N is fixed and thus $dNm = Ndm$ and that $\delta(Nm - x) = \delta(m - \frac{x}{N})$ as they both concentrate at $x = Nm$. Thus:

$$\int dm f(m) \delta(Nm - x) = \frac{1}{N} \int dNm f(m) \delta(Nm - x) \tag{2.97}$$

$$= \frac{1}{N} \int dNm f(m) \delta(Nm - x) = \frac{1}{N} f(x) \tag{2.98}$$

□

The second result of Theorem 2.16 is the most used version by Physicists. Notice that if we take the logarithm of such identity we find that:

$$\lim_{N\to\infty} \frac{\log \frac{f(x)}{N}}{N} = \lim_{N\to\infty} \frac{\log f(x)}{N} - \frac{\log N}{N} \to \frac{\log f(x)}{N} \tag{2.99}$$

So that the N in front can be ignored.

Lemma 2.17 (A *useful to expand* hyperbolic identity)**.** *For a set of N sized Potts spins $\{\boldsymbol{S}\} = \{-1,1\}^N$ notice that:*

$$\sum_{\{\boldsymbol{S}\}} e^{-\kappa \sum_i S_i} = (e^{\kappa} + e^{-\kappa})^N = (2\cosh\kappa)^N \quad \forall \kappa \in \mathbb{C}, \forall N \in \mathbb{N}^+ \tag{2.100}$$

Which is trivial but difficult to explain in one line steps for proofs.

Proof. We prove the claim by induction on N. For $N = 0$ it is trivial, so starting from 1 **(Base case** $N = 1$**)** Observe that:

$$\{\mathbf{S}\} = \{-1,1\} \implies \sum_{\{\mathbf{S}\}} e^{-\kappa \sum_i S_i} = \sum_{\{-1,1\}} e^{-\kappa \sum_i S_i} = e^{-\kappa} + e^{\kappa} = (e^{-\kappa} + e^{\kappa})^1 = (2\cosh\kappa)$$

(Induction Hypothesis) Assume it is true $\forall N \in \mathbb{N}^+$
(Inductive Case $N + 1$**)** We aim to evaluate:

$$\begin{aligned}
\sum_{\{\mathbf{S}\}} e^{-\kappa \sum_i S_i} &= \sum_{\{-1,1\}^{N+1}} e^{-\kappa \sum_{i=1}^{N+1} S_i} && \text{with } N+1 \text{ pedix in the sums} \\
&= \sum_{\{-1,1\}^N \times \{-1,1\}} e^{-\kappa(\sum_{i=1}^{N} S_i + S_{N+1})} && \text{sums split with } S_{N+1} \in \{-1,1\} \\
&= \sum_{\{-1,1\}^N} \sum_{\{-1,1\}} e^{-\kappa S_{N+1}} e^{-\kappa \sum_{i=1}^{N} S_i} \\
&= \sum_{\{-1,1\}^N} (e^{-\kappa} + e^{\kappa}) e^{-\kappa \sum_{i=1}^{N} S_i} \\
&= (e^{-\kappa} + e^{\kappa}) \sum_{\{-1,1\}^N} e^{-\kappa \sum_{i=1}^{N} S_i} \\
&= (e^{-\kappa} + e^{\kappa})(e^{-\kappa} + e^{\kappa})^N && \text{inductive hypothesis} \\
&= (e^{-\kappa} + e^{\kappa})^{N+1} \\
&= (2\cosh\kappa)^{N+1} && \text{trigonometric identities}
\end{aligned}$$

□

We consider again the Hamiltonian with N particles and the partition sum.

Assume we wish to compute probabilities but do not have access to $\Omega(m)$ for a magnetization m. Expressing the partition sum with the help of Theorem 2.16 we can concentrate the evaluation around the magnetization m using $\delta(Nm - \sum_i S_i)$ where $\frac{\sum_i S_i}{N} = m$:

$$Z_N = \sum_{\{\mathbf{S}\}} e^{\frac{N\beta}{2}(\frac{\sum_i S_i}{N})^2 + N\beta h \sum_i \frac{S_i}{N}} \tag{2.101}$$

$$= N \sum_{\{\mathbf{S}\}} \int dm\, \delta(Nm - \sum_i S_i) e^{\frac{N\beta}{2}m^2 + N\beta hm} \tag{2.102}$$

$$= N \int dm\, e^{\frac{N\beta}{2}m^2 + N\beta hm} \sum_{\{\mathbf{S}\}} \delta(Nm - \sum_i S_i) \tag{2.103}$$

Where in the last passage we exchange sum and integral since we recognize that the delta function *selects* configurations that attain a magnetization m and zeroes out the others, so that the integral is done for each correct configuration that attains a specific m magnetization. Further assuming we cannot compute the exact integral as it is impossible to evaluate the entropy at fixed m, we implement a Fourier Transform of the delta function and write:

$$Z_N = N \int dm \int d\lambda \, e^{\frac{N\beta}{2}m^2 + N\beta hm} \sum_{\{\mathbf{S}\}} e^{i2\pi\lambda N(m - \sum_i \frac{S_i}{N})} \qquad \text{substitute } \widehat{m} = i2\pi\lambda : d\lambda = \frac{d\widehat{m}}{2i\pi} \tag{2.104}$$

$$= \frac{N}{2i\pi} \int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\widehat{m} \, e^{\frac{N\beta}{2}m^2 + N\beta hm + Nm\widehat{m}} \sum_{\{\mathbf{S}\}} e^{-\widehat{m}\sum_i S_i} \quad \text{where integrals are over } \mathbb{R} \text{ and } \mathbb{C} \tag{2.105}$$

$$= \frac{N}{2i\pi} \int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\widehat{m} \, e^{\frac{N\beta}{2}m^2 + N\beta hm + Nm\widehat{m}} (2\cosh\widehat{m})^N \qquad \text{Lem 2.17} \tag{2.106}$$

And taking the density of the log at the limit $N \to \infty$ to obtain $\Phi(\beta, h)$:

$$\lim_{N\to\infty} \frac{\log Z_N}{N} = \lim_{N\to\infty} \frac{\log\left[\frac{N}{2i\pi} \int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\widehat{m} \, e^{\frac{N\beta}{2}m^2 + N\beta hm + Nm\widehat{m}} (2\cosh\widehat{m})^N\right]}{N} \tag{2.107}$$

$$\to \frac{1}{N} \log\left[\int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\widehat{m} \, e^{\frac{N\beta}{2}m^2 + N\beta hm + Nm\widehat{m}} (2\cosh\widehat{m})^N\right] \tag{2.108}$$

Where letting $(2\cosh\widehat{m})^N = e^{N\log 2 + N\log\cosh\widehat{m}}$ we get that:

$$\Phi(\beta, h) = \frac{1}{N} \log\left[\int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\widehat{m} \, e^{\frac{N\beta}{2}m^2 + N\beta hm + Nm\widehat{m} + N\log 2 + N\log\cosh\widehat{m}}\right] \tag{2.109}$$

While this might look difficult to evaluate, there is a simple way. Laplace theorem cannot be directly used as the integral for $\widehat{m}$ is over the complex plane. We will instead use the saddle point method, which generalizes Laplace, exploiting results by Cauchy and was developed by Debye and Riemann [Deb09; RWD13]. Essentially, the concentration will take place at a saddle point of the $(m, \widehat{m})$ space:

$$\Phi(\beta, h) = \underset{m,\widehat{m}}{Ext}\{g(m, \widehat{m})\} \; : \; g(m, \widehat{m}) = \frac{\beta}{2}m^2 + \beta hm + \widehat{m}m + \log 2 + \log\cosh\widehat{m} \tag{2.110}$$

Taking the extremum of g w.r.t. $\widehat{m}$ we get:

$$\frac{\partial g}{\partial \widehat{m}} = m + \frac{1}{\cosh\widehat{m}} \sinh\widehat{m} = m + \tanh\widehat{m} = 0 \iff m = -\tanh\widehat{m} \tag{2.111}$$

Inverting the relation as $\widehat{m} = -\tanh^{-1} m$ and substituting into g:

$$g(m, \widehat{m}) = g(m) = \frac{N\beta}{2}m^2 + \beta hm - m(\tanh^{-1} m) + \log(2\cosh\tanh^{-1} m) \tag{2.112}$$

Where using the identity:

$$\log(2\cosh\tanh^{-1} m) - m(\tanh^{-1} m) = -\left(\frac{1+x}{2}\log(\frac{1+x}{2}) + \frac{1-x}{2}\log(\frac{1-x}{2})\right) = H(x) \tag{2.113}$$

We get that:

$$g(m) = \frac{\beta N}{2} + \beta hm + H(m) = \phi(\beta, h, m) \tag{2.114}$$

And as previously found $\Phi(\beta, h) = \underset{m}{Extr}\{\phi(\beta, h, m)\}$ with a simpler version that avoided the combinatoric evaluation of $\Omega(m)$, directly obtaining the maximization of $\phi(\beta, h, m)$.

Chapter 3

The Random Field Ising Model

This is the first solution to the Replica Model we will see. In Chapter 2 we introduced basic concepts such as understanding how to compute Z by means of field theory and the cavity trick. Now, we will discuss the concept of disorder and the replica method. All of the discussion will be based on a simple but more convoluted model.

Definition 3.1 (Random Field Ising Model (RFIM))**.** Consider the following Hamiltonian:

$$\mathscr{H}_{RFIM}(\mathbf{s}, \mathbf{h}) = -\frac{N}{2}\left(\frac{\sum_i^N s_i}{2}\right)^2 - \sum_i^N h_i s_i \quad h_i \overset{iid}{\sim} \mathcal{N}(0, \Delta) \tag{3.1}$$

Where we notice that unlike before h_i is inside the sum, and follows a normal distribution. On one hand, alignment is desired: the first term is minimized for $s_i = 1 \forall i \wedge -1 \forall i$. On the other hand, we wish that the s_i align to the respective h_i, with a *probabilistic* misalignment.

Assumption 3.2 (Averaging Notation)**.** Whenever we average over the Boltzmann distribution we use the symbol $\langle \cdot \rangle$, otherwise, we use $\mathbb{E}$.

The question of determining the minimum energy configuration clearly depends on $\mathbf{h}$. Following the discussion made in the definition, if $\Delta = 0$ we get a Curie Weiss Model as in Equation 2.14, while if $\Delta \to \infty$ the first term is dominated by the second.

In a thermodynamic fashion, we will inspect the behavior $\forall \Delta$ as $N \to \infty$. By the discussion of Section 2.3, it is sufficient to compute Φ. Recalling Definition 2.9:

$$\log\left[Z(\beta, \Delta, \mathbf{h})\right] = \log\left[\sum_{\{\mathbf{s}\}} e^{-\beta\mathscr{H}}\right] \tag{3.2}$$

$$\implies \frac{\partial}{\partial\beta}\log(Z) = \frac{1}{Z}\frac{\partial Z}{\partial\beta} \tag{3.3}$$

$$= \frac{\sum_{\{\mathbf{s}\}} \mathscr{H}(\mathbf{s}) e^{-\beta\mathscr{H}(\mathbf{s})}}{Z} \tag{3.4}$$

$$= -\langle \mathscr{H}(\mathbf{s}) \rangle \tag{3.5}$$

We have that by Proposition 2.2 the Boltzmann distribution will concentrate on the minimum energy configuration as $\beta \to \infty$. What about $\log(Z)$? The sum involves 2^N terms and is inherently hard to compute. Moreover, the h_i are random, and if we sampled once, the validity would not extend to different samples.

For this reason, we instead inspect $\mathbb{E}_{\mathbf{h}}[\mathscr{H}_{min}(\mathbf{s})]$ or equivalently:

$$\Phi^{avg}(\beta, \Delta) := \lim_{N\to\infty}\left\{\mathbb{E}_{\mathbf{h}}\left[\Phi_N(\beta, \Delta, \mathbf{h})\right]\right\} = \lim_{N\to\infty}\left\{\mathbb{E}_{\mathbf{h}}\left[\frac{\log(Z(\beta, \Delta, \mathbf{h}))}{N}\right]\right\} \tag{3.6}$$

claiming that, as $N \to \infty$ it will not be much different than the minimum energy configuration. In physics, this is referred to as **self-averaging**. The next theorem formally proves our claim.

Lemma 3.3 (Gaussian Poincaré Inequalities)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be smooth and* $X \sim \mathcal{N}(0, \Gamma)$. *Then:*

$$V[f(X)] \leq \mathbb{E}[\langle \Gamma \nabla f(X) \cdot \nabla f(X) \rangle] \tag{3.7}$$

Theorem 3.4 (Self-Averaging)**.**

$$V\left[\Phi_N(\Delta, \beta, \boldsymbol{h})\right] \leq \frac{\Delta \beta^2}{N} \tag{3.8}$$

Proof. Using Lemma 3.3 we observe that:

$$\begin{aligned}
\frac{\partial \Phi_N(\mathbf{h}, \beta)}{\partial h_i} &= \frac{1}{N} \frac{\partial \log[Z(\mathbf{h}, \beta)]}{\partial h_i} = \frac{1}{N} \frac{1}{Z} \frac{\partial Z(\mathbf{h}, \beta)}{\partial h_i} \\
&= \frac{1}{N} \frac{e^{-\beta \mathscr{H}(\mathbf{s}}}{Z} (\beta s_i) \\
&= \frac{\beta}{N} \langle S_i \rangle
\end{aligned}$$

Thus, developing the LHS:

$$\nabla \Phi_N(\beta, \mathbf{h}) \cdot \nabla \Phi_N(\beta, \mathbf{h}) = \begin{bmatrix} \frac{\beta}{N} \langle S_1 \rangle \\ \cdots \\ \frac{\beta}{N} \langle S_N \rangle \end{bmatrix} \cdot \begin{bmatrix} \frac{\beta}{N} \langle S_1 \rangle \\ \cdots \\ \frac{\beta}{N} \langle S_N \rangle \end{bmatrix} = \frac{\beta^2}{N} \sum_i \frac{\langle S_i \rangle^2}{N} \leq \frac{\beta^2}{N}$$

As $\sum_i \frac{\langle S_i \rangle^2}{N} \leq 1$. Then:

$$V[\Phi_N(\Delta, \beta, \mathbf{h})] \leq \mathbb{E}_{\mathbf{h}}[\langle \Delta \nabla \Phi_N(\beta, \mathbf{h}) \cdot \nabla \Phi_N(\beta, \mathbf{h}) \rangle] = \frac{\Delta \beta^2}{N}$$

□

By Theorem 3.4 we conclude that:

$$\mathop{lim}_{N \to \infty} V[\Phi_N(\Delta, \beta, \mathbf{h})] = 0 \implies \Phi^{avg}(\beta, \Delta) = \lim_{N \to \infty} \mathbb{E}_{\mathbf{h}} \Phi_N(\beta, \mathbf{h}, \Delta) \to \Phi(\beta, \Delta, \mathbf{h}) \tag{3.9}$$

And it is sufficient to compute the expectation of the logarithm:

$$\mathbb{E}_{\mathbf{h}}\left[\frac{\log(Z(\beta, \mathbf{h})}{N}\right] \tag{3.10}$$

3.1 The Replica Heuristic

This method was proposed by Edwards and Kac, and later served as the basis for the research contributions of Parisi and Mezard [MPV86].

Using a Taylor approximation:

$$n \;\; small \implies Z^n = e^{n \log(Z)} \cong 1 + n \log(Z) + O(n^2) \implies \log(Z) = \frac{Z^n - 1}{n} \;\; for \;\; n \to 0 \tag{3.11}$$

Thus, we could argue that by deliberately exhanging limit and expectation:

$$\mathbb{E}_{\mathbf{h}}[\log(Z)] = \lim_{n\to 0} \frac{\mathbb{E}_{\mathbf{h}}[Z^n] - 1}{n} \tag{3.12}$$

And we aim to calculate Z^n for $n \in \mathbb{N}^+$, generalizing it for $n \in \mathbb{R}$. Obviously, the generalization and the limit/expectation switch are seemingly not rigorous, but it turns out that the solution is always equivalent to exact computations, whenever they are available. This trick lays the basis for the replica method.

Definition 3.5 (Replica Method Steps)**.** To compute $\mathbb{E}_{\mathbf{h}}\left[\frac{\log(Z_N(\beta, \mathbf{h}, \Delta)}{N}\right] = \Phi_N(\beta, \Delta)$:

1. compute $\mathbb{E}_{\mathbf{h}}[Z^n]$ for $n \in \mathbb{N}^+$
2. Assume it is valid for $n \in \mathbb{R}$
3. let $n \to 0$ in the limit and apply Equation 3.12

Computing the powers of Z we index them by $\alpha = 1, \dots, n$, and follow the steps for the Curie-Weiss model of Chapter 2

$$Z^n = \left(\sum_{\{\mathbf{s}\}} e^{-\beta\mathscr{H}(\mathbf{s})}\right)^n \tag{3.13}$$

$$= \prod_{\alpha=1}^{n} \left(\sum_{\{\mathbf{s}^{(\alpha)}\}} e^{-\beta\mathscr{H}(\mathbf{s}^{(\alpha)})} \right) \tag{3.14}$$

$$= \prod_{\alpha=1}^{n} \left(\sum_{\{\mathbf{s}^{(\alpha)}\}} \int dm^{(\alpha)} \delta\left(Nm^{(\alpha)} - \sum_i s_i^{(\alpha)}\right) e^{\beta\frac{N}{2}(m^{(\alpha)})^2 + \beta\sum_i h_i s_i^{(\alpha)}} \right) \quad \text{fixing magnetization} \tag{3.15}$$

$$= \sum_{\{\mathbf{s}^{(\alpha)}\}_{\alpha=1}^n} \int \prod_{\alpha=1}^{n} \left[dm^{(\alpha)} \delta\left(Nm^{(\alpha)} - \sum_i s_i^{(\alpha)}\right)\right] e^{\beta\frac{N}{2}\sum_\alpha(m^{(\alpha)})^2 + \beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}} \Big) \tag{3.16}$$

$$= \sum_{\{\mathbf{s}^{(\alpha)}\}_{\alpha=1}^n} \int \prod_{\alpha=1}^{n} \left[dm^{(\alpha)} d\widehat{m}^{(\alpha)}\right] e^{\sum_\alpha \widehat{m}^{(\alpha)}[\sum_i s_i^{(\alpha)} - Nm^{(\alpha)}]} e^{\beta\frac{N}{2}\sum_\alpha(m^{(\alpha)})^2 + \beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}} \quad \text{Fourier} \tag{3.17}$$

Were it is useful recalling that $\widehat{m}^{(\alpha)} = 2\pi i\lambda^{(\alpha)}$. The expectation with respect to the $\mathbf{h}$ variables can be moved by linearity of $\mathbb{E}$, as the other terms are not dependent on $\mathbf{h}$:

$$\mathbb{E}_{\mathbf{h}}[Z^n] = \int \prod_{\alpha=1}^{n} \left[dm^{(\alpha)} d\widehat{m}^{(\alpha)}\right] e^{\beta\frac{N}{2}\sum_\alpha(m^{(\alpha)})^2 - N\sum_\alpha \widehat{m}^{(\alpha)} m^{(\alpha)}} \mathbb{E}_{\mathbf{h}}\left[\sum_{\{\mathbf{s}^{(\alpha)}\}_{\alpha=1}^n} e^{\sum_\alpha \widehat{m}^{(\alpha)}\sum_i s_i^{(\alpha)} + \beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}} \right] \tag{3.18}$$

Where for the term inside the expectation it can be argued that:

$$\sum_{\{\mathbf{s}^{(\alpha)}\}_{\alpha=1}^{n}} e^{\sum_\alpha \widehat{m}^{(\alpha)} \sum_i s_i^{(\alpha)} + \beta \sum_\alpha \sum_i h_i s_i^{(\alpha)}} = \sum_{\{\mathbf{s}^{(\alpha)}\}_{\alpha=1}^{n}} \prod_{\alpha=1}^{n} \prod_{i=1}^{N} e^{\widehat{m}^{(\alpha)} s_i^{(\alpha)} + \beta h_i s_i^{(\alpha)}} \tag{3.19}$$

$$= \prod_{i=1}^{N} \prod_{\alpha=1}^{n} \sum_{s_i^{(\alpha)}=\pm 1} e^{\widehat{m}^{(\alpha)} s_i^{(\alpha)} + \beta h_i s_i^{(\alpha)}} \tag{3.20}$$

$$= \prod_{i=1}^{N} \prod_{\alpha=1}^{n} \sum_{s_i^{(\alpha)}=\pm 1} e^{s_i^{(\alpha)} (\widehat{m}^{(\alpha)} + \beta h_i)} \tag{3.21}$$

$$= \prod_{i=1}^{N} \prod_{\alpha=1}^{n} 2 \cosh(\widehat{m}^{(\alpha)} + \beta h_i) \tag{3.22}$$

Where in the second line the sum of products is expressed as a product of sums, which is a direct consequence of:

- commutative products
- an adaptation of Lemma 2.17 in its middle steps

And we eventually recover a hyperbolic form as in the result of Lemma 2.17. Going back to the expectation:

$$\mathbb{E}_{\mathbf{h}}[Z^n] = \int \prod_{\alpha=1}^{n} \left[dm^{(\alpha)} d\widehat{m}^{(\alpha)} \right] e^{\beta \frac{N}{2} \sum_\alpha (m^{(\alpha)})^2 - N \sum_\alpha \widehat{m}^{(\alpha)} m^{(\alpha)}} \left\{ \mathbb{E}_h \left[\prod_{\alpha=1}^{n} 2 \cosh(\widehat{m}^{(\alpha)} + \beta h) \right] \right\}^N \quad h_i \overset{i.i.d.}{\sim} \mathcal{N}(\tag{3.23}$$

$$= \int \prod_{\alpha=1}^{n} \left[dm^{(\alpha)} d\widehat{m}^{(\alpha)} \right] e^{N \left\{ \beta \frac{1}{2} \sum_\alpha (m^{(\alpha)})^2 - \sum_\alpha \widehat{m}^{(\alpha)} m^{(\alpha)} + \log[\mathbb{E}_h (\prod_{\alpha=1}^{n} 2 \cosh(\widehat{m}^{(\alpha)} + \beta h))] \right\}} \tag{3.24}$$

Note that in the first passage the expectation is turned to being over scalars h, and not the whole vector $\mathbf{h}$, as we can factorize it into N equal expectations by the *iid* assumption. The last version is somewhat more comfortable as we bring the expectation inside the exponential with $N \log$ factor in front, and collect the N factor across all terms.

Note that the integral is over $m^{(\alpha)} \in \mathbb{R}^N, \widehat{m}^{(\alpha)} \in \mathbb{C}^N$. It is also clear that we will implement a saddle point by the form of Equation 3.24. A reasonable[1] assumption is used to simplify the job.

Assumption 3.6 (Replica Symmetry Ansatz)**.** For an estimation problem as that of Equation 3.24, let:

$$m^{(\alpha)} = m \quad \widehat{m}^{(\alpha)} = \widehat{m} \quad \forall \alpha \in \{1, \dots, n\} \tag{3.25}$$

Assumption 3.6 leads us to the following integral:

$$\mathbb{E}_{\mathbf{h}}[Z^n] = \int dm d\widehat{m} \; e^{N \left[\frac{\beta}{2} n m^2 - n \widehat{m} m + \log(\mathbb{E}_h [2^n \cosh^n(\beta h + \widehat{m})]) \right]} \tag{3.26}$$

[1] Looks like it is not, but it works!

And the formula for the average free energy becomes:

$$\Phi(\beta,\Delta) = \lim_{N\to\infty}\left\{\frac{1}{N}\log[Z(\beta,\mathbf{h})]\right\} \tag{3.27}$$

$$= \lim_{N\to\infty}\left\{\frac{1}{N}\lim_{n\to 0}\left[\frac{\mathbb{E}_{\mathbf{h}}Z(\beta,\mathbf{h})^n-1}{n}\right]\right\} \qquad \text{replica trick} \tag{3.28}$$

$$= \lim_{n\to 0}\frac{1}{n}\lim_{N\to\infty}\frac{\mathbb{E}_{\mathbf{h}}Z(\beta,\mathbf{h})^n-1}{N} \qquad \text{exchange limits} \tag{3.29}$$

$$= \lim_{n\to 0}\left\{\frac{1}{n}\underset{m,\widehat{m}}{Ext}\left[N\left(\frac{\beta}{2}nm^2-n\widehat{m}m+\log(\mathbb{E}_h[2^n\cosh^n(\beta h+\widehat{m})])\right)\right]\right\} \qquad \text{saddle point method} \tag{3.30}$$

Where the red comment is a non rigorous step.

It is now possible to apply a second time the replica trick to get rid of the n powers inside the logarithm. Indeed:

$$\begin{aligned}\mathbb{E}[X^n] &= \mathbb{E}[e^{\log(X^n)}] && (3.31)\\ &\overset{Taylor}{\approx} \mathbb{E}[1+\log(X^n)] && (3.32)\\ &= 1+n\mathbb{E}[\log(X)] && (3.33)\\ &\overset{inverse\ Taylor}{\approx} e^{n\mathbb{E}[\log(X)]} && (3.34)\end{aligned}$$

Where by inverse Taylor we just mean the Taylor contraction from linearization to function.

$$\implies \log(\mathbb{E}[X^n]) \approx \log(1+n\mathbb{E}[\log(X)]) \approx n\mathbb{E}[\log(X)] \tag{3.35}$$

$$\implies \log\Big(\mathbb{E}_h[(2\cosh(\beta h+\widehat{m}))n]\Big) \approx n\mathbb{E}_h\Big[\log(2\cosh(\beta h+\widehat{m}))\Big] \tag{3.36}$$

Using it inside the extremum as claimed before we will get that:

$$\begin{aligned}\Phi(\beta,\Delta) &= \underset{m,\widehat{m}}{Ext}\left[\frac{\beta}{2}nm^2-n\widehat{m}m+n\mathbb{E}_h[\log(2\cosh(\beta h+\widehat{m}))]\right] && (3.37)\\ &= \underset{m,\widehat{m}}{Ext}\left[\frac{\beta}{2}m^2-\widehat{m}m+\mathbb{E}_h[\log(2\cosh(\beta h+\widehat{m}))]\right] && \text{delete } n \perp\!\!\!\perp m,\widehat{m} \quad (3.38)\\ &= \underset{m,\widehat{m}}{Ext}\Big[M(m,\widehat{m})\Big] && \text{just notation} \quad (3.39)\end{aligned}$$

Taking the derivative with respect to m:

$$\frac{\partial M(m,\widehat{m})}{\partial m} = \beta m-\widehat{m} = 0 \implies \widehat{m} = \beta m \tag{3.40}$$

Which plugged into Equation 3.38 becomes:

$$\begin{aligned}\Phi(\beta,\Delta) &= \underset{m,\widehat{m}}{Ext}\left[\frac{\beta}{2}m^2-\beta m^2+\mathbb{E}_h[\log(2\cosh(\beta h+\beta m))]\right] && (3.41)\\ &= \underset{m}{Ext}\left[-\frac{\beta m^2}{2}+\mathbb{E}_h[\log(2\cosh(\beta(h+m)))]\right] && (3.42)\\ &= \underset{m}{Ext}\Big[\Phi_{RS}(m)\Big] = \Phi_{RS}(m^*) && \text{just notation} \quad (3.43)\end{aligned}$$

Where we can further impose the second extremality condition:

$$\frac{\partial M(m)}{\partial m} = -\beta m + \mathbb{E}_h\left[\frac{1}{2\cosh(\beta(h+m))} 2\sinh(\beta(h+m))(\beta)\right] \tag{3.44}$$
$$= -m + \mathbb{E}_h[\tanh(\beta(h+m))] \qquad \beta \neq 0 \tag{3.45}$$
$$= 0 \implies m = \mathbb{E}_h[\tanh(\beta(h+m))] \tag{3.46}$$

If we could come back to the $\underset{m,\widehat{m}}{Ext}$ point, since $\widehat{m}$ is a special function where $\widehat{m} = 2i\pi\lambda$ we could get by first imposing the saddle point condition on $\widehat{m}$ that:

$$m = \mathbb{E}_h[\tanh(\beta h + \widehat{m})] \tag{3.47}$$

And it is possible to derive that the replica symmetry approach has a large deviation behavior:

$$\mathbb{P}(\overline{\mathbf{S}} = m) \asymp e^{N\Phi(m,\beta,\Delta)} = e^{N[\phi(m) - \max_m(\phi(m))]} \tag{3.48}$$
$$\Psi(m,\beta,\Delta) = \frac{\beta m^2}{2} - m\widehat{m}^* + \mathbb{E}_h[\log(2\cosh(\beta h + \widehat{m}^*))] \tag{3.49}$$
$$m = \mathbb{E}_h\left[\tanh(\beta h + \widehat{m}^*\right] \tag{3.50}$$

3.2 Rigorous RFIM solution by the Interpolation Method

As previously argued, many steps are questionable. We do one limit exchange, Assumption 3.6, the Replica method approximation, and the $n \in \mathbb{N}^+ \to \mathbb{R}$ generalization with no apparent justification.

This section is devoted to proving that for the RFIM of Definition 3.1 the solution proposed is exact. To do so, we will use the Interpolation Method developed by Francesco Guerra [Gue03].

Instead of solving the hard version of this problem, we consider a simpler formulation:

$$\mathscr{H}_0(\mathbf{s},\mathbf{h},m) = -\sum_i s_i(h_i + m) \tag{3.51}$$

$$Z_0(\beta,\mathbf{h},m) = \sum_{\{\mathbf{s}\}} e^{-\beta\mathscr{H}_0} = \prod_{i=1}^{N}\sum_{s_i=\pm 1} e^{\beta s_i(h_i+m)} = \prod_{i=1}^{N} 2\cosh[\beta(h_i+m)] \tag{3.52}$$

$$\Phi_0(\beta,\Delta,m) = \lim_{N\to\infty}\left\{\mathbb{E}_{\mathbf{h}}\left[\frac{\log(Z_0)}{N}\right]\right\} = \mathbb{E}_{\mathbf{h}}\left[\frac{\sum_i 2\cosh[\beta(h_i+m)]}{N}\right] \overset{i.i.d.}{=} \mathbb{E}_h\left[2\cosh[\beta(h+m)]\right] \tag{3.53}$$

Where the partition sum naturally arises from the hamiltonian and the free energy form[2] is a consequence of the $\mathbf{h}$ vector being independent and identically distributed (i.e. the N at the denominator cancels out).

We opt to solve a slighltly more complicated problem at fixed magnetization, to explore the large deviation behavior:

$$Z_0(\beta,\mathbf{h},\overline{\mathbf{S}} = m) = \sum_{\{\mathbf{s}\}} \mathbf{1}(\overline{\mathbf{s}} = m) e^{-\beta\mathscr{H}_0} \tag{3.54}$$

[2]Again, notice that the last expectation is over h and not $\mathbf{h}$!

Equation 3.54 is not entirely trivial but the Legendre Transform is of help. We first define the new partition Hamiltonian (like in Equation 2.57), which modifies the partition function:

$$\widetilde{Z_0}(\beta, \mathbf{h}, m, k) = \sum_{\{\mathbf{s}\}} e^{\beta \sum_i s_i(h_i+m)+k\sum_i s_i} \tag{3.55}$$

$$\implies \widetilde{\Phi_0}(\beta, m, \Delta, k) \to \mathbb{E}_h\Big[2\cosh[\beta(m+h)+k]\Big] \tag{3.56}$$

And doing a Legendre transform with $s = m$:

$$\Phi_0(\beta, m, \Delta) = \lim_{N\to\infty}\left[\frac{\log[Z_0(\beta, \mathbf{h}, m, \overline{\mathbf{S}} = m)]}{N}\right] = \mathbb{E}_h\Big[\log[2\cosh(\beta(h+m)+k^*)]\Big] - k^*m \tag{3.57}$$

$$= \widetilde{\Phi_0}(\beta, m, \Delta, k^*) - k^*m \tag{3.58}$$

$$= Ext_k\Big[\widetilde{\Phi_0}(\beta, m, \Delta, k) - km\Big] \tag{3.59}$$

$$\implies m = \mathbb{E}_h\Big[\tanh(\beta(h+m)+k^*)\Big] \tag{3.60}$$

Where, if we change variables $k^* = \widehat{m} - \beta m$:

$$\Phi_0(\beta, m, \Delta) = Ext_{\widehat{m}}\left\{\mathbb{E}_h\Big[\log[2\cosh(\beta h + \widehat{m})]\Big] - m\widehat{m} + \beta m^2\right\} \tag{3.61}$$

Which is close to the desired result!

In this context, Guerra's Interpolation Method provides a solution. First of all, an *abstractly* time dependent[3] Hamiltonian is considered:

$$\mathscr{H}_t(\mathbf{s}, \mathbf{h}, m) = -\sum_i s_i[h_i + m(1-t)] - t\frac{N}{2}\left(\sum_i \frac{s_i}{N}\right)^2 \tag{3.62}$$

$$Z_t(\beta, \mathbf{h}, m) = \sum_{\{\mathbf{s}\}} \mathbf{1}(\overline{\mathbf{S}} = m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)} \quad t \in [0,1] \tag{3.63}$$

Where:

- For $t = 0$ we get the simplified model of Equations 3.51, 3.52 & further
- For $t = 1$ we get the RFIM of Definition 3.1

We can then recover the free entropy density at $t = 1$ as:

$$\Phi(\beta, m, \Delta) = \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\frac{\log[Z_1(\beta, \mathbf{h}, m)]}{N}\right] \quad \text{RFIM } \Phi \tag{3.64}$$

$$= \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\frac{\log[Z_0(\beta, \mathbf{h}, m)]}{N} + \int_0^1 d\tau \frac{\partial}{\partial t}\frac{\log[Z_t(\beta, \mathbf{h}, m)]}{N}\right] \quad \text{fund Th. of calculus} \tag{3.65}$$

$$= \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\frac{\log[Z_0(\beta, \mathbf{h}, m)]}{N}\right] + \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau \frac{\partial}{\partial t}\frac{\log[Z_t(\beta, \mathbf{h}, m)]}{N}\right] \quad \text{linearity} \tag{3.66}$$

$$= \Phi_0(\beta, m, \Delta) + \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau \frac{\partial}{\partial t}\frac{\log[Z_t(\beta, \mathbf{h}, m)]}{N}\right] \quad \text{Eq. 3.53} \tag{3.67}$$

[3] Basically: parametrized

Where we only need to compute the derivative and its integral:

$$\frac{\partial}{\partial t}\frac{\log[Z_t(\beta,\mathbf{h},m)]}{N} = \frac{1}{N}\frac{1}{Z_t(\beta,\mathbf{h},m)}\frac{\partial}{\partial t}\Big(Z_t(\beta,\mathbf{h},m)\Big) \tag{3.68}$$

$$= \frac{1}{N}\frac{1}{Z_t(\beta,\mathbf{h},m)}\frac{\partial}{\partial t}\Big(\sum_{\{\mathbf{s}\}}\mathbf{1}(\overline{\mathbf{S}}=m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)}\Big) \tag{3.69}$$

Eq 3.6

$$= \frac{1}{N}\frac{1}{Z_t(\beta,\mathbf{h},m)}\sum_{\{\mathbf{s}\}}\mathbf{1}(\overline{\mathbf{S}}=m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)}(-\beta)\frac{\partial}{\partial t}\Big(\mathscr{H}_t(\mathbf{s},\mathbf{h},m)\Big) \tag{3.70}$$

$$= \frac{1}{N}\frac{1}{Z_t(\beta,\mathbf{h},m)}\sum_{\{\mathbf{s}\}}\mathbf{1}(\overline{\mathbf{S}}=m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)}(-\beta) \tag{3.71}$$

$$\times\frac{\partial}{\partial t}\Big(-\sum_i s_i[h_i+m(1-t)] - t\frac{N}{2}(\sum_i\frac{s_i}{N})^2\Big)$$

Eq. 3.6

$$= \frac{1}{N}\frac{1}{Z_t(\beta,\mathbf{h},m)}\sum_{\{\mathbf{s}\}}\mathbf{1}(\overline{\mathbf{S}}=m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)}(\beta)\Big[-m\sum_i s_i+\frac{N}{2}\Big(\sum_i\frac{s_i}{N}\Big)^2\Big] \tag{3.72}$$

$$= \frac{1}{N}\sum_{\{\mathbf{s}\}}\underbrace{\frac{\mathbf{1}(\overline{\mathbf{S}}=m)e^{-\beta\mathscr{H}_t(\mathbf{s},\mathbf{h},m)}}{Z_t(\beta,\mathbf{h},m)}}_{=\mathbb{P}_{\beta,t,\mathbf{h},m}(\mathbf{s})}(\beta)\Big[-m\sum_i s_i+\frac{N}{2}\Big(\sum_i\frac{s_i}{N}\Big)^2\Big] \tag{3.73}$$

$$= \frac{\beta}{N}\sum_{\{\mathbf{s}\}}\mathbb{P}_{\beta,t,\mathbf{h},m}(\mathbf{s})\Big[-m\sum_i s_i+\frac{N}{2}\Big(\sum_i\frac{s_i}{N}\Big)^2\Big] \tag{3.74}$$

β out

$$= \beta\sum_{\{\mathbf{s}\}}\mathbb{P}_{\beta,t,\mathbf{h},m}(\mathbf{s})\Big[\frac{1}{N}(-m)\sum_i s_i+\frac{1}{N}\frac{N}{2}\Big(\sum_i\frac{s_i}{N}\Big)^2\Big] \tag{3.75}$$

N in

$$= -\beta m\sum_{\{\mathbf{s}\}}\mathbb{P}_{\beta,t,\mathbf{h},m}(\mathbf{s})\Big[\sum_i\frac{s_i}{N}\Big]+\frac{\beta}{2}\sum_{\{\mathbf{s}\}}\mathbb{P}_{\beta,t,\mathbf{h},m}(\mathbf{s})\Big[\Big(\sum_i\frac{s_i}{N}\Big)^2\Big] \tag{3.76}$$

split su

$$= \beta\Big\{-m\Big\langle\sum_i\frac{s_i}{N}\Big\rangle_{\beta,t,\mathbf{h},m}+\frac{1}{2}\Big\langle\Big(\sum_i\frac{s_i}{N}\Big)^2\Big\rangle_{\beta,t,\mathbf{h},m}\Big\} \tag{3.77}$$

$$= \beta\Big\{\frac{1}{2}\Big\langle\Big(m-\frac{\sum_i s_i}{N}\Big)^2\Big\rangle_{\beta,t,\mathbf{h},m}-\frac{1}{2}m^2\Big\} \tag{3.78}$$

Where in the last passage we complete the square of the binomial as:

$$\frac{1}{2}(x-y)^2-\frac{1}{2}y^2=-\frac{1}{2}2xy+\frac{1}{2}x^2$$

Integrating over $\tau \in [0,1]$, with limit and expectation:

$$\lim_{N\to\infty} \mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau \frac{\partial}{\partial t}\frac{\log[Z_t(\beta,\mathbf{h},m)]}{N}\right] = \lim_{N\to\infty} \mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau\beta\left\{\frac{1}{2}\left\langle\left(m-\frac{\sum_i s_i}{N}\right)^2\right\rangle_{\beta,t,\mathbf{h},m} - \frac{1}{2}m^2\right\}\right] \tag{3.79}$$

$$= -\beta\frac{m^2}{2} + \frac{1}{2}\lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau \underbrace{\left\langle\left(m-\frac{\sum_i s_i}{N}\right)^2\right\rangle_{\beta,t,\mathbf{h},m}}_{m=\sum_i \frac{s_i}{N}}\right] \tag{3.80}$$

$$= -\beta\frac{m^2}{2} + \frac{1}{2}\lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau\left\langle 0\right\rangle_{\beta,t,\mathbf{h}}\right] \tag{3.81}$$

$$= -\beta\frac{m^2}{2} \tag{3.82}$$

Where we exploited the indicator function $\mathbf{1}(\overline{\mathbf{S}} = m)$, which restricts magnetizations of the spins to the fixed m. Then:

$$\Phi(\beta,m,\Delta) = \Phi_0(\beta,m,\Delta) + \lim_{N\to\infty}\mathbb{E}_{\mathbf{h}}\left[\int_0^1 d\tau\frac{\partial}{\partial t}\frac{\log[Z_t(\beta,\mathbf{h},m)]}{N}\right] \qquad \text{Eq. 3.67} \tag{3.83}$$

$$= \Phi_0(\beta,m,\Delta) - \beta\frac{m^2}{2} \qquad \text{Eq. 3.82} \tag{3.84}$$

$$= Ext_{\widehat{m}}\left\{\mathbb{E}_h\left[\log[2\cosh(\beta h + \widehat{m})]\right] - m\widehat{m} + \beta m^2\right\} - \beta\frac{m^2}{2} \qquad \text{Eq. 3.61} \tag{3.85}$$

$$= Ext_{\widehat{m}}\left\{\mathbb{E}_h\left[\log[2\cosh(\beta h + \widehat{m})]\right] - m\widehat{m} + \beta\frac{m^2}{2}\right\} \tag{3.86}$$

Which is the same as the results of Section 3.1 after imposing the condition $\frac{\partial M(m,\widehat{m})}{\partial m} = \beta m - \widehat{m} = 0 \implies \widehat{m} = \beta m$. Using the notation of Equation 3.43, we can eventually state that:

$$\Phi(\beta,m,\Delta) = Ext_{\widehat{m}}\left\{\Phi_{RS}(m)\right\} = \Phi_{RS}(m^*) \tag{3.87}$$

And the solutions coincide!

To summarize, we conclude that:

- there is a way to compute the free energies with a heuristic, the Replica Method
- despite the questionable steps, it coincides with the exact solution
- not all replica solutions are guaranteed as some models have no known exact form
- notice that there is no overlap, and disorder is not in the couplings. If this were the case, there would be no way to obtain the replica **symmetric** solution. This problem will be dealt with in later discussions, when considering Replica Symmetry Breaking approaches.

To read about an extensive application of the Replica Method, see Appendix B.

Chapter 4

Graphical Models & Belief Propagation

The purpose of this chapter is introducing a well known heuristic technique known as **Belief Propagation** (BP, from now on) to work on the problems we discussed in Chapter 1. For this purpose, it is worth discussing graphical models first, as the working framework of reference.

4.1 Graphical Models

Graphs are well known objects in multiple fields. The next statement quickly summarizes the notation used:

Definition 4.1 (Graphs Notation)**.** A graph $G(V, E)$ is an object with:

- $i \in V$ vertices (nodes), where $|V| = N$
- $(ij) \in E$ edges (connections), where $|E| = M$

It is represented through an adjacency matrix $A \in \{0, 1\}^{N \times N}$ where:

$$A_{ij} = \begin{cases} 1 & (ij) \in E \\ 0 & otherwise \end{cases} \tag{4.1}$$

A very common operator associated to a graph is its neighborhood function $\partial : V \to 2^V$ where:

$$\partial(i) = \partial i = \{j \mid (ij) \in E\} \tag{4.2}$$

Another widely used function is the degree function $d : V \to \mathbb{N}^+$ where:

$$d(i) = d_i = \#\Big(j \in V \mid (ij) \in E\Big) = \sum_{j=1}^{N} A_{ij} \tag{4.3}$$

Where obviously $d_i = |\partial i|$.

Having introduced graphs in general, it is now easier to reason about Factor Graphs.

Definition 4.2 (Factor Graphs)**.** Factor graphs are bipartite graphs (with disjoint vertex sets, i.e. no edges within, only "across"). One set is that of variable nodes i, j, k (circles), another is composed of factor nodes a, b, c (squares). We further have that:

- variable nodes represent variables $s_i \in \Lambda$, a feature space, where $i \in \{1, \dots, N\}$

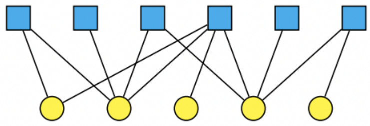


Figure 4.1: Factor graph in *Bipartite* view

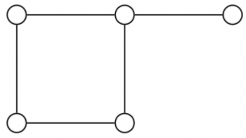


Figure 4.2: Graph, Example 4.4

- factor nodes represent **non negative** functions $f_a(\{s_i\}_{i\in\partial a})$, where $a \in \{1,\dots,M\}$

Observation 4.3 (On the bipartiteness of factor graphs)**.** *A requirement for factor graphs is being bipartite. This ensures that the domain of the factor nodes f_a is well specified, since ∂a will be made of only variable nodes (neighbors of the factor). Similarly ∂i is only made of factor nodes! Up to orientation, each of the instances of this model can be seen as per the usual bipartite graph visualization (Figure 4.1)*

The main purpose of factor graphs is representing probability distributions. For a set of samples $\{s_i\}_{i=1}^N$ where $s_i \in \Lambda \forall i$ we could have a distribution of the form:

$$\mathbb{P}(S = \{s_i\}_{i=1}^N) = \frac{1}{Z_N}\prod_{a=1}^{M} f_a(\{s_j\}_{j\in\partial a}) \tag{4.4}$$

$$Z_N = \sum_{i=1}^{N}\prod_{a=1}^{M} f_a(\{S_j\}_{j\in\partial a}) \tag{4.5}$$

In order to better understand how the factors arise, some intuitive examples are added below.

Example 4.4 (Spin Glass)**.** A spin glass model generalizes the Curie Weiss model of Equation 2.14, Chapter 2. Its Hamiltonian is of the form:

$$\mathscr{H}(\{s_i\}_{i=1}^N) = -\sum_{(ij)\in E} J_{ij}s_is_j - \sum_i h_is_i \quad J_{ij}\ \textit{interactions} \quad h_i\in\mathbb{R}\ \textit{magnetic field} \tag{4.6}$$

The target of the analysis is finding a minimum energy configuration. We then assign a Boltzmann measure:

$$\mathbb{P}(\{s_i\}_{i=1}^N) = \frac{1}{Z_N}e^{-\beta\mathscr{H}(\{s_i\}_{i=1}^N)} \tag{4.7}$$

$$= \frac{1}{Z_N}\prod_{i=1}^{N} e^{\beta h_i s_i}\prod_{(ij)\in E} e^{\beta J_{ij}s_is_j} \tag{4.8}$$

Where the split highlights two different types of factor (square) nodes. The former is only related to the magnetic field $h_i \in \mathbb{R}$, the latter is joint between two nodes and is a function of the interaction J_{ij}. A visualization is found in Figures 4.2, 4.3. We have in total N variables nodes and $N + \dfrac{N(N+1)}{2}$ factor nodes (N from h_i, and all possible pairs $\frac{N(N+1)}{2}$ from J_{ij}).

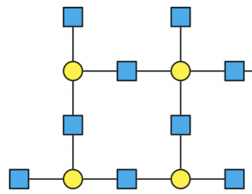


Figure 4.3: Factor graph, Example 4.4

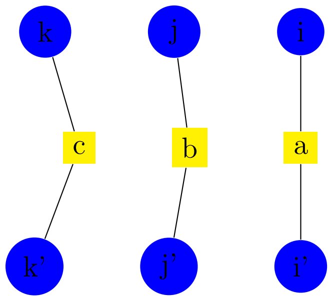


Figure 4.4: Factor graph, Example 4.5

Example 4.5 (Graph Coloring)**.** Discussed in Chapter 1, it is slightly different, as $s_i \in \{1, \dots, q\}$ is not restricted to being a binary choice. The problem of finding the number of proper colorings of a graph can be formalized as:

$$Z_N = \sum_{\{s_i\}_{i=1}^N} \prod_{(ij)\in E} (1 - \delta_{s_i,s_j}) \tag{4.9}$$

$$\delta(s_i, s_j) = \begin{cases} 1 & s_i = s_j \\ 0 & otherwise \end{cases} \tag{4.10}$$

Where clearly connected nodes with the same color cancel out the contribution of the $\{s_i\}$ configuration. We can then impose a uniform distribution over colorings as:

$$\mathbb{P}(\{s_i\}_{i=1}^N) = \frac{1}{Z_N} \prod_{(ij)\in E} 1 - \delta_{s_i,s_j} \tag{4.11}$$

Which is relaxed with a β factor to avoid the case in which $Z_N = 0$:

$$\mathbb{P}(\{s_i\}_{i=1}^N) = \frac{1}{Z_N(\beta)} \prod_{(ij)\in E} e^{-\beta\delta_{s_i,s_j}} \overset{\beta\to\infty\equiv T\to 0}{\to} \frac{1}{Z_N} \prod_{(ij)\in E} 1 - \delta_{s_i,s_j} \tag{4.12}$$

Its factor graph is like that of Example 4.4 without node factors, and with edge factors representing the function $e^{-\beta\delta_{s_i,s_j}}$. A minimal example is proposed in Figure 4.4.

Other such examples can be found in the original Lecture Notes [KZ21b].

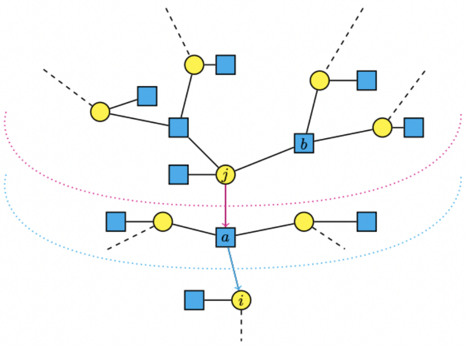


Figure 4.5: $R_{s_j}^{j\to a}, V_{s_i}^{a\to i}$ Partition Functions

4.2 Belief Propagation

The main topic of this Chapter can now be introduced. As experienced in Example 4.4, we will restrict our analysis to factor graphs representing probability laws of the form:

$$\mathbb{P}(\{s_i\}_{i=1}^N) = \frac{1}{Z}\prod_{i=1}^{N} g_i(s_i)\prod_{a=1}^{M} f_a(\{s_i\}_{i\in\partial a}) \tag{4.13}$$

$$Z = \sum_{\{s_i\}_{i=1}^N}\prod_{i=1}^{N} g_i(s_i)\prod_{a=1}^{M} f_a(\{s_i\}_{i\in\partial a}) \tag{4.14}$$

The restriction is reasonable. Such graphical models represent all probabilistic phenomenas that can be described by element/element and element/system interactions. As seen in Section 2.3, having access to the free entropy $\Phi = \frac{\log(Z)}{N}$ is sufficient.
Another interesting probabilistic function is that of the marginals of a specific node:

$$\mu_i(s_i) = \sum_{\{s_j\}_{j=1}^N\ j\neq i} \mathbb{P}(\{s_j\}_{j=1}^N) \tag{4.15}$$

The problem with computing Equations 4.14, 4.15 is that both sums are exponentially[1] large in N. While this result holds in general cases, this Chapter is devoted to deriving a method which is exact for trees (no loops enforced) and becomes a good heuristic for graphs, in linear time $O(N)$.

A first step towards this direction is laying the grounds for a more understandable version of the partition function of Equation 4.14.

Definition 4.6 (Auxiliary Partition Functions $R_{s_j}^{j\to a}, V_{s_i}^{a\to i}$)**.** For a factor Tree G , consider the following equations:

$$R_{s_j}^{j\to a} := g_j(s_j)\sum_{\{s_k\}_{k\ above\ j}}\ \prod_{k\ above\ j} g_k(s_k)\prod_{b\ above\ j} f_b(\{s_l\}_{l\in\partial b}) \tag{4.16}$$

$$V_{s_i}^{a\to i} := \sum_{\{s_j\}_{j\ above\ a}} f_a(\{s_k\}_{k\in\partial a})\prod_{j\ above\ a} g_j(s_j)\prod_{b\ above\ a} f_b(\{s_k\}_{k\in\partial b}) \tag{4.17}$$

Where $R_{s_j}^{j\to a}$ is the partition function of the tree above a variable/factor connection, with variable s_j fixed, while $V_{s_i}^{a\to i}$ is the partition function of the tree above a factor/variable connection with fixed s_i value in the neighbors $k \in \partial a$ of f_a.

[1]The former has q^N terms, the latter has q^{N-1} since one is fixed

Looking at Figure 4.5, $R_{s_j}^{j\to a}$ is the partition function of the system above the red dotted line, with node j having value s_j. Similarly, $V_{s_i}^{a\to i}$ is the partition function of the system above the blue dotted line, with variable i fixed to s_i.

The two custom created objects are easily related in tree models.

Lemma 4.7 (Linking $R_{s_j}^{j\to a}, V_{s_i}^{a\to i}$ in a Tree)**.** *Consider a factor tree G, then:*

$$\implies R_{s_j}^{j\to a} = g_j(s_j) \prod_{b\in\partial j\setminus a} V_{s_j}^{b\to j} \tag{4.18}$$

$$\implies V_{s_i}^{a\to i} = \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a} \tag{4.19}$$

For both equations, we restrict to ***immediate neighbors*** *of the node j/factor a but the one in which the edge is directed (i.e. not a/not i).*

Proof. **(Equation 4.18)** Node j is restricted to taking value s_j. By the graph being a tree, there are no loops above or below j. Moreover, the restriction to a precise value makes all the branches above **independent**. Again, for a tree, the only branch below contains a, all the others are above. Considering all the branches denoted by b the first factor node encountered (the immediate neighbor) we obtain:

$$R_{s_j}^{j\to a} = g_j(s_j) \overbrace{\prod_{b\in\partial j\setminus a}}^{\textit{indep branches}} \underbrace{\left[\sum_{\{s_k\}_{k \textit{ above } b}} f_b(\{s_k\}_{k\in\partial b}) \prod_{k \textit{ above } b} g_k(s_k) \prod_{c \textit{ above } b} f_c(\{s_l\}_{l\in\partial c})\right]}_{=V_{s_j}^{b\to j}} \tag{4.20}$$

$$= g_j(s_j) \prod_{b\in\partial j\setminus a} V_{s_j}^{b\to j} \tag{4.21}$$

Where we used Equation 4.17 from Definition 4.6.

(Equation 4.19) Using the same reasoning of the first Claim, we split the branches above a, which are all but i, to obtain:

$$V_{s_i}^{a\to i} = \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \tag{4.22}$$

$$\times \overbrace{\prod_{j\in\partial a\setminus i}}^{\textit{indep branches}} \left[\sum_{\{s_k\}_{k \textit{ above } j}} g_j(s_j) \prod_{k \textit{ above } j} g_k(s_k) \prod_{b \textit{ above } j} f_b(\{s_l\}_{l\in\partial b})\right] \quad j \text{ fixed in outer sum}$$

$$= \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \tag{4.23}$$

$$\times \prod_{j\in\partial a\setminus i} \underbrace{\left[g_j(s_j) \sum_{\{s_k\}_{k \textit{ above } j}} \prod_{k \textit{ above } j} g_k(s_k) \prod_{b \textit{ above } j} f_b(\{s_l\}_{l\in\partial b})\right]}_{=R_{s_j}^{j\to a}} \quad \text{taking out } j$$

$$= \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a} \tag{4.24}$$

Where we used Equation 4.16 from Definition 4.6. □

These relations suggest an interesting result for the partition function.

Theorem 4.8 (Partition Function for Tree Models)**.** *For a factor tree G, consider its leaves indexed by j. Then:*

$$Z = \sum_{s_j} g_j(s_j) \prod_{b \in \partial j} V_{s_j}^{b \to j} \tag{4.25}$$

Proof. For a given reference, a factor tree G has a set of leaves with features $\{s_j\}$ and a root. For nodes j, it holds that:

$$R_{s_j}^{j \to a} = g_j(s_j) \forall j$$

Since there are no factor nodes *outgoing* from j, and only one father factor node b feeds them. Recursing the relation to higher levels, it is possible to alternately recover all the layers of factor and variable nodes. Using Lemma 4.7, we will have that Equation 4.14 is simplified to:

$$Z = \sum_{s_j} g_j(s_j) \prod_{b \in \partial j} V_{s_j}^{b \to j} \tag{4.26}$$

Where the $V_{s_j}^{b \to j}$ can be expanded further using again Lemma 4.7.
Notice that in the proof we choose a *reference* for leaves and root, but this does not influence the specification of Z, since all factors are accounted anyway. □

Theorem 4.8 provides a method for trees to compute their partition function in linear time, namely $T(N) \in O(|V|)$. Yet, the size of V is tipically exponentially exploding, with a $exp(cN)$ rate[2]. To overcome the hardness of summing over an exponentially exploding number of terms we define messages (i.e. probabilities).

Definition 4.9 (Messages $\chi_{s_j}^{j \to a}, \psi_{s_i}^{a \to i}$)**.** Drawing from Definition 4.6, interpret $\chi_{s_j}^{j \to a}$ as the probability that variable j takes value s_j in the restricted system of predecessors. Work out a similar reasoning for $\psi_{s_i}^{a \to i}$.

$$\chi_{s_j}^{j \to a} = \frac{R_{s_j}^{j \to a}}{\sum_s R_s^{j \to a}} \qquad \sum_s \chi_s^{j \to a} = 1 \quad \forall (ja) \in E \tag{4.27}$$

$$\psi_{s_i}^{a \to i} = \frac{V_{s_i}^{a \to i}}{\sum_s V_s^{a \to i}} \qquad \sum_s \psi_s^{a \to i} \quad \forall (ia) \in E \tag{4.28}$$

This reworking of variables comes with the need to express previous results in terms of $R_{s_j}^{j \to a}, V_{s_i}^{a \to i}$, finding *self consistent equations.* First, we their connection as in Lemma

[2] At each split the number of nodes in the lower layer is at least doubled

4.7:

$$
\begin{aligned}
\chi_{s_j}^{j\to a} &= \frac{R_{s_j}^{j\to a}}{\sum_s R_s^{j\to a}} && \text{Def. 4.9} && (4.29)\\
&= \frac{g_j(s_j)\prod_{b\in\partial j\backslash a} V_{s_j}^{b\to j}}{\sum_s g_j(s)\prod_{b\in\partial j\backslash a} V_s^{b\to j}} && \text{Lem. 4.7} && (4.30)\\
&= \frac{g_j(s_j)\prod_{b\in\partial j\backslash a} V_{s_j}^{b\to j}}{\sum_s g_j(s)\prod_{b\in\partial j\backslash a} V_s^{b\to j}}\underbrace{\frac{\prod_{b\in\partial j\backslash a}\sum_{s'} V_{s'}^{b\to j}}{\prod_{b\in\partial j\backslash a}\sum_{s''} V_{s''}^{b\to j}}}_{=1} && && (4.31)\\
&= \frac{g_j(s_j)\prod_{b\in\partial j\backslash a}\frac{V_{s_j}^{b\to j}}{\sum_{s''} V_{s''}^{b\to j}}}{\sum_s g_j(s)\prod_{b\in\partial j\backslash a}\frac{V_s^{b\to j}}{\sum_{s'} V_{s'}^{b\to j}}} && \text{pushed in prod} && (4.32)\\
&= \frac{g_j(s_j)\prod_{b\in\partial j\backslash a}\psi_{s_j}^{b\to j}}{\sum_s g_j(s)\prod_{b\in\partial j\backslash a}\psi_s^{b\to j}} && \text{Def 4.9} && (4.33)\\
&= \frac{1}{Z^{j\to a}} g_j(s_j)\prod_{b\in\partial j\backslash a}\psi_{s_j}^{b\to j} && && (4.34)\\
\text{where}\quad Z^{j\to a} &= \sum_s g_j(s)\prod_{b\in\partial j\backslash a}\psi_{s_j}^{b\to j} && && (4.35)
\end{aligned}
$$

Similarly but slightly more convoluted:

$$\psi_{s_i}^{a\to i} = \frac{V_{s_i}^{a\to i}}{\sum_s V_s^{a\to i}} \qquad \text{Def 4.9} \tag{4.36}$$

$$= \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}} \qquad \text{Lem 4.7} \tag{4.37}$$

$$= \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}} \underbrace{\frac{\prod_{j\in\partial a\setminus i}\sum_{s'_j} R_{s'_j}^{j\to a}}{\prod_{j\in\partial a\setminus i}\sum_{s''_j} R_{s''_j}^{j\to a}}}_{=1} \tag{4.38}$$

$$= \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \frac{R_{s_j}^{j\to a}}{\sum_{s''_j} R_{s''_j}^{j\to a}}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \frac{R_{s_j}^{j\to a}}{\sum_{s'_j} R_{s'_j}^{j\to a}}} \qquad \text{pushed in p} \tag{4.39}$$

$$= \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}} \qquad \text{Def. 4.9} \tag{4.40}$$

$$= \frac{1}{Z^{a\to i}} \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} \tag{4.41}$$

$$where \quad Z^{a\to i} = \sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} = \sum_{\{s_j\}_{j\in\partial a}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} \tag{4.42}$$

Where in the last line we just notice that inner and outer sum cover all neighbors of the factor node a for any choice of j.

The marginals are now easier to compute. Consider a tree model where the variable node i has incoming factor nodes indexed by a, then:

$$\begin{aligned}
\mu_i(s_i) &= \frac{1}{Z} g_i(s_i) \prod_{a\in\partial i} V_{s_i}^{a\to i} && (4.43)\\
&= \frac{g_i(s_i)\prod_{a\in\partial i} V_{s_i}^{a\to i}}{\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}} && (4.44)\\
&= \frac{g_i(s_i)\prod_{a\in\partial i} V_{s_i}^{a\to i}}{\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}} \underbrace{\frac{\prod_{a\in\partial i}\sum_{s'} V_{s'}^{a\to i}}{\prod_{a\in\partial i}\sum_{s''} V_{s''}^{a\to i}}}_{=1} && (4.45)\\
&= \frac{g_i(s_i)\prod_{a\in\partial i}\frac{V_{s_i}^{a\to i}}{\sum_{s''} V_{s''}^{a\to i}}}{\sum_s g_i(s)\prod_{a\in\partial i}\frac{V_s^{a\to i}}{\sum_{s'} V_{s'}^{a\to i}}} && (4.46)\\
&= \frac{g_i(s_i)\prod_{a\in\partial i}\psi_{s_i}^{a\to i}}{\sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i}} && \text{Def 4.9} \quad (4.47)\\
&= \frac{1}{Z^{(i)}} g_i(s_i)\prod_{a\in\partial i}\psi_{s_i}^{a\to i} && (4.48)\\
\textit{where} \quad Z^{(i)} &= \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i} && (4.49)
\end{aligned}$$

Lastly, the result for Z is presented as a Theorem.

Theorem 4.10 (Partition Function for tree models, messages version)**.** *Consider a factor tree G. Define:*

$$\begin{aligned}
Z^{(i)} &= \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i} && (4.50)\\
Z^{(a)} &= \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a})\prod_{i\in\partial a}\chi_{s_i}^{i\to a} && (4.51)\\
Z^{(ia)} &= \sum_s \chi_s^{i\to a}\psi_s^{a\to i} && (4.52)
\end{aligned}$$

Then:

$$\implies Z = \frac{\prod_i Z^{(i)}\prod_a Z^{(a)}}{\prod_{(ia)} Z^{(ia)}} \tag{4.53}$$

Proof. Starting from the claim we slowly derive a known version of Z:

$$\frac{\prod_i Z^{(i)} \prod_a Z^{(a)}}{\prod_{(ia)} Z^{(ia)}} = \frac{\prod_{i=1}^N \sum_s g_i(s) \prod_{a\in\partial i} \psi_s^{a\to i} \prod_{a=1}^M \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a}) \prod_{i\in\partial a} \chi_{s_i}^{i\to a}}{\prod_{(ai)\in E} \sum_s \chi_s^{i\to a}\psi_s^{a\to i}} \tag{4.54}$$

$$= \frac{\prod_{i=1}^N \sum_s g_i(s) \prod_{a\in\partial i} \frac{V_{s_i}^{a\to i}}{\sum_s V_s^{a\to i}} \prod_{a=1}^M \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a}) \prod_{i\in\partial a} \frac{R_{s_i}^{i\to a}}{\sum_s R_s^{i\to a}}}{\prod_{(ai)\in E} \sum_s \frac{V_s^{a\to i}}{\sum_{s'} V_{s'}^{a\to i}} \frac{R_s^{i\to a}}{\sum_{s''} R_{s''}^{i\to a}}} \tag{4.55}$$

$$= \frac{\frac{\prod_{i=1}^N \sum_s g_i(s) \prod_{a\in\partial i} V_{s_i}^{a\to i}}{\prod_{(ia)\in E} \sum_s V_s^{a\to i}} \frac{\prod_{a=1}^M \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a}) \prod_{i\in\partial a} R_{s_i}^{i\to a}}{\prod_{(ia)\in E} \sum_s R_s^{i\to a}}}{\frac{\sum_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}}{\prod_{(ia)\in E} \sum_{s'} V_{s'}^{a\to i} \prod_{(ia)\in E} \sum_{s''} R_{s''}^{i\to a}}} \tag{4.56}$$

$$= \frac{\prod_{i=1}^N \sum_s g_i(s) \prod_{a\in\partial i} V_{s_i}^{a\to i} \prod_{a=1}^M \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a}) \prod_{i\in\partial a} R_{s_i}^{i\to a}}{\sum_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}} \tag{4.57}$$

$$= \left[\sum_s g_j(s) \prod_{a\in\partial j} V_s^{a\to j}\right]\left[\prod_{i\neq j, i=1}^N \sum_s g_i(s) V_s^{b\to i} \prod_{a\in\partial i\setminus b} V_s^{a\to i}\right] \tag{4.58}$$

$$\times \left[\prod_{a=1}^M \sum_{s_i} \sum_{\{s_k\}_{k\in\partial a\setminus i}} f_a(\{s_i\}_{i\in\partial a}) R_{s_i}^{i\to a} \prod_{k\in\partial a\setminus i} R_{s_k}^{k\to a}\right]$$

$$\times \left[\prod_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}\right]^{-1}$$

$$= \Bigg[\overbrace{\sum_s g_j(s) \prod_{a\in\partial j} V_s^{a\to j}}^{Z}\Bigg]\Bigg[\prod_{i\neq j, i=1}^N \sum_s V_s^{b\to i} \overbrace{g_i(s) \prod_{a\in\partial i} V_s^{a\to i}}^{R_s^{i\to b}}\Bigg] \tag{4.59}$$

$$\times \Bigg[\prod_{a=1}^M \sum_{s_i} R_{s_i}^{i\to a} \overbrace{\sum_{\{s_k\}_{k\in\partial a\setminus i}} f_a(\{s_i\}_{i\in\partial a}) \prod_{k\in\partial a\setminus i} R_{s_k}^{k\to a}}^{V_s^{a\to i}}\Bigg]$$

$$\times \left[\prod_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}\right]^{-1}$$

$$= Z \frac{\prod_{i=1, i\neq j}^N \sum_s R_s^{i\to b} V_s^{b\to i} \prod_{a=1}^M \sum_s V_s^{a\to i} R_s^{i\to a}}{\prod_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}} \tag{4.60}$$

$$= Z \tag{4.61}$$

Where:

- in the first passage we apply the definition of $\chi_{s_j}^{j\to a}, \psi_{s_i}^{a\to i}$ with appropriate indices
- in the second we collect together the products over all i and neighbors (i.e. all edges) and over all a and neighbors (i.e. all edges)
- in the third passage we cancel the highlighted elements as they are equal
- in the fourth step we split the products and sums into three, and further inside make explicit some products and sum indices j, i, where j is set as a root of the tree, and i are the non nodes at all the ther layers.

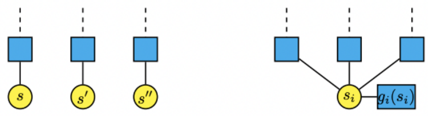


Figure 4.6: $Z^{(i)}$ graphically

Figure 4.7: $Z^{(a)}$ graphically

- which serve us in the fifth equality to identify definitions of already known objects by Lemma 4.7, Theorem 4.8
- eventually, we realize that the denominator and the numerator of the factor at the left of Z just cancel. Indeed, covering all variable nodes i but the root and all factors amounts to covering all the edges of the tree, and so we are summing over s inside, and doing the same product outside.

□

We stress the interpretation of the just introduced decomposition factors of the partition function for a factor graph:

- $Z^{(i)}$ is the change of Z when $g_i(s_i)$ is added (Figure 4.6)
- $Z^{(a)}$ is the change of Z when f_a is added (Figure 4.7)
- $Z^{(ia)}$ is the change of Z when $g_i(s_i), f_a$ are connected (Figure 4.8)

Observation 4.11 (Link between partition functions)**.** *We recognize that:*

- $Z^{j\to a} \sim Z^i$ *except we did not connect to* a
- $Z^{a\to i} \sim Z^a$ *except* i *is not present*

The result of Theorem 4.10 is in simple terms a decomposition of the energy configuration of a system into:

- the sum of factors and variables
- minus the edge connections (ia) which are counted twice in the sum

A summary of all the results is found in the box below.

Summary of Belief Propagation Equations Consider a graphical model with distribution:

$$\mathbb{P}(\{s_i\}_{i=1}^{N}) = \frac{1}{Z}\prod_{i=1}^{N} g_i(s_i)\prod_{a=1}^{M} f_a(\{s_i\}_{i\in\partial a}) \tag{4.62}$$

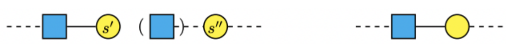


Figure 4.8: $Z^{(ia)}$ graphically

The BP messages are:

$$\chi_{s_j}^{j\to a} = \frac{1}{Z^{j\to a}} g_j(s_j) \prod_{b\in\partial j\setminus a} \psi_{s_j}^{b\to j} \qquad \sum_s \chi_s^{j\to a} = 1 \tag{4.63}$$

$$\psi_{s_i}^{a\to i} = \frac{1}{Z^{a\to i}} \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a(\{s_j\}_{j\in\partial a}) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} \qquad \sum_s \psi_s^{a\to i} = 1 \tag{4.64}$$

The free entropy density Φ is recovered using the identity:

$$\Phi = \frac{\log(Z)}{N} = \frac{1}{N}\left[\sum_i \log(Z^{(i)}) + \sum_a \log(Z^{(a)}) - \sum_{(ia)} \log(Z^{(ia))}\right] \tag{4.65}$$

Where:

$$Z^{(i)} = \sum_s g_i(s) \prod_{a\in\partial i} \psi_s^{a\to i}, \quad Z^{(a)} = \sum_{\{s_i\}_{i\in\partial a}} f_a(\{s_i\}_{i\in\partial a}) \prod_{i\in\partial a} \chi_{s_i}^{i\to a}, \tag{4.66}$$

$$Z^{(ia)} = \sum_s \chi_s^{i\to a} \psi_s^{a\to i} \tag{4.67}$$

And marginals are recovered using:

$$\mu_i(s_i) = \frac{1}{Z} g_i(s_i) \prod_{a\in\partial i} \psi_{s_i}^{a\to i} \tag{4.68}$$

We can now exploit another example where this formalization arises naturally to build our heuristic algorithm.

Example 4.12 (Generalized Linear Model)**.** Consider n samples indexed by μ of the form:

$$X_\mu \in \mathbb{R}^{n\times d} \quad y_\mu \in \{-1,1\}^n$$

Generalized Linear Regression aims at minimizing a parametrized loss of the form:

$$\mathcal{L}(\mathbf{w}) = \sum_{\mu=1}^{n} \ell(y_\mu, X_\mu \cdot \mathbf{w}) + \sum_{i=1}^{d} r(w_i)$$

Where the second term is a regularization term. To represent this as a factor graph we endow the parameters with a Boltzmann measure and let $\beta \to \infty$, which by Proposition 2.2, means that the distribution will concentrate at the minimum value of $\mathcal{L}(\cdot)$. Thus, we let:

$$\begin{aligned}
\mathbb{P}(\mathbf{w}) &= \frac{1}{Z_N(\mathbf{X},\mathbf{y},\beta)} e^{-\beta\mathcal{L}(\mathbf{w})} \\
&= \frac{1}{Z_N(\mathbf{X},\mathbf{y},\beta)} exp\left\{-\beta\left[\sum_{\mu=1}^{n} \ell(y_\mu, X_\mu\cdot\mathbf{w}) + \sum_{i=1}^{d} r(w_i)\right]\right\} \\
&= \frac{1}{Z_N(\mathbf{X},\mathbf{y},\beta)} \prod_{i=1}^{d} e^{-\beta r(w_i)} \prod_{\mu=1}^{n} e^{-\beta\ell(y_\mu, X_\mu\cdot\mathbf{w})}
\end{aligned}$$

Where for clarity:

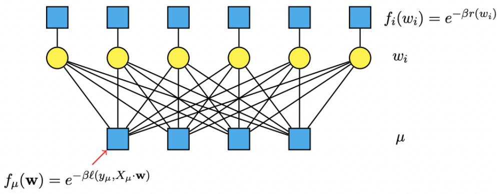


Figure 4.9: GLM factor graph

- N is the number of configurations of the possible values $\mathcal{L}$ can take
- d is again the dimension of the samples, indexed b i
- n is the number of samples, indexed by μ

And moreover:

- We recognize for w_i that $g_i(w_i) = e^{-\beta r(w_i)}$
- factors are indexed by μ and $f_a = f_\mu$. Where each of the factors connects with all w_i

We can then build the factor graph for the GLM as in Figure 4.9.

The result of Example 4.12, though interesting, proves that not all factor graphs are trees, and that the reasoning made so far is not exact for any kind of problem. After presenting the procedure for an exact tree, we will show a result that avoids this issue.

4.3 BP on trees and sparse graphs

For a tree, the BP algorithm needs only one iteration, and will exploit the equations summary, with initialization at the leaves:

$$\chi_{s_j}^{j \to a} = g_j(s_j) \tag{4.69}$$

Which follows by Equation 4.63 on leaves. In case of non tree like models, the independence of branches up to node conditioning generally fails to hold. Yet, it is possible to heuristically iterate until convergence the equations parametrized by a time index t as:

$$\chi_{s_j}^{j \to a}(t+1) = \frac{1}{Z^{j \to a}(t)} g_j(s_j) \prod_{b \in \partial j \setminus a} \psi_{s_j}^{b \to i}(t) \tag{4.70}$$

$$\psi_{s_i}^{a \to i}(t) = \frac{1}{Z^{a \to i}(t)} \sum_{\{s_j\}_{j \in \partial a \setminus i}} f_a\left(\{s_j\}_{j \in \partial a}\right) \prod_{j \in \partial a \setminus i} \chi_{s_j}^{j \to a}(t) \tag{4.71}$$

Where it is required to initialize $\chi_{s_j}^{j \to a}$ at some points which are not leaves, leading to *inexact* choices such as:

- the prior approach

$$\chi_{s_j}^{j \to a}(t=0) = \frac{g_j(s_j)}{\sum_s g_j(s)} \tag{4.72}$$

- a perturbed prior, yet to normalize

$$\chi_{s_j}^{j \to a}(t=0) = g_j(s_j) + \varepsilon_{s_j}^{j \to a} \tag{4.73}$$

- random, yet to normalize

$$\chi_{s_j}^{j \to a}(t=0) = \varepsilon_{s_j}^{j \to a} \tag{4.74}$$

- planted initialization

$$\chi_{s_j}^{j\to a}(t=0) = \delta_{s_j, s_j^*} \tag{4.75}$$

Pros and cons of these approaches will be discussed in the next Chapters.

Similarly, we will concentrate on cases in which the independence between incoming messages $\psi_{s_j}^{b\to j}$ for $b \in \partial j \setminus a$ and messages $\chi_{s_j}^{j\to a}$ for $j \in \partial a \setminus i$ holds approximately and leads to exact results to leading order in N. One such case is that of sparse graphs.

Getting back to the graph coloring problem of Section 1.1, we could construct and adjacency matrix for a graph with N nodes of the form:

$$\begin{cases} i < j \\ A_{ij} = 1 \quad w.p. \ \dfrac{c}{N} \\ A_{ij} = 0 \quad otherwise \end{cases} \tag{4.76}$$

Where clearly as $N \to \infty$ we have that the average number of neighbors $\dfrac{c}{N-1} \sim \dfrac{c}{N}$ the probability of having a single connection. We could then let the degree of any node be equal to the factor $c = \overline{d_i} \in O(1) \forall i$. Informally, *sparsity* refers to the fact that while the degree is kept constant at c, the size of the graph diverges. Again informally, we say that a graph will be *locally tree like* if for almost all nodes, where $N \to \infty$, the neighborhood at a diverging distance $d \to \infty$ is a tree.

Theorem 4.13 (Sparse random graphs are locally tree like)**.** *For a sparse random graph of size N, consider its neighborhood. As $N \to \infty$ it is tree like up to distance $d \to \infty$.*

Proof. For a sparse graph, we aim to check that the distance for a loop (i.e. not a tree) diverges. To do so, we consider a node i, then in d steps a loop has probability:

$$\begin{aligned} \mathbb{P}(loop\ i) &= 1 - \mathbb{P}(no\ return\ to\ i) \\ &= 1 - \Big(1 - \mathbb{P}(i\ neighbor \forall steps)\Big) \\ &\approx 1 - \left(1 - \frac{1}{N}\right)^{c^d} && \text{where } c^d \approx \overline{d_i}^d = \mathbb{E}[\#nodes\ in\ d\ steps] \end{aligned}$$

Letting $N \to \infty, c \in O(1)$ we have that:

- for d small, the probability is exponentially close to zero
- for d large, it is exponentially close to one

We further look for a distance d such that the probability becomes $\mathbb{P}(loop\ i) \in O(1)$, which marks the order of length of the shortest loop. Thus, we inspect the logarithm of the second term for better clarity[3] and we have that:

$$\begin{aligned} \log\left[\left(1 - \frac{1}{N}\right)^{c^d}\right] &= c^d \log\left(1 - \frac{1}{N}\right) \\ &\approx c^d \left(-\frac{1}{N} - \frac{1}{2N^2} + \cdots\right) \in O(1) && \text{Taylor} \\ \iff c^d &\approx O(N) \\ \iff d &\approx \frac{\log(N)}{\log(c)} = \log(N) && \text{by } \log(c) \in O(1) \end{aligned}$$

[3] Indeed, enforcing a constant w.r.t. size $O(1)$ means the same on the linear or logarithmic scale. In other terms, a constant has a constant logarithm and viceversa.

And for a distance lower than $\log(N)$ the neighborhood is with high probability a tree. Notice that as $N \to \infty \implies d \to \infty$ and we have a divergence property of the neighborhood ensuring that sparse graphs are locally tree like. $\square$

In later Chapters these concepts will be expanded further for greater understanding.

Chapter 5

Solving Graph Coloring: Belief Propagation

In this Chapter, we will focus on giving a fresh start to the graph coloring problem introduced in Chapter 1. We will just see the tip of the iceberg of using a graphical model for graph coloring, and reroute the reader to later Chapters for its conclusion.

Recalling the discussion of Section 1.1, given a graph $G(V, E)$ where $|V| = N$ and $|E| = M$. For each vertex $i \in V$, we aim to assign a color, represented as an integer $s_i \in \{1, \dots, q\}$. In Physics, this formalization is commonly referred to as **Potts spins**. In this setting, the Boltzmann measure takes the form:

$$\mathbb{P}\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z_G(\beta)} \prod_{(ij)\in E} e^{-\beta\delta_{s_i s_j}} \tag{5.1}$$

Just to recall key concepts, below are some noteworthy quick observations:

- interactions are determined by colors, encoded in s_i, s_j variables for an edge $(ij) \in E$
- the β factor governs the strength in probability of such interaction and is *artificially introduced*

Going further into their effects in the probabilistic context, we can see that for two connected nodes $i, j \in V$:

$$\left\{s_i = s_j \iff \delta_{s_i s_j} = 1\right\} \implies \mathbb{P}\left(\{s_i\}_{i=1}^N\right) \overset{\beta\to\infty}{\to} 0 \tag{5.2}$$

$$\beta = 0 \implies \mathbb{P}\left(\{s_i\}_{i=1}^N\right) \sim \mathcal{U}(q^N \textit{ colorings}) \tag{5.3}$$

$$\left\{s_i = s_j \iff \delta_{s_i s_j} = 1\right\}\forall (ij) \in E \implies \mathbb{P}\left(\{s_i\}_{i=1}^N\right) \overset{\beta\to-\infty}{\to} 1 \tag{5.4}$$

These three conditions read as follows:

- In case of equal colors and $\beta \gg 1$, the probability drowns to zero. It is very unlikely that two connected spins will be equally colored. This type of interaction is **antiferromagnetic**
- In any case, $\beta = 0$ causes the probability measure to become uniform across all the possible configurations
- A very low inverse temperature $\beta \ll 1$ induces alignment in all[1] the connected spins with a probability that tends to unity. This type of interaction is **ferromagnetic**

[1]note that in Equation 5.4 there is a *for all edges* term before the implication. Namely, all clusters of connected nodes must agree on a color

In this Chapter, we will focus on the latter case, which will be analyzed without much algorithmic flavour. A naturally arising question could be:

Why is the Boltzmann measure introduced in the first place?

Hopefully, the next arguments will provide evidence for why this is useful.

5.1 Energy, Entropy and Counting

Given a graph instance G and a cost/energy counting the average number of monochromatic edges of the form:

$$\mathfrak{e} = \frac{1}{N} \sum_{(ij)\in E} \delta_{s_i s_j} \tag{5.5}$$

Consider a function $\mathscr{N}(\cdot) : \mathbb{R}^+ \to \mathbb{N}^+$ which given an energy value returns the number of configurations attaining it. Since $\mathscr{N}(\mathfrak{e})$ is tipycally exponentially large, we will work with a defined value.

Definition 5.1 (Entropy $s(\cdot)$)**.** For a given energy level $\mathfrak{e}$ a system has entropy:

$$\mathscr{N}(\mathfrak{e}) = e^{Ns(\mathfrak{e})} \iff s(\mathfrak{e}) = \frac{1}{N} \log \Big[\mathscr{N}(\mathfrak{e})\Big] \tag{5.6}$$

The importance of such quantity lies in the fact that for any problem we wish to examine feasible solutions at a given energy level. In other words, for graph coloring, we might be interested in the existance of satisfactory configurations at a given energy level, i.e. is it feasible to have a valid graph coloring with necessarily zero energy? If not, then graph coloring for the considered instance is not solvable. Again, with access to $s(\cdot)$, one could examine almost exact results with low energy levels as well.

Observation 5.2 (Entropy and free entropy)**.** *The previously introduced free entropy deonted as Φ is not equal to the entropy s. Just to recall Definition 2.9 tells us that:*

$$\Phi(\beta) = \log[Z_N(\beta)] \tag{5.7}$$

Nevertheless, it is used to make easier calculations to determine s.

Using the just reminded fact, we observe that the partition function can be expressed as:

$$\begin{aligned}
e^{N\Phi_N(\beta)} &= e^{N\frac{1}{N}\log[Z_G(\beta)]} && \text{Definition 2.9} && (5.8)\\
&= Z_G(\beta) && && (5.9)\\
&= \sum_{\{s_i\}_{i=1}^N} e^{-\beta \sum_{(ij)\in E} \delta_{s_i s_j}} && \text{recall Example 4.5} && (5.10)\\
&= \sum_{\mathfrak{e}\ values} \sum_{\{s_i\}_{i=1}^N | \mathfrak{e}\ energy} e^{-\beta N \mathfrak{e}} && \text{splitting the sum} && (5.11)\\
&= \sum_{\mathfrak{e}\ values} \mathscr{N}(\mathfrak{e}) e^{-\beta N \mathfrak{e}} && \text{fixing energy (the value is the same)} && (5.12)\\
&= \sum_{\mathfrak{e}\ values} e^{Ns(\mathfrak{e}) - \beta N \mathfrak{e}} && \text{Definition 5.1} && (5.13)\\
Z_G(\beta) &= \int d\mathfrak{e}\ e^{Ns(\mathfrak{e}) - \beta N \mathfrak{e}} && \text{as } N \to \infty && (5.14)
\end{aligned}$$

Where in particular we substitute with an integral since we are interested in the large size behavior and only look for leading order trends in the quantities $\Phi, \mathfrak{e}, s$ of interest.

Equation 5.14 can be solved using the saddle point method introduced in Chapter 2, which reads:

$$\begin{cases} \dfrac{\partial[Ns(\mathfrak{e}) - N\beta\mathfrak{e}]}{\partial\mathfrak{e}} = 0 \\ e^{N\Phi_N(\beta)} = e^{Ns(\mathfrak{e})-\beta N\mathfrak{e}} \end{cases} \iff \begin{cases} \left.\dfrac{\partial s(\mathfrak{e})}{\partial\mathfrak{e}}\right|_{\mathfrak{e}=\mathfrak{e}^*} = \beta \\ \Phi_N(\beta) = s(\mathfrak{e}^*) - \beta\mathfrak{e}^* \end{cases} \tag{5.15}$$

While this last result might seem useless, connecting some dots from previous arguments, it will delineate a clear method to derive the entropy of a system and thus the number of configurations at a given energy.

Now, we recall that for a Boltzmann measure, we have that :

$$\frac{d\Phi_N(\beta)}{d\beta} = \frac{1}{N}\frac{d}{d\beta}\log[Z_G(\beta)] \tag{5.16}$$

$$= \frac{1}{N}\frac{d}{d\beta}\log\left[\sum_{\{s_i\}_{i=1}^N} e^{-\beta\sum_{(ij)\in E}\delta_{s_is_j}}\right] \tag{5.17}$$

$$= \frac{1}{N}\frac{1}{\underbrace{\sum_{\{s_i\}_{i=1}^N} e^{-\beta\sum_{(ij)\in E}\delta_{s_is_j}}}_{=Z_G(\beta)}}\sum_{\{s_i\}_{i=1}^N}(-1)\sum_{(ij)\in E}\delta_{s_is_j}e^{-\beta\sum_{(ij)\in E}\delta_{s_is_j}} \tag{5.18}$$

$$= (-1)\sum_{\{s_i\}_{i=1}^N}\underbrace{\frac{1}{N}\sum_{(ij)\in E}\delta_{s_is_j}}_{=\mathfrak{e}}\underbrace{\frac{e^{-\beta\sum_{(ij)\in E}\delta_{s_is_j}}}{\sum_{\{s_i\}_{i=1}^N} e^{-\beta\sum_{(ij)\in E}\delta_{s_is_j}}}}_{=\mathbb{P}(\{s_i\}_{i=1}^N)} \quad \text{reordering \& bringining inside the su} \tag{5.19}$$

$$= -\langle\mathfrak{e}\rangle_{Boltzmann} \tag{5.20}$$

At the large limit, the average energy will also be the final energy as discussed in Theorem 3.4. Thus, if we compute $\Phi_N(\beta)$ we can explicitly derive $\mathfrak{e}$ and solve for $s(\mathfrak{e})$ in Equation 5.15:

$$s(\mathfrak{e}^*) = \Phi_N(\beta) + \beta\mathfrak{e}^* \tag{5.21}$$

Doing these calculations exactly is in most cases unfeasible. In Chapter 4 we saw how this is independent of Belief Propagation, where we were computing the approximate Bethe free entropy and messages $\chi^{i\to j}$ until convergence (a fixed point) satisfying (*Bethe free Entropy*, exercise 4.2 [KZ21b]):

$$\frac{d\Phi_{bethe}(\beta)}{d\beta} = \underbrace{\frac{\partial\Phi_{bethe}}{\partial\chi}}_{=0}\frac{\partial\chi}{\partial\beta} + \frac{\partial\Phi_{bethe}}{\partial\beta} \quad \text{at a fixed point} \tag{5.22}$$

$$= \frac{\partial\Phi_{bethe}}{\partial\beta} \tag{5.23}$$

Making it possible to derive the energy directly from the BP equations through Φ_{bethe} to solve for $s(\mathfrak{e}^*)$ in Equation 5.21

5.2 Adapting BP equations to graph coloring

For a graphical model as that of a graph coloring problem, we recognize that the node functions and edge functions take the form:

$$\begin{cases} g_i(s_i) = 1 & \forall i \\ f_{(ij)}(s_i, s_j) = e^{-\beta\delta_{s_i s_j}} & \forall a = (ij) \in E \end{cases} \tag{5.24}$$

Where we stress that every edge is a factor, i.e. $(ij) = a$. Plugging these notions into the Equations of Chapter 4 we get:

$$\chi_{s_j}^{j\to(ij)} = \frac{1}{Z^{j\to(ij)}} \prod_{(kj)\in\partial j\setminus(ij)} \psi_{s_j}^{(kj)\to j} \tag{5.25}$$

$$\psi_{s_i}^{(ij)\to i} = \frac{1}{Z^{(ij)\to i}} \sum_{\{s_j\}_{j\in\partial(ij)\setminus i}} f_{(ij)}(\{s_j\}_{j\in\partial(ij)}) \prod_{j\in\partial(ij)\setminus i} \chi_{s_j}^{j\to(ij)} \tag{5.26}$$

$$= \frac{1}{Z^{(ij)\to i}} \sum_{s_j} f_{(ij)}(s_i, s_j)\chi_{s_j}^{j\to(ij)} \qquad \text{since } \{s_j\}_{j\in\partial(ij)\setminus i} = s_j \tag{5.27}$$

$$= \frac{1}{Z^{(ij)\to i}} \sum_{s_j} e^{-\beta\delta_{s_i s_j}} \chi_{s_j}^{j\to(ij)} \qquad \text{Eqn. 5.24} \tag{5.28}$$

$$= \frac{1}{Z^{(ij)\to i}} \Bigg[\underbrace{e^{-\beta}\chi_{s_i}^{i\to j}}_{case\ s_i=s_j} + \sum_{s_j\neq s_i} e^{-\beta\cdot 0}\chi_{s_j}^{i\to j} \Bigg] \qquad \text{splitting} \tag{5.29}$$

$$= \frac{1}{Z^{(ij)\to i}} \Bigg[\underbrace{e^{-\beta}\chi_{s_i}^{i\to j}}_{case\ s_i=s_j} + \underbrace{1 - \chi_{s_i}^{j\to(ij)}}_{1-case\ s_i=s_j} \Bigg] \tag{5.30}$$

$$= \frac{1}{Z^{(ij)\to i}} \Bigg[1 - \left(1 - e^{-\beta}\right)\chi_{s_i}^{j\to(ij)} \Bigg] \qquad \text{reordering} \tag{5.31}$$

Now, substituting back the value of ψ into χ we get that:

$$\chi_{s_j}^{j\to(ij)} = \frac{1}{Z^{j\to(ij)}} \prod_{(kj)\in\partial j\setminus(ij)} \frac{1}{Z^{(kj)\to i}} \Bigg[1 - \left(1 - e^{-\beta}\right)\chi_{s_j}^{k\to(kj)} \Bigg] \qquad \text{Eqns 5.25, 5.31} \tag{5.32}$$

$$= \frac{1}{Z^{j\to(ij)}} \frac{\prod_{(kj)\in\partial j\setminus(ij)} 1 - \left(1 - e^{-\beta}\right)\chi_{s_j}^{k\to(kj)}}{\prod_{(kj)\in\partial j\setminus(ij)} Z^{(kj)\to i}} \qquad \text{reordering} \tag{5.33}$$

$$= \underbrace{\frac{1}{Z^{j\to(ij)} \prod_{(kj)\in\partial j\setminus(ij)} Z^{(kj)\to i}}}_{=Z^{j\to i}\ new\ normalization} \prod_{(kj)\in\partial j\setminus(ij)} 1 - \left(1 - e^{-\beta}\right)\chi_{s_j}^{k\to(kj)} \tag{5.34}$$

$$\chi_{s_j}^{j\to i} = \frac{1}{Z^{j\to i}} \underbrace{\prod_{k\in\partial j\setminus i} 1 - \left(1 - e^{-\beta}\right)\chi_{s_j}^{k\to j}}_{(kj)\in\partial j\setminus(ij)\equiv k\in\partial j\setminus i} \tag{5.35}$$

Where in the last passage we exploit the fact that the factor graph has a neighbor structure such that variable nodes are connected to each other only through factors. Due to this, we can resort to using a **graph** notation instead of a **factor graph** notation for neighbors and write:

$$(kj) \in \partial j \setminus (ij) \equiv k \in \partial j \setminus i$$

Indeed, they are the same as messages between nodes with the newly defined χ are necessarily one to one with messages between nodes in the factor graph, which connects variable nodes only through the edge (ij).
For the same reasons, since all factors have two neighbors, we can ignore $\psi_{s_i}^{(ij)\to i}$.

To further understand, following the interpretations proposed in Chapter 4, we recall that $\chi_{s_j}^{j\to i}$ is the probability that j takes value s_j if the connection (ij) was removed. Then:

- $$1-(1-e^{-\beta})\chi_{s_j}^{k\to j} = 1-\chi_{s_j}^{k\to j}+e^{-\beta}\chi_{s_j}^{k\to j} = \sum_{s_k\neq s_j} \chi_{s_k}^{k\to i}+e^{-\beta}\chi_{s_j}^{k\to i} = \mathbb{P}(k \textit{ allows } j \textit{ taking } s_j) \tag{5.36}$$
  Noticing that $e^{-\beta\delta_{s_i s_j}} = e^{-\beta\cdot 0} = e^0 = 1$ in the sum making it disappear, while out of the sum we get $e^{\delta_{s_i s_j}} = e^{-\beta}$. Then, inside the product we have the probability of a neighbor k allowing node j to take the value s_j
- the product $\prod_{k\in\partial j\setminus i}$ spreads across all neighbors of j but the connection to i, thus returning the probability that all neighbors allow j to take value s_j
- the product makes sense since upon conditioning on s_j there neighbors are assumed to be independent

In an algorithmic setting, Belief Propagation is used to extract solutions to computational problems. In this case, the messages are to be interpreted as probabilities. To return a solution (i.e. a graph coloring), one needs to **decimate** BP, resorting to an iteration of the algorithm and a choice of colors depending on the final marginal probabilities. We will get back to this procedure in later Chapters.
For the time being, we will focus on using BP as a tool for analysis. The missing piece is computing Φ_{bethe}, which will be the objective of the next section.

5.3 Free energy for graph coloring

It is possible to show with similar arguments that for graph coloring it holds that[2]:

$$N\Phi_{N,bethe}(\beta) = \Phi_{bethe}(\beta) = \sum_i \log[Z^i] - \sum_{(ij)\in E} \log[Z^{(ij)}] \tag{5.37}$$

$$Z^i = \sum_s \prod_{k\in\partial i} \left[1 - \left(1-e^{-\beta}\right)\chi_s^{k\to i}\right] \tag{5.38}$$

$$Z^{(ij)} = \sum_{s_i,s_j} e^{-\beta\delta_{s_i s_j}} \chi_{s_i}^{i\to j}\chi_{s_j}^{j\to i} = 1-(1-e^{-\beta})\sum_s \chi_s^{i\to j}\chi_s^{j\to i} \tag{5.39}$$

We will evaluate this expression at a fixed point of the BP equations as argued before, and stress that such a quantity might be seen in terms of the messages χ or other parametrizations.

Observation 5.3 (Exactness of Φ_{bethe}). *While on a tree we showed in Section 4.3 that $\Phi_{bethe} = \Phi$, this is not always the case. In the next Chapters, the main differences in general cases will be dealt with in a clear manner.*

Observation 5.4 (Disorder and BP). *All of the above discussion is based on a graph instance $G(V,E)$, and no average over the probabilistic disorder was taken yet. In the replica method this was the first step.*

[2] *Bethe free Entropy*, exercise 5.1 [KZ21b]

5.3.1 Paramagnetic Interaction

Searching for a closed form of fixed points of the BP equations it is rather easy to check that there is always one.

Theorem 5.5 (Easy BP fixed point for graph coloring). *For graph coloring, it holds that:*

$$\chi_{s_j}^{j\to i} = \frac{1}{q} = \frac{1}{q}\frac{\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i-1}}{\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i-1}} \quad \forall (ij) \in E, \forall s_j \in \{1,\dots,q\} \tag{5.40}$$

Where d_i is the degree of node i, and q is the number of colors, is a fixed point for Equation 5.35

Proof. We check that the claim is correct by substituting into Equation 5.35:

$$\chi_{s_j}^{j\to i} = \frac{1}{Z^{j\to i}} \prod_{k\in\partial j\setminus i} 1-\left(1-e^{-\beta}\right)\chi_{s_j}^{k\to j} \tag{5.41}$$

$$\frac{1}{q} = \frac{1}{Z^{j\to i}}\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i-1} \qquad |k\in\partial j\setminus i| = d_i-1 \text{ all neighbors of } i \text{ but itself} \tag{5.42}$$

$$= \underbrace{\frac{1}{q\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i-1}}}_{=Z}\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i-1} \qquad Z \text{ is uniform probability normalization} \tag{5.43}$$

$$= \frac{1}{q} \tag{5.44}$$

□

Using the result of Theorem 5.5, and letting $c = \sum_i \frac{d_i}{N} = \frac{2M}{N}$ be the average degree of the graph we recover for the Bethe free entropy of Equation 5.37:

$$N\Phi_{N,bethe}(\beta) = \sum_i \log[Z^i] + \sum_{(ij)\in E} \log[Z^{(ij)}] \tag{5.45}$$

$$= \sum_i \log\left\{\underbrace{\sum_s}_{q\ values}\ \underbrace{\prod_{k\in\partial i}}_{d_i\ times}\left[1-(1-e^{-\beta})\frac{1}{q}\right]\right\} \tag{5.46}$$

$$- \sum_{(ij)\in E}\log\left\{1-(1-e^{-\beta})\underbrace{\sum_s}_{q\ values}\frac{1}{q}\frac{1}{q}\right\}$$

$$= \sum_i \log\left\{q\left[1-(1-e^{-\beta})\frac{1}{q}\right]^{d_i}\right\} \tag{5.47}$$

$$- \sum_{(ij)\in E}\log\left\{1-(1-e^{-\beta})\frac{1}{q}\right\}$$

Now, we move N to the LHS and use some logarithm properties to simplify the equation, along with the previouslt defined $c = \sum_i \frac{d_i}{N} = \frac{2M}{N}$:

$$\Phi_{N,bethe}(\beta) = \frac{1}{N}\sum_i \log(q) + d_i \log[1-(1-e^{-\beta})q^{-1}] - \frac{1}{N}\sum_{(ij)\in E} \log\left[1-(1-e^{-\beta})\frac{1}{q}\right] \tag{5.48}$$

$$= \frac{1}{N} N\log(q) + \log[1-(1-e^{-\beta})q^{-1}]\underbrace{\frac{1}{N}\sum_i d_i}_{=c} - \underbrace{\frac{1}{N}M}_{=\frac{c}{2}}\log\left[1-(1-e^{-\beta})\frac{1}{q}\right] \tag{5.49}$$

$$= \log(q) + c\log[1-(1-e^{-\beta})q^{-1}] - \frac{c}{2}\log[1-(1-e^{-\beta})q^{-1}] \tag{5.50}$$

$$= \log(q) + \frac{c}{2}\log\left[1-(1-e^{-\beta})\frac{1}{q}\right] \tag{5.51}$$

With this value in hand, we can go on inspecting the energy and the entropy using Equations 5.15, 5.20, 5.23:

$$-\frac{d\Phi_{bethe}}{d\beta} = -\frac{\partial \Phi_{bethe}}{\partial\beta} \qquad \text{Eqn. 5.23} \tag{5.52}$$

$$= \langle \mathfrak{e} \rangle_{Boltzmann} \qquad \text{Eqn. 5.20} \tag{5.53}$$

$$= \mathfrak{e}^* \qquad \text{Thm. 3.4} \tag{5.54}$$

$$= \frac{c}{2}\frac{1}{1-(1-e^{-\beta})\frac{1}{q}}(-1)(-1)e^{-\beta}\frac{1}{q} \qquad \text{expand } \partial \text{ derivative} \tag{5.55}$$

Which, working further the final fraction becomes:

$$\implies \mathfrak{e}^* = \frac{c}{2}\frac{e^{-\beta}\frac{1}{q}}{1-\frac{1}{q}+e^{-\beta}\frac{1}{q}} \tag{5.56}$$

$$= \frac{c}{2}\frac{\frac{1}{q}}{\frac{1}{q}}\frac{e^{-\beta}\frac{1}{q}}{1-\frac{1}{q}+e^{-\beta}\frac{1}{q}} \tag{5.57}$$

$$= \frac{c}{2}\underbrace{\frac{e^{-\beta}}{(q-1)+e^{-\beta}}}_{=\mathbb{P}(edge\ monochromatic)} \tag{5.58}$$

Where by $\mathbb{P}(edge\ monochromatic)$ we mean the probability that two nodes have the same color, i.e. the inverse of the probability that they have any pair of non equal colors.

Similarly, inspecting the entropy:

$$s(\mathfrak{e}^*) = \Phi_{bethe}(\beta) + \beta\mathfrak{e}^* \qquad \text{Eqn. 5.15} \tag{5.59}$$

$$= \log(q) + \frac{c}{2}\log\left[1-(1-e^{-\beta})\frac{1}{q}\right] + \beta\frac{c}{2}\frac{e^{-\beta}}{(q-1)+e^{-\beta}} \tag{5.60}$$

We recall the parameters into play are:

- β the inverse temperature
- q the number of available colors
- $c = \dfrac{2M}{N}$ the average degree of the graph
- the fixed point is $\chi^{i\to j}_{s_j} = \dfrac{1}{q}$

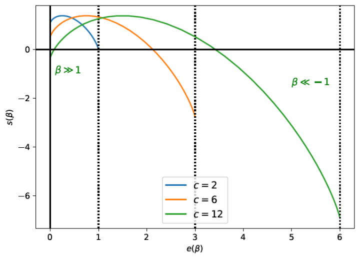


Figure 5.1: Entropy vs energy at different c values

And choose to plot the entropy on the y-axis against the energy parametrized by β, for a fixed number of colors $q = 4$. Different average degrees (i.e. values of c) are plotted in different colors. The result is that of Figure 5.1.

The plot can be explained observing the following facts:

- the energy $\mathfrak{e} \in [0, \frac{c}{2}] = [0, \frac{M}{N}] = [all\ different,\ allsame]$
- the entropy is maximum at the tipycal case for the energy, which is $\mathfrak{e} = \frac{c}{2q} = \dfrac{M}{Nq}$ and $s(\mathfrak{e}) = \dfrac{1}{N}\log(\mathcal{N}(\mathfrak{e})) = \dfrac{1}{N}\log(q^N) = \log(q)$ by a simple maximization argument. Intuitively, a uniform distribution is the *most chaotic.*
- the slope of the curve $s(\mathfrak{e})$ is $\dfrac{\partial s(\mathfrak{e})}{\partial \mathfrak{e}} = \beta$ by the Legendre transform

Observation 5.6 (Entropy and its possible values)**.** *Observing the plot of Figure 5.1 and the entropy from Definition 5.1 the two are not in accordance. Indeed, $s(\mathfrak{e})$ is defined only on the positive section of the plane, being the logarithm of positive numbers, i.e. $\mathcal{N}(\mathfrak{e})$. Thus, the section of the curve where $s(\mathfrak{e}) < 0$ make no sense, and either does not exist, or highlights a mistake in the just made reasoning.*

Observation 5.7 (Achievable but non sense energy levels)**.** *Consider a graph, set all the colors to the same index, then $\mathfrak{e} = \dfrac{c}{2} = \dfrac{M}{N}$. Yet, for $c = 12$, as Figure 5.1 suggests, this is not feasible, as it has negative entropy (see Observation 5.6).*
*Additionally notice that the assumption on the fixed point roughly means that all of the colors will be covered by the same amount of nodes. Clearly, this does not coincide with an assignment of only one color to all nodes, achieving $\mathfrak{e} = \dfrac{c}{2}$. The **paramagnetic** i.e. $\chi_{s_j}^{i\to j}$ fixed point assumption is not valid and one must inspect a **ferromagnetic** setting where there is a **probabilistic preference** on a color.*

5.3.2 Ferromagnetic Interaction

Assumption 5.8 (Only random graphs)**.** To simplify the analysis, only focus on random sparse graphs.

Enforcing a preference on a specific color, which for simplicity is indexed by 1:

$$\chi_1^{i\to j} = a \quad \chi_s^{i\to j} = b = \frac{1-a}{q-1} \quad \forall s \neq 1 \tag{5.61}$$

And adjusting the BP equations we get that the new form of Equation 5.41 is:

$$\begin{cases} a = \dfrac{1}{Z^{j\to i}} A^{d_j - 1} := \dfrac{1}{Z^{i\to j}} \left[1 - (1 - e^{-\beta})a\right]^{d_j - 1} \\ b = \dfrac{1}{Z^{j\to i}} B^{d_j - 1} := \dfrac{1}{Z^{j\to i}} \left[1 - (1 - e^{-\beta})b\right]^{d_j - 1} \\ Z^{j\to i} = (q-1)B^{d_j - 1} + A^{d_j - 1} \end{cases} \tag{5.62}$$

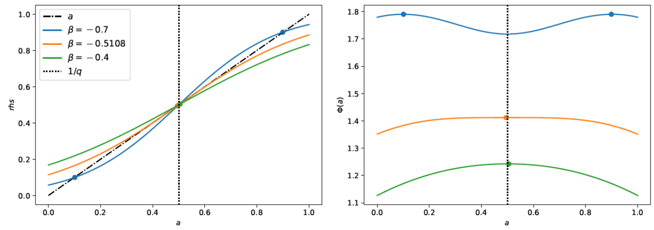


Figure 5.2: Fixed points and Φ in the imbalanced fixed point Equation

Readjusting the fixed point requirements as before we can work out the expressions:

$$\begin{cases} a = \dfrac{A^{d_j-1}}{(q-1)B^{d_j-1} + A^{d_j-1}} \\ b = \dfrac{B^{d_j-1}}{(q-1)B^{d_j-1} + A^{d_j-1}} \end{cases} \tag{5.63}$$

Which are satisfied $\forall (ij) \in E \iff d_j \equiv d \ \forall j$. This condition coincides with the definition of **random regular graphs**, an instance of the family of graphs which have always d connections for each node.

Observation 5.9 (On the fixed point solution)**.** *It is worth noticing that this is only an option. Another possibility is specifying further the granularity of a as a function of its (ij) connections for each node. This is a simpler treatment.*

Expressing the a quantity as a function of $d_j \equiv d \ \forall j, \beta, q$ we get the ferromagnetic fixed point condition:

$$a = RHS(a; \beta, d, q) := \frac{\left[1 - (1 - e^{-\beta})a\right]^{d_j - 1}}{(q-1)B^{d_j-1} + A^{d_j-1}} \tag{5.64}$$

$$= \frac{[1 - (1 - e^{-\beta})a]^{d-1}}{(q-1)[1 - (1 - e^{-\beta})b]^{d-1} + [1 - (1 - e^{-\beta})a]^{d-1}} \tag{5.65}$$

$$= \frac{[1 - (1 - e^{-\beta})a]^{d-1}}{(q-1)[1 - (1 - e^{-\beta})\frac{1-a}{q-1}]^{d-1} + [1 - (1 - e^{-\beta})a]^{d-1}} \tag{5.66}$$

Which plugged into the Bethe free entropy results in:

$$\Phi_{N,bethe}(\beta) = \frac{1}{N}\sum_i \log[Z^i] - \frac{1}{N}\sum_{(ij)\in E} \log[Z^{(ij)}] \tag{5.67}$$

$$= \log\left\{(q-1)[1 - (1 - e^{-\beta})\frac{1-a}{q-1}]^d + [1 - (1 - e^{-\beta})a]^d\right\} \tag{5.68}$$

$$- \frac{d}{2}\log\left\{1 - (1 - e^{-\beta})\left[\frac{(1-a)^2}{q-1} + a^2\right]\right\}$$

We choose to plot the behavior of the fixed point equation $a = RHS(a, \beta, d, q)$ and the $a, \Phi_{N,bethe}(a)$ relationship. These two are the left and right side plot of Figure 5.2. Both are proposed for different values of β, which correspond to different colors. The former highlights the fact that as $\beta \to -1$ the stable fixed points get far away, while for a bigger $\beta = 0/4$ such fixed point is paramagnetic (i.e. $\frac{1}{q}$), previously discussed. The latter is even more interesting: we notice that as $\beta \to -1$ clearly the paramagnetic fixed point becomes unstable (a minima) and two new (maximal) fixed points appear.

It is interesting to explore at what precise β value such a behavior starts showing up. The upward trend of the slope of the $RHS(a, \beta, d, q)$ curve suggests inspecting when

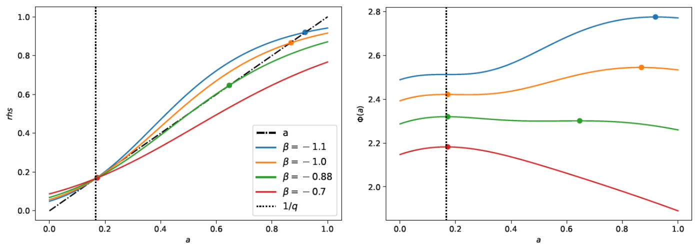


Figure 5.3: The β_{stab} switch

such slope becomes unity at the paramagnetic fixed point to see the boundary point from straight line to curve. We call this value β_{stab} and formalize it as:

$$\frac{\partial RHS(a)}{\partial a}\bigg|_{a=\frac{1}{q}} = 1 = \frac{(d-1)(1-e^{-\beta})}{e^{-\beta}+q-1} \qquad \beta_{stab} \text{ condition} \tag{5.69}$$

$$\implies \beta_{stab} = -\log\left[1+\frac{q}{d-2}\right] \tag{5.70}$$

Equation 5.70 is useful for examinining other cases in which $q = 2$ but $d \geq 3$. All of those present the same transition from one to two fixed points[3], equivalent to the second order phase transition of the Curie Weiss model of Section 2.2.

Observation 5.10 (Curie Weiss model vs Graph coloring)**.** *We can interpret the Curie Weiss model as the limiting form of the graph coloring problem for $d \to \infty$. Recall that the Curie Weiss model explains the solution on a fully connected graph, while in the case of d regular graphs the number of connections is d for each node.*

To visualize the switch that happens at β_{stab} refer to the example for $q = 2, d = 5$ of Figure 5.3, where the horizontal straight line is $\frac{1}{q}$. Remember that in section 2.2 the probability a was a magnetization: the plots are equivalent!

Coming back to Observation 5.10, more things can be said. Sticking to $q = 2, d \geq 3$ we have that in another interpretation the colors can be encoded differently as $q = \pm 1$ and $\delta_{s_i s_j} \in \{0, 1\}$, with a factor 2 floating around. This is exactly the **Bethe approximation** of the **Ising model** on sparse graphs with d neighbors on a lattice of dimension D where the following trivial identity holds:

$$d = 2D \tag{5.71}$$

Describing the properties of such models is usually very difficult. The Bethe approximation is a *quick* approach to propose a solution assuming that the lattice is tree like[4] d regular sparse graphs are locally tree like from the discussion carried out in Section 4.3, and we can inspect the degree of correctness of the approximation[5].

- for $d = 2 \implies \beta_{stab} = -\infty$ which makes sense as in a one dimensional Ising model there is no phase transition. Observe also that in this case the approximation is **exact** as a one dimensional grid is a tree
- for $d = 4 \implies \beta_{stab} = -0.693$ which is rather close to the exact solution by Onsager [Ons44] $\beta_{Onsager} = -\log[1+\sqrt{2}] = -0.881$

[3]From no preference on colors to color preference, from uniform probability to spike probability & so on...

[4]To be tree like, it must have no loops, this is a very strong adjustment! A grid has many loops.

[5]The sign of the critical temperature is opposite to the usual literature in this case since the β temperature was oriented as an antiferromagnet. Also the Hamiltonian misses the usual 2 factor in the Ising model. All of these formalizations do not influence the equivalence between the two approaches.

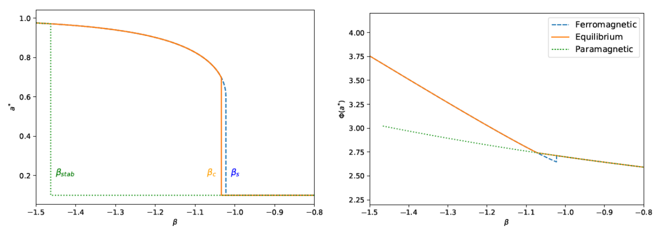


Figure 5.4: Phase transition, $q = 10, d = 5$

- for $d = 6 \implies \beta_{stab} = -0.4433$ and while there are no exact solutions, numerical simulation suggests that $\beta_{d=6} \approx -0.4055$, which is very close.
- It can be proved (*Fully connected limit from BP*, exercise 5.2 [KZ21b]) that as $d \to \infty$ the Bethe approximation becomes exact, up to a rescaling in the interaction strength[6].

5.3.3 More colors: the Potts ferromagnetic interaction

Consider the cases in which $q \geq 3$. In Figure 5.4, we show an example of the $RHS(a, \beta, q, d) = f(a)$ plot and $\Phi_{N,bethe}(a)$ plot as before for $q = 6, d = 5$. The situation is different. The $\frac{1}{q}$ fixed point is present, but there is no symmetry of the appearing fixed points when $\beta \to -1$. We recognize regions of β:

- for $\beta > \beta_{stab}$ inverse temperatures, the $\frac{1}{q}$ fixed point is a stable unique maximizer
- up to a certain point, as $\beta \to -1, \beta < \beta_{stab}$ another local maxima of the free entropy comes into play, but is globally lower than the level attained at the $\frac{1}{q}$ option. This region is the interval $[\beta_c, \beta_{stab}]$ where β_c denotes a the critical temperature, which was not present before
- at another critical value, denoted as β_s, the **spinodal inverse temperature**, a **discontinuous second order phase transition** induces the change of maximizer from $\frac{1}{q}$ to an imbalanced setting.
- Remember that with more than one fixed point, the highest in entropy is the dominant one when applying the saddle point method.

Observation 5.11 (More colors enhance the difference)**.** *Figure 5.4 is an example for $q = 10, d = 5$ of the magnetization level a^* (the fixed point) as β changes. The legend is shared across the two plots, and the change is clearly discontinuous.*

We can now solve the problem outlined in Observation 5.6, justifying a change of the fixed point from paramagnetic to ferromagnetic, attaining positive entropy at larger energy values. The newly adjusted plot is proposed in Figure 5.5. Above a certain energy level, the balanced color fixed point is no longer valid and a preference must be introduced, leading to the orange line.

At this point, a noteworthy question might be:

How do we ensure there is no other solution for $\beta < 0$?

Fortunately, two important theoretical results guarantee the uniqueness claim in the limit.

[6] The previously mentioned 2 factor

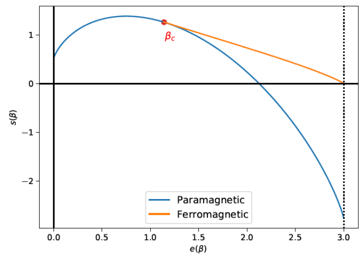


Figure 5.5: Adjusted Entropy vs energy plot

Theorem 5.12 (Exactness of BP on random graphs)**.** *Let $\beta < 0$, then the results obtained with BP on random graphs are exact. Namely:*

$$\forall \epsilon > 0 \quad \mathbb{P}\bigg[|\Phi(\beta) - \Phi_{bethe}(\beta)| < \epsilon\bigg] \overset{N\to\infty}{\to} 1 \quad w.h.p \tag{5.72}$$

In the cases:

- *$q = 2$, sparse random graphs (not necessarily regular) [DM09]*
- *$q = 3$ d-regular graphs [Dem+12]*

What we are missing is a (more convolved) treatment of the $\beta > 0$ case, which is developed in the next section.

5.4 Anti-ferromagnetic interaction

So far, with $\beta < 0$ we enforced attraction between nodes, and argued that nodes have a tendency to align. Now we let $\beta > 0$, where $\beta \to \infty$ is the original coloring problem. In particular, the focus is on finding a condition for sparse graphs that ensures existance of colorings with high probability, and the magnitude of such colorings, the entropy at zero energy[7] $s(\mathfrak{e} = 0)$.

Recall the general form of the probability distribution we are considering:

$$\mathbb{P}\Big(\{s_i\}_{i=1}^N\Big) = \frac{1}{Z_G}\prod_{i=1}^{N} g_i(s_i)\prod_{a=1}^{M} f_a\Big(\{s_i\}_{i\in\partial a}\Big) \tag{5.73}$$

In this setting the free entropy is explicitly dependent on the graph considered and we expect that it will self average:

$$\Phi_N = \frac{1}{N}\log[Z_G] \quad | \quad \mathbb{P}\Big(|\Phi_N - \mathbb{E}[\Phi_N]| > \epsilon\Big) \overset{N\to\infty}{\to} 0 \forall \epsilon > 0 \tag{5.74}$$

In the limit, the expectation and the actual value of the free entropy are assumed to have the same value. For the purpose of further understanding the differences into play, we introduce two useful probabilistic quantities:

Definition 5.13 (Quenched and Annealed free entropy)**.** Define two types of evaluating the expectation, either on the disorder of possible graphs as in the replica method, or on the given graph as in BP. We hvae that:

$$\Phi_{quench} := \mathbb{E}_G\left[\frac{1}{N}\log[Z_G]\right] \tag{5.75}$$

$$\Phi_{anneal} := \frac{1}{N}\log\Big[\mathbb{E}_G[Z_G]\Big] \tag{5.76}$$

[7] In other words, the number of valid ($\mathfrak{e} = 0 \implies$ *no errors*) colorings if they exist at all.

Proposition 5.14 (Properties of annealed and quenched free entropy)**.** *It holds that:*

- $\Phi_{anneal} \geq \Phi_{quench}$ *by applying Jensen's Inequality*
- $\exists G \mid \Phi_{anneal} > \Phi_{quench}$ *as* $N \to \infty$ *so the two do not necessarily coincide in the limit. Indeed, the exponential size order of magnitude of the partition function breaks the concentration that holds for* $\frac{1}{N}\log[Z]$*, and the annealed free entropy is dominated by rare instances of the graph, and not by the average!*

Example 5.15 (Easy rare dominance case)**.** Consider the simple artificial partition function:

$$Z_G = \begin{cases} e^N & w.p.\ 1 - e^{-N} \\ e^3 N & w.p.\ e^{-N} \end{cases} \tag{5.77}$$

Then with straigthforward calculations one gets:

$$\Phi_{quench} = \mathbb{E}\left[\frac{1}{N}\log[Z_G]\right] = 1 + e^{-N}2 \overset{N\to\infty}{\to} 1 \tag{5.78}$$

$$\Phi_{anneal} = \frac{1}{N}\log\left[\mathbb{E}[Z_G]\right] = 2 + \frac{1}{N}\log[1 + e^{-N} - e^{-2N}] \overset{N\to\infty}{\to} 2 \tag{5.79}$$

And the two averages are different! The quenched entropy is representative of typical values, while the annealed version is highly influenced by exponentially rare events.

Other than being an upper bound to the quenched entropy (first claim of Proposition 5.14), the annealed entropy is usually easier to compute, and can still be inspected to understand the problem further. Moreover, a nice result pops up for graph coloring:

Theorem 5.16 (Graph coloring Bethe and annealed entropy)**.** *For the graph coloring model, the Bethe free entropy at the fixed point is equivalent to the value of the annealed entropy:*

$$\Phi_{anneal} = \Phi_{bethe}\Big|_{\chi=\frac{1}{q}} \tag{5.80}$$

Proof. Let $G(N, M)$ denote a random graph with N nodes and M edges chosen at random. Then:

$$\Phi_{anneal} = \frac{1}{N}\log\left[\mathbb{E}_{G(N,M)}[Z_G(\beta)]\right] \tag{5.81}$$

$$\implies \mathbb{E}_{G(N,M)}\left[Z_G(\beta)\right] = \mathbb{E}_{G(N,M)}\left[\sum_{\{s_i\}_{i=1}^N} e^{-\beta\sum_{(ij)\in E}\delta_{s_i s_j}}\right] \qquad \text{Eqn. 5.10} \tag{5.82}$$

$$= q^N \underbrace{\mathbb{E}_{\{s_i\}_{i=1}^N}}_{over\ nodes}\left[\underbrace{\mathbb{E}_M}_{over\ edges}\ [e^{-\beta\sum_{(ij)\in E}\delta_{s_i s_j}}]\right] \quad \text{expanding expectation across edges and no} \tag{5.83}$$

$$= q^N \underbrace{\left[e^{-\beta}\frac{1}{q} + (1 - \frac{1}{q})\right]^M}_{single\ edge\ entropy} \tag{5.84}$$

Where we exploited the fact that edges are independently sampled, and contributions are dominated by equally represented colors.

Then:

$$\Phi_{anneal} = \frac{1}{N}\log\left[q^N\left[e^{-\beta}\frac{1}{q} + (1-\frac{1}{q})\right]^M\right] \tag{5.85}$$

$$= \log(q) + \frac{M}{N}\log\left[e^{-\beta}\frac{1}{q} + \left(1-\frac{1}{q}\right)\right] \tag{5.86}$$

$$= \log(q) + \frac{c}{2}\log\left[e^{-\beta}\frac{1}{q} + \left(1-\frac{1}{q}\right)\right] \qquad c = \frac{2M}{N} \tag{5.87}$$

Where Equation 5.87 is equal to Equation 5.51, proving the claim. □

At this point one could ask:

Is the paramagnetic fixed point Bethe approximation correct $\forall\beta > 0, \forall c$?

First of all notice that:

$$\Phi_{anneal}(\beta) = 0 \iff c_{anneal}(\beta) = \frac{2\log[q]}{\log[1-(1-e^{-\beta})\frac{1}{q}]} \tag{5.88}$$

Additionally, for $\beta \to 0^+$ research shows that proper colorings disappear with high probability if $c < c_{anneal}(\beta \to \infty) \mid \Phi(\beta \to \infty) < 0$ [CV14]. Then, it is not possible that $\Phi_{anneal} = \Phi_{quenched}$ at all average degree values. The problem lies in the fact that we basically **postulated** the $\frac{1}{q}$ fixed point, which in some cases for $\beta < 0$ was not even the maximizer. Running BP on a simple graph as a thought experiment, or with a formalized analysis as in the lecture notes[KZ21b], it is possible to identify an average degree level such that iterations of BP starting below eventually converge to $\frac{1}{q}$. Average degress c above the threshold will not converge to the paramagnetic fixed point obtained by BP. This value denoted as $\widetilde{c}_{KS}$ is in the limit [KS66; AT78]:

$$\widetilde{c}_{KS} \overset{\beta\to\infty}{\to} (q-1)^2 \tag{5.89}$$

And for Erdös-Rényi graphs where $q = 3, \beta \to \infty$ the excess degree $\widetilde{c}$ is equal to the average degree c so that:

$$c < \widetilde{c}_{KS} \implies \textit{BP conv to } \frac{1}{q} \tag{5.90}$$

$$c > \widetilde{c}_{KS} \implies \textit{BP not convergent} \tag{5.91}$$

Where when BP does not converge, finer methods such as replica symmetry breaking will be explained in later Chapters.

On the contrary, for $q \geq 4, \beta \to \infty$ we have that $\widetilde{c}_{KS} > c_{anneal}$ and no information can be extracted from the colorability investigation. It is known by probabilistic lower bounds [CV14] that the colorability threshold scales like $c_{anneal} \in O(2q\log[q])$, yet $\widetilde{c}_{KS} \in O(q^2)$ and the two will be apart as $q \to \infty$. Additionally, algorithms finding proper colorings only support levels up to half of the colorable region, i.e. $c \leq q\log[q]$. Thus, as shown in Figure 5.6, depending on c, there is:

- an easy region where problems can be algorithmically solved
- a hard region where there is a solution but not polynomial[8]
- An impossible region where being above the threshold $\widetilde{c}_{KS}$ there is no proper coloring since there are too many connections.

In the following Chapters, we will present more details about the hard region, exploring the reasons why it is difficult and how this links to BP.

[8]This is an open problem, part of the NP-Hard family

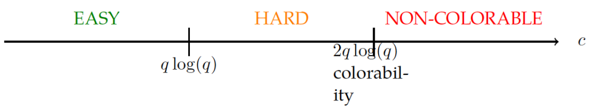


Figure 5.6: Colorability map for c average degree values

Chapter 6

Bayesian Inference

This Chapter is devoted to applying Belief Propagation (Chapter 4), the replica method (subsection 2.4.2) and cavity method (subsection 2.4.1) to solve a **statistical inference**[1] problem in the optimal Bayesian setting.

Assumption 6.1 (Notation setting)**.** Throughout the Chapter, for easier understanding of when to distinguish vectors and scalars, vectors are denoted in **bold**.

To give the reader a taste, a toy example is reported.

Example 6.2 (Vector denoising)**.** Consider a vector $\mathbf{v}^* \in \mathbb{R}^{200}$ where all the entries but one are null:

$$\mathbf{v}^* \in \mathbb{R}^{200} \mid \exists! i \ : \ v_i \neq 0$$

It is a **sparse** signal that can be visualized in a plot as in Figure 6.1. Add gaussian noise to each entry of v^*:

$$\mathbf{v} = \mathbf{v}^* + \mathbf{z}\sqrt{\Delta} \ : \ \mathbf{z} \sim \mathcal{N}^{200}(\mathbf{0}, I_{200})$$

An example is Figure 6.2, with the sparse signal perturbed by gaussian noise. The objective is finding the true non zero entry.
The task becomes harder when the variance Δ or the size of the vector N increase.

Thanks to the rich analytical theory Statistical Physics brings to the table, it is actually possible to recover a confidence level for Δ to make the problem easily solvable. Above it, random guessing will be no worse than any possible approach. In the flavour we saw previously, this will coincide with a **phase transition**.

[1] Also Known As **inverse** or **denoising** problem

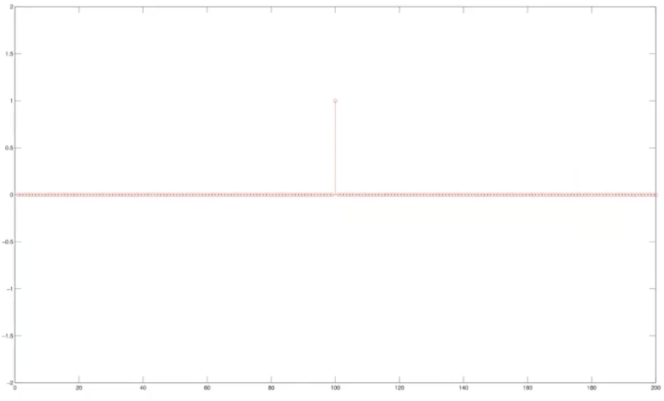


Figure 6.1: Original vector $\mathbf{v}^* \in \mathbb{R}^d$

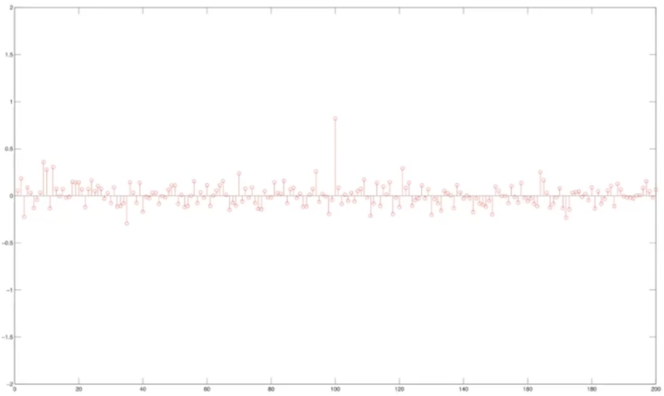


Figure 6.2: Vector with added Gaussian Noise

Assumption 6.3 (Always Gauss)**.** For simplicity, whenever noise is added to a vector, it will be gaussian, zero centered, with parameter Δ, the variance.

6.1 Bayesian Probability 101

Consider the slightly simpler problem of denoising a scalar $x \in \mathbb{R}$ coming from a **prior** distribution:

$$x \sim \mathbb{P}_X(x) \tag{6.1}$$

Example 6.4 (The 3 toy priors)**.** To see in practice how Bayesian calculations work, consider the three prior distributions:

1. **Rademacher** $X = \begin{cases} 1 & w.p.\ \frac{1}{2} \\ -1 & w.p.\ \frac{1}{2} \end{cases}$
2. **Gaussian** $X \sim \mathcal{N}(0,1)$
3. **Gauss Bernoulli process** $X = \begin{cases} 0 & w.p. \frac{1}{2} \\ \sim \mathcal{N}(0,1) & w.p. \frac{1}{2} \end{cases}$

When considering the true value of an observation, a $*$ apex will be added. Denoising can be rephrased as:

Definition 6.5 (Denoising in Optimal Bayesian inference)**.** Knowing the truth distribution of $x^* \sim \mathbb{P}_X$ (our Example 6.4), and the gaussian noise structure generating a sequence of observations:

$$\{y_i\}_{i=1}^N \ : \ y_i = x^* + z_i\sqrt{\Delta} \quad z \sim \mathcal{N}(0,1), \Delta \in \mathbb{R} \tag{6.2}$$

Find the realized value of x^*.
With this breadth of knowledge the setting is often named Optimal Bayesian inference, and it presents:

- a **prior** $\mathbb{P}_X$
- a **likelihood** $\mathbb{P}_{Y|X}$ given by Equation 6.2, known in our case to be a product of gaussians centered at the measurement x^* of the form:

$$\mathbb{P}_{Y|X}(\mathbf{y}|x) = \prod_{i=1}^{N} \frac{e^{-\frac{(y_i-x)^2}{2\Delta}}}{\sqrt{2\pi\Delta}} \tag{6.3}$$

These two measures allow us to apply the simple Bayes Theorem to recover the **posterior** as:

$$\mathbb{P}_{X|Y}(x|\mathbf{y}) = \frac{\mathbb{P}_X(x)\mathbb{P}_{Y|X}(\mathbf{y}|x)}{\underbrace{\mathbb{P}(y)}_{=\mathcal{N}\ normalization}} \tag{6.4}$$

Or in other words:

$$posterior = \frac{prior \times likelihood}{evidence}$$

For the three distributions of Example 6.4 it is possible to recover a closed form of the posterior with some work.
A Rademacher prior returns a distribution of the kind:

$$\mathbb{P}^{Rad}_{X|Y}(x|\mathbf{y}) = \frac{1}{1 + e^{-x\sum_i \frac{y_i}{\Delta}}} = \sigma\left(x \sum_i \frac{y_i}{\Delta}\right) \tag{6.5}$$

Where $\sigma(\cdot)$ is the sigmoid function.
For the Gaussian prior we have that the posterior is gaussian as well:

$$x \sim \mathcal{N}(0,1), \mathbf{y} \sim \mathcal{N}^N(x, \Delta) \implies x|\mathbf{y} \sim \mathcal{N}\left(\mu = \frac{\sum_i y_i}{n+\Delta}, \sigma^2 = \frac{\Delta}{n+\Delta}\right) \tag{6.6}$$

Lastly, the Gauss-Bernoulli process is slightly more convoluted, with a posterior density of the form:

$$\mathbb{P}^{GaussBern}_{X|Y}(x|\mathbf{y}) = \frac{\delta(x)}{1 + \sqrt{\frac{\Delta}{N+\Delta}}e^{\frac{(\sum_i y_i)^2}{2\Delta(N+\Delta)}}} + \frac{\mathcal{N}\left(x;\ \frac{\sum_i y_i}{N+\Delta}, \frac{\Delta}{N+\Delta}\right)}{1 + \sqrt{\frac{N+\Delta}{\Delta}}e^{\frac{-(\sum_i y_i)^2}{2\Delta(N+\Delta)}}} \tag{6.7}$$

Essentially, not much more can be said about these cases. Having access to the prior and the likelihood, the posterior is computed. In cases in which either of the two are not available, there are methods to recover bounds on the result.

The posterior $\mathbb{P}_{X|Y}$ is a distribution over values. In some cases **pointwise estimation** is preferred.

Definition 6.6 (Estimator $\widehat{x}(\mathbf{y})$)**.** Given a set of observations $\{y_i\}_{i=1}^N \in \mathcal{Y} = \mathbb{R}^N$ an estimator is a function:

$$\widehat{x} : \mathcal{Y} \to \mathcal{X} \tag{6.8}$$

Definition 6.7 (Maximum a Posteriori, MAP)**.** The most intuitive estimator is the MAP:

$$MAP := \arg\max_x \left\{\mathbb{P}_{X|Y}(x, \mathbf{y})\right\} \tag{6.9}$$

It is very useful tool, especially when the density is concave and gradient descent can be run.

A naturally arising question could be:

Is MAP the best possible estimator? If so, in which sense?

To answer this, we need to define a measure of correctness for estimators and a systematic way to evaluate comparisons. For this reason, the next definitions are reported.

Definition 6.8 (Error Function $\mathcal{L}(\cdot,\cdot)$)**.** The Error[2] is a function:

$$\mathcal{L} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \quad (\widehat{x}(\mathbf{y}), x^*) \to \mathcal{L}(\widehat{x}(\mathbf{y}), x^*) \tag{6.10}$$

Example 6.9 (Common Error functions)**.** Among the most famous costs we list:

- L2 or squared error $\mathcal{L}(\widehat{x}(\mathbf{y}), x^*) = (\widehat{x}(\mathbf{y}) - x^*)^2$
- the mean average error $\mathcal{L}(\widehat{x}(\mathbf{y}), x^*) = |\widehat{x}(\mathbf{y}) - x^*|$
- the discrete error counter (a Kroenecker delta), well versed for discrete spaces $\mathcal{L}(\widehat{x}(\mathbf{y}), x^*) = \delta_{\widehat{x}x^*}$

While the estimator is for a single instance $(x^*, \mathbf{y})$ of the problem, we are interested in the probabilistic stability of our estimator across all the possible scenarios. This motivates the introduction of an *averaged* object.

Definition 6.10 (Averaged risk/error $\mathcal{R}^{av}(\cdot)$, Bayes risk $\mathcal{R}^{bayes}(\widehat{x})$)**.** Given an estimator as in Definition 6.6 and an error function as in Definition 6.8, the averaged risk is over many experiments and signals, eventually taking the form of of a functional:

$$\begin{aligned}
\mathcal{R}^{av}(\widehat{x}) &= \mathbb{E}_{x^*,\mathbf{y}}\Big[\mathcal{L}(\widehat{x}(\mathbf{y}), x^*)\Big] && (6.11)\\
&= \int dx^* d\mathbf{y} \mathbb{P}_{X,Y}(x^*, \mathbf{y}) \mathcal{L}(\widehat{x}(\mathbf{y}), x^*) && (6.12)\\
&= \int d\mathbf{y} \mathbb{P}_Y(\mathbf{y}) \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) \mathcal{L}(\widehat{x}(\mathbf{y}), x^*) && (6.13)\\
&= \mathbb{E}_Y\Big[\mathcal{R}^{posterior}(\widehat{x}, \mathbf{y})\Big] && (6.14)
\end{aligned}$$

Where the objective is minimizing the averaged error, at a value denoted as Bayes risk:

$$\mathcal{R}^{Bayes}(\widehat{x}) = \min_{\widehat{x}} \Big\{\mathcal{R}^{av}(\widehat{x})\Big\} \tag{6.15}$$

And the optimal estimator attaining it.

Clearly, the risks of Definition 6.10 depend heavily on the choice of $\mathcal{L}(\cdot,\cdot)$. Some easy results provide further information for the analysis.

Proposition 6.11 (Best estimators for different error functions)**.** *Go back to Example 6.9, then:*

1. $\mathcal{L}(\widehat{x}(\boldsymbol{y}), x^*) = (\widehat{x} - x^*)^2 \implies \widehat{x}_{MMSE}(\boldsymbol{y}) = \arg\min_{\widehat{x}} \Big\{\mathcal{R}^{av}(\widehat{x})\Big\} = \mathbb{E}_{X|Y}[x|\boldsymbol{y}]$
2. $\mathcal{L}(\widehat{x}(\boldsymbol{y}), x^*) = |\widehat{x} - x^*| \implies \widehat{x}_{MMAE}(\boldsymbol{y}) = \arg\min_{\widehat{x}} \Big\{\mathcal{R}^{av}(\widehat{x})\Big\} = median_{X|Y}(x|\boldsymbol{y})$
3. $\mathcal{L}(\widehat{x}(\boldsymbol{y}), x^*) = \delta_{\widehat{x}x^*} \implies \widehat{x}_{OBD}(\boldsymbol{y}) = \arg\min_{\widehat{x}} \Big\{\mathcal{R}^{av}(\widehat{x})\Big\} = MAP$

Where OBD is short for Optimal Bayesian Decision.

Proof. **(Claim** 1**)** Consider Equation 6.14, then:

$$\implies \arg\min_{\widehat{x}} \Big\{\mathcal{R}^{av}(\widehat{x})\Big\} = \arg\min_{\widehat{x}} \Big\{\mathcal{R}^{posterior}(\widehat{x})\Big\} \tag{6.16}$$

[2] Also Known As Loss, Cost, Energy, depending on the context

To minimize the expression, we look for a zero derivative point and find that:

$$\frac{\partial \mathcal{R}^{posterior}(\widehat{x})}{\partial \widehat{x}} = 0 \tag{6.17}$$

$$= \frac{\partial}{\partial \widehat{x}}\Big(\int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})(\widehat{x}-x^*)^2\Big) \tag{6.18}$$

$$= \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) 2(\widehat{x}-x^*) \tag{6.19}$$

$$\iff \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})\widehat{x} = \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) x^* \tag{6.20}$$

$$\iff \underbrace{\widehat{x}\int dx \mathbb{P}_{X|Y}(x|\mathbf{y})}_{=1} = \int dx \mathbb{P}_{X|Y}(x|\mathbf{y}) x \qquad \text{take out } \widehat{x}\text{, change dummy } x^* \tag{6.21}$$

$$\iff \widehat{x}(\mathbf{y}) = \mathbb{E}_{X|Y}[x|\mathbf{y}] \tag{6.22}$$

Another proof is proposed in the original lecture notes [KZ21b].
(Claim 2**)** Following the same fashion we first express the posterior risk

$$\mathcal{R}^{posterior}(\widehat{x}) = \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})|\widehat{x}-x^*| \tag{6.23}$$

$$= \int_{-\infty}^{\widehat{x}} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})(-\widehat{x}+x^*) + \int_{\widehat{x}}^{\infty} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})(\widehat{x}-x^*) \quad \text{split integral} \tag{6.24}$$

and then derive it:

$$\frac{\partial \mathcal{R}^{posterior}(\widehat{x})}{\partial \widehat{x}} = 0 \tag{6.25}$$

$$= \frac{\partial}{\partial \widehat{x}}\Big(\int_{-\infty}^{\widehat{x}} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})(-\widehat{x}+x^*) + \int_{\widehat{x}}^{\infty} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})(\widehat{x}-x^*)\Big) \tag{6.26}$$

$$= -\int_{-\infty}^{\widehat{x}} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) + \int_{\widehat{x}}^{\infty} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) \tag{6.27}$$

$$\iff \int_{-\infty}^{\widehat{x}} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) = \int_{\widehat{x}}^{\infty} dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y}) \tag{6.28}$$

$$\iff \mathbb{P}_{X|Y}(x \leq \widehat{x}) = \mathbb{P}_{X|Y}(x > \widehat{x}) \tag{6.29}$$

$$\iff \widehat{x}(\mathbf{y}) = median_{X|Y}(x|\mathbf{y}) \tag{6.30}$$

(Claim 3**)** For x^* discrete we have:

$$\begin{aligned}
\mathcal{R}^{posterior}(\widehat{x}) &= \int dx^* \mathbb{P}_{X|Y}(x^*|\mathbf{y})\delta_{\widehat{x}x^*} \\
&= \sum_{x^*} \mathbb{P}_{X|Y}(x^*|\mathbf{y})\delta_{\widehat{x}x^*} \\
&= \sum_{x^*} \mathbb{P}_{X|Y}(x^*|\mathbf{y})\mathbb{I}(x^* \neq \widehat{x}) && \text{discrete error counter} \\
&= \underbrace{\sum_{x^*} \mathbb{P}_{X|Y}(x^*|\mathbf{y})}_{=1} - \sum_{x^*} \mathbb{P}_{X|Y}(x^*|\mathbf{y})\mathbb{I}(x^* = \widehat{x}) && \text{errors = 1- correct} \\
&= 1 - \mathbb{P}_{X|Y}(\widehat{x}|\mathbf{y})
\end{aligned}$$

Minimizing such function is equivalent to finding the maximizer of the posterior, which is the MAP (Definition 6.7). □

6.2 The Statistical Physics Perspective and two Tools

It might be tempting to observe that Bayes' rule takes the form:

$$\mathbb{P}_{X|Y}(x|\mathbf{y}) = \frac{\mathbb{P}_X(x)\mathbb{P}_{Y|X}(\mathbf{y}|x)}{\mathbb{P}_Y(\mathbf{y})} \tag{6.31}$$

$$= \underbrace{\frac{1}{\mathbb{P}_Y(\mathbf{y})}}_{=Z(\mathbf{y})} exp\Big\{ \underbrace{\log[\mathbb{P}_X(x)\mathbb{P}_{Y|X}(\mathbf{y}|x)]}_{=-\beta\mathscr{H}} \Big\} \tag{6.32}$$

$$= \mathbb{P}_{Gibbs,\mathbf{y}}(x) \tag{6.33}$$

Yet, we will use a slightly different convention that turns up being easier for computations.

Definition 6.12 (Bridge from Bayesian Theory to Statistical Physics)**.** Using the fact that noise is Gaussian (Assumption 6.3), aim to completely isolate x and obtain:

$$\mathbb{P}_{X|Y}(x|\mathbf{y}) = \frac{1}{\mathbb{P}_Y(\mathbf{y})}\mathbb{P}_X(x)\frac{\prod_i e^{-\frac{(y_i-x)^2}{2\Delta}}}{\sqrt{2\pi\Delta}} \tag{6.34}$$

$$= \underbrace{\frac{e^{-\frac{\sum_i y_i}{2\Delta}}}{(\sqrt{2\pi\Delta})^N\mathbb{P}_Y(\mathbf{y})}}_{:=Z(\mathbf{y})}\mathbb{P}_X(x)e^{\sum_i -\frac{x^2}{2\Delta}+\frac{xy_i}{\Delta}} \qquad \text{isolate } x \tag{6.35}$$

$$:= \frac{\mathbb{P}_X(x)e^{\sum_i -\frac{x^2}{2\Delta}+\frac{xy_i}{\Delta}}}{Z(\mathbf{y})} \tag{6.36}$$

Where we recognize a normalization term which takes the form:

$$Z(\mathbf{y}) = \frac{e^{-\frac{\sum_i y_i}{2\Delta}}}{(\sqrt{2\pi\Delta})^N\mathbb{P}_Y(\mathbf{y})} = \int dx\frac{e^{-\frac{\sum_i y_i}{2\Delta}}}{(\sqrt{2\pi\Delta})^N\mathbb{P}_Y(\mathbf{y})}\mathbb{P}_X(x) = \int dx e^{\frac{\sum_i -x^2}{2\Delta}+\frac{\sum_i y_i x}{\Delta}}\mathbb{P}_X(x) \tag{6.37}$$

$$\beta = 1 \tag{6.38}$$

$$\mathscr{H}(\mathbf{x}) = \frac{\sum_i -x^2}{2\Delta} + \frac{\sum_i y_i x}{\Delta} \tag{6.39}$$

And can be read as the sum over possible configurations x of the numerator.

Observation 6.13 ($Z(\cdot)$ as a Likelihood Ratio)**.** *It is remarkable that such a formalism is equivalent to asserting that:*

$$Z(\boldsymbol{y}) = \frac{\mathbb{P}_Y^{random}(y)}{\mathbb{P}_Y(\boldsymbol{y})} \tag{6.40}$$

Where random means a purely random noise $y \sim \mathcal{N}(0,\Delta)$ and on the denominator we find the probability of $\boldsymbol{y}$ *coming from the true data generation process.*

Remark (Notation for Averages)**.** An average of a quantity over the posterior is of the form:

$$\mathbb{E}_{X|Y}[\dots|\mathbf{y}] = \int dx\mathbb{P}_{X|Y}(x|\mathbf{y})\Big[\dots\Big]$$

$$= \frac{1}{Z(\mathbf{y})}\int dx\mathbb{P}_X(x)\Big[\dots\Big] \qquad \text{Def 6.12}$$

$$= \langle\dots\rangle_{Boltzmann}$$

And will be denoted by brackets as in Chapter 3 and the whole document.
Similarly, an average over the disorder[3] $(X^*, \mathcal{Y})$ will be denoted by the $\mathbb{E}_Y$ script.

[3] i.e. true value, configurations attaining it.

We are ultimately interested in using the setting of Definition 6.12 to recover the free entropy:

$$\Phi_N = \mathbb{E}_Y\left[\log[Z(\mathbf{y})]\right] \tag{6.41}$$

Other quantities are equivalent to Equation 6.41.

Definition 6.14 (Statistical Entropy $H(\cdot)$)**.** For a random variable $Y \in \mathcal{Y}$ the statistical entropy is:

$$H(Y) := -\mathbb{E}_Y[\log(Y)] \tag{6.42}$$

Definition 6.15 (Kullback-Leibler Divergence $D_{KL}(\cdot||\cdot)$)**.** For two distributions f, g the KL divergence is:

$$D_{KL}(f||g) := \int dx f(x) \log\left[\frac{f(x)}{g(x)}\right] \tag{6.43}$$

Definition 6.16 (Mutual Information $I(\cdot,\cdot)$)**.** For two random variables $X \in \mathcal{X}, Y \in \mathcal{Y}$ the mutual information is:

$$I(X,Y) := D_{KL}\left(\mathbb{P}_{XY}||\mathbb{P}_X\mathbb{P}_Y\right) \tag{6.44}$$

Proposition 6.17 (Equivalence of free entropy, statistical entropy, mutual information)**.** *It holds that* $\forall X \in \mathcal{X}, Y \in \mathcal{Y}$*:*

1. $H(Y) = N\frac{\mathbb{E}_Y[y^2]}{2\Delta} + \frac{N}{2}\log[2\pi\Delta] - \Phi_N$
2. $I(X,Y) = H(Y) - H(Y|X) = -\Phi_N + N\frac{\mathbb{E}_X[x^2]}{2\Delta}$

The path to obtaining a Φ_N formula passes through the knowledge of two important mathematical facts, which are presented below.

First of all recall that by A.4:

$$\int dx(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\sigma^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.45}$$

Which will be used in the next result.

Lemma 6.18 (Stein's Lemma)**.** *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$*. Let* g *be differentiable and such that* $\exists\mathbb{E}[(X-\mu)g(X)]$ *and* $\exists\mathbb{E}[|g'(X)|]$*. Then:*

$$\mathbb{E}[g(X)(X-\mu)] = \sigma^2\mathbb{E}[g'(X)] \tag{6.46}$$

Which in the particular case of a gaussian standard distribution means:

$$X \sim \mathcal{N}(0,1) \implies \mathbb{E}[Xg(X)] = \mathbb{E}[g'(X)] \tag{6.47}$$

Proof. Expressing the LHS of Equation 6.46 we get applying integration by parts[4]:

$$\mathbb{E}[g(X)(X-\mu)] = \int dx \mathbb{P}_X(x)[g(x)(x-\mu)] \tag{6.48}$$

$$= \int dx \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (x-\mu)}_{f'} \underbrace{g(x)}_{h} \tag{6.49}$$

$$= \left[g(x) \int dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (x-\mu) \right]\Bigg|_{-\infty}^{\infty} - \int dx g'(x) \Bigg(\int d\chi \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\chi-\mu)^2}{2\sigma^2}} (\chi-\mu) \Bigg) \tag{6.50}$$

$$= \left[g(x)(-\sigma^2) e^{\frac{-(x-\mu)^2}{2\sigma^2}} \right]\Bigg|_{-\infty}^{\infty} - \int dx g'(x)(-\sigma^2) e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{6.51}$$

$$= \left[g(x)(-\sigma^2) e^{\frac{-(x-\mu)^2}{2\sigma^2}} \right]\Bigg|_{-\infty}^{\infty} + \sigma^2 \mathbb{E}[g'(X)] \tag{6.52}$$

Where in Equation 6.51 we applied Lemma A.4 on both the $d\chi$ integral and the integral in the first term. What is missing is proving that the first term is null, this can be realized through some additional arguments.
First of all, we denote a normal distribution as $s(\cdot)$ and inspect the product of $s(\cdot)g(\cdot)$ at the extremes $x \to \pm\infty$ where we need to evaluate it. Recall that for finite μ (the mean of $h(\cdot)$) we will have that:

$$\lim_{x \to \pm\infty} h(x) = 0$$

It is also useful to express $g(\cdot)$ differently as:

$$g(x) = g(z) + + \int_z^x dy g'(y)$$

Now, let $z > \mu$ ensuring that $\forall x > z$:

$$\begin{aligned} g(x)s(x) &= g(z)h(x) + s(x) \int_z^x dy g'(y) \\ &\leq g(z)s(x) + \int_z^x dy s(y) g'(y) \qquad \text{basic integral properties} \\ \implies \lim_{x\to\infty} \left\{ \sup g(x)h(x) \right\} &\leq \int_z^\infty dy s(y) g'(y) \end{aligned}$$

When $x \to \infty$ we have that $s(x) \to 0$. Moreover, to evaluate the term at $+\infty$ we need also $z \to \infty$. In this setting, the derivation above ensures that:

$$\lim_{x\to\infty} \left\{ \sup g(x)h(x) \right\} = 0$$

The same reasoning can be made for $x \to -\infty$, and the claim is proved since:

$$\begin{aligned} \mathbb{E}[g(X)(X-\mu)] &= \underbrace{\left[g(x)(-\sigma^2) e^{\frac{-(x-\mu)^2}{2\sigma^2}} \right]\Bigg|_{-\infty}^{\infty}}_{=0} + \sigma^2 \mathbb{E}[g'(X)] \\ &= \sigma^2 \mathbb{E}[g'(X)] \end{aligned}$$

□

Another famous result by the Japanese Statistical Physics pioneer Nishimori is of pivotal importance.

[4] $\int f'h = [fh] - \int fh'$

Theorem 6.19 (Nishimori Identity)**.** *Given k samples from the posterior distribution:*

$$\left\{X^{(i)}\right\}_{i=1}^{k} \ : \ X^{(i)} \sim \mathbb{P}_{X|Y} \forall i \tag{6.53}$$

Then for all f continuous and bounded we can "switch" one copy of $X^{(i)}$ with the true variable X^*:

$$\mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \ldots, X^{(k)}\right)\right\rangle_k\right] = \mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \ldots, X^{(k-1)}, X^*\right)\right\rangle_{k-1}\right] \tag{6.54}$$

Proof. We have that:

$$\mathbb{E}_{X^*,Y}[\langle f(Y, X^{(1)}, \ldots, X^{(k-1)}, X^*)\rangle_{k-1}] = \int dx^* dy \mathbb{P}_{X^*,Y} \langle f(Y, X^{(1)}, \ldots, X^{(k-1)}, X^*)\rangle_{k-1} \quad \text{expand } \mathbb{E} \tag{6.55}$$

$$= \int dy \mathbb{P}_Y \int dx^* \mathbb{P}_{X^*|Y} \langle f(Y, X^{(1)}, \ldots, X^{(k-1)}, X^*)\rangle_{k-1} \tag{6.56}$$

$$= \int dy \mathbb{P}_Y \int dx^{(k)} \mathbb{P}_{X^{(k)}|Y} \langle f(Y, X^{(1)}, \ldots, X^{(k-1)}, X^{(k)})\rangle_{k-1} \tag{6.57}$$

$$= \int dy \mathbb{P}_Y \langle f(Y, X^{(1)}, \ldots, X^{(k-1)}, X^{(k)})\rangle_k \tag{6.58}$$

$$= \mathbb{E}[\langle f(Y, X^{(1)}, \ldots, X^{(k)})\rangle_k] \tag{6.59}$$

Where the change from X^* to $X^{(k)}$ can be done since it is just a dummy index in the integral. □

Theorem 6.19 has interesting consequences!

Observation 6.20 (Magnetization with Nishimori)**.** *In the context of overlaps for Optimal Bayesian inference we can see that a magnetization reduces to:*

$$m = \mathbb{E}[\langle x\rangle x^*] = \mathbb{E}[\langle x\rangle\langle x'\rangle] = \mathbb{E}[\langle x\rangle^2] = q \tag{6.60}$$

Since X^* is equivalent in expectation to a copy X' which is itself identitic to the X considered.

Lemma 6.21 (Optimal Bayesian inference MMSE)**.** *In Optimal Bayesian inference it holds that:*

$$MMSE = \mathbb{E}_{X^*,Y}\left[(x^* - \langle x\rangle)^2\right] = \rho - m \quad : \quad \rho = \mathbb{E}_{X^*,Y}\left[(x^*)^2\right] \tag{6.61}$$

Where ρ can be seen as the self overlap.

Proof. We just need to expand the product inside the expectation to find that:

$$\begin{aligned}
\mathbb{E}_{X^*,Y}\left[(x^* - \langle x\rangle)^2\right] &= \mathbb{E}_{X^*,Y}\left[(x^*)^2\right] + \mathbb{E}_{X^*,Y}[\langle x\rangle^2] - 2\mathbb{E}_{X^*,Y}[\langle x\rangle x^*] \\
&= \rho + q - 2m \\
&= \rho + m - 2m && \text{Obs 6.20 by Theorem 6.19} \\
&= \rho - m
\end{aligned}$$

□

Lemma 6.21 guarantees that we only need to compute m (the overlaps), since ρ is rather easy to get. We are only missing one piece in the puzzle, presented as a Theorem below.

Theorem 6.22 (*I-MMSE* Theorem)**.** *For a single disturbed measurement:*

$$Y = X^* + \sqrt{\Delta}Z \quad Z \sim \mathcal{N}(0,1) \tag{6.62}$$

It holds that:

$$\frac{\partial \Phi_N(\Delta)}{\partial \Delta^{-1}} = \frac{1}{2}m \tag{6.63}$$

$$\frac{\partial I(\Delta)}{\partial \Delta^{-1}} = \frac{1}{2}(\rho - m) \qquad \textit{the MMSE} \tag{6.64}$$

Proof. Firstly, we express the free entropy as a function of the noise to take its expectation:

$$\Phi_N = \mathbb{E}_Y\Big[\log[Z(y)]\Big] \tag{6.65}$$

$$= \mathbb{E}_Y\Big[\log \int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xy}{\Delta}}\Big] \qquad \text{Eqn. 6.37} \tag{6.66}$$

$$= \mathbb{E}_{X^*,Z}\Big[\log \int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{xz}{\sqrt{\Delta}}}\Big] \qquad \text{Eqn. 6.62} \tag{6.67}$$

Then, the partial derivative becomes[5]:

$$\frac{\partial \Phi_N(\Delta)}{\partial \Delta^{-1}} = \mathbb{E}_{X^*,Z}\Big\{ \underbrace{\frac{1}{Z(y)} \int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{xz}{\sqrt{\Delta}}}}_{\textit{Boltzmann weight}} \Big[-\frac{x^2}{2} + xx^* + \frac{xz}{2}\sqrt{\Delta}\Big]\Big\} \tag{6.68}$$

$$= \mathbb{E}_{X^*,Z}\Big[\Big\langle -\frac{x^2}{2} + xx^* + \frac{xz}{2}\sqrt{\Delta}\Big\rangle\Big] \tag{6.69}$$

$$= -\frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\langle x^2\rangle\Big] + \mathbb{E}_{X^*,Z}\Big[x^*\langle x\rangle\Big] + \frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\underbrace{z}_{Z\sim\mathcal{N}(0,1)} \underbrace{\sqrt{\Delta}\langle x\rangle}_{=g(z)}\Big] \tag{6.70}$$

$$= -\frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\langle x^2\rangle\Big] + \underbrace{\mathbb{E}_{X^*,Z}\Big[x^*\langle x\rangle\Big]}_{\textit{Nishimori Thm } 6.19} + \frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\underbrace{\langle x^2\rangle - \langle x\rangle^2}_{=g'(z)}\Big] \tag{6.71}$$

$$= -\frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\langle x^2\rangle\Big] + \mathbb{E}_{X^*,Z}\Big[\langle x\rangle^2\Big] + \frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\langle x^2\rangle\Big] - \frac{1}{2}\mathbb{E}_{X^*,Z}\Big[\langle x\rangle^2\Big] \tag{6.72}$$

$$= \frac{1}{2}m \tag{6.73}$$

Where, expanding the passage from Equation 6.70 to 6.71 Stein's Lemma 6.18 is used where we have[6]:

$$\langle {\color{blue}x}\rangle = \frac{1}{Z(y)} \int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{xz}{\sqrt{\Delta}}} {\color{blue}x} = \frac{\Big[\int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{xz}{\sqrt{\Delta}}} {\color{blue}x}\Big]}{\int dx \mathbb{P}_X(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{xz}{\sqrt{\Delta}}}} \tag{6.74}$$

[5] Ignoring some regularity conditions which we enforce

[6] ${\color{blue}x}$ is used to distinguish the weights from the function in the expectation

So that using the classic differentiation rule for fractions:

$$
\begin{aligned}
&\frac{\partial}{\partial z}\left(\frac{num(z)}{den(z)}\right) = \frac{[\partial_z num(z)]den(z) - [\partial_z den(z)]num(z)}{[den(z)]^2}\\
&num(z) = \int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\\
&den(z) = Z(y) \mid y = x^* + \sqrt{\Delta}z\\
&\partial_z num(z) = \int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\underbrace{\frac{\partial}{\partial z}\left(-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}\right)}_{=\frac{x}{\sqrt{\Delta}}}\\
&\partial_z den(z) = \int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}\underbrace{\frac{\partial}{\partial z}\left(-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}\right)}_{=\frac{x}{\sqrt{\Delta}}}
\end{aligned}
$$

We can easily compute:

$$
\begin{aligned}
\frac{\partial\sqrt{\Delta}\langle x\rangle}{\partial z} &= \frac{\sqrt{\Delta}}{\underbrace{[Z(y)]^2}_{=[den(z)]^2}}\Bigg\{\underbrace{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\frac{x}{\sqrt{\Delta}}}_{=\partial_z num(z)}\underbrace{Z(y)}_{=den(z)} \qquad (6.75)\\
&\quad - \left[\underbrace{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}\frac{x}{\sqrt{\Delta}}}_{=\partial_z den(z)}\right]\underbrace{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x}_{=num(z)}\Bigg\}\\
&= \frac{\sqrt{\Delta}}{[Z(y)]^2}\frac{Z(y)}{\sqrt{\Delta}}\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x^2 \qquad (6.76)\\
&\quad - \sqrt{\Delta}\frac{1}{\sqrt{\Delta}}\frac{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x}{Z(y)}\frac{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x}{Z(y)}\\
&= \frac{1}{Z(y)}\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x^2 - \left[\frac{\int dx\mathbb{P}_X(x)e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x}{Z(y)}\right]^2 \qquad (6.77)\\
&= \langle x^2\rangle - \langle x\rangle^2 \qquad (6.78)
\end{aligned}
$$

Which plugged into the expectation returns what was written in Equation 6.71.
The result of Equation 6.64 can be obtained using Proposition 6.17 and the just proved Equation 6.63. □

6.3 Denoising Sparse Vectors

Thanks to Theorem 6.22, upon having knowledge of Φ_N $\forall\Delta$, it is rather straightforward to find the magnetization m.

Having discussed all the necessary theory, we now move from $x^* \in \mathbb{R}$ to $\mathbf{x}^* \in \mathbb{R}^d$ where the original vector is sparse as before meaning:

$$\mathbf{x}^* \in \mathbb{R}^d \mid \exists! i \;:\; x_i \neq 0 \tag{6.79}$$

Assumption 6.23 (Power of two size). To simplify calculations, assume that $d = 2^N, N \in \mathbb{N}$.

Considering the rescaled noisy variable[7]:

$$Y = X^* + \frac{\sqrt{\Delta}}{N} Z \quad Z \sim \mathcal{N}(0,1) \tag{6.80}$$

We are asked to compute the MMSE and find a *critical* Δ_c, above which the problem becomes impossible (i.e. where the phase transition takes place).

Using the result of Theorem 6.22 we start from the free energy for a set of observations $\mathbf{y}, \mathbf{x}$:

$$\Phi(\Delta) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{X|Y}\Big[\log(Z(\mathbf{y})\Big] \quad \mathbb{P}_{X|Y}(x|\mathbf{y}) = \frac{1}{Z(\mathbf{y})} \mathbb{P}_X(\mathbf{x}) exp\Big\{ - \frac{\langle \mathbf{x}, \mathbf{x}\rangle}{\frac{2\Delta}{N}} + \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\frac{\Delta}{N}} \Big\} \tag{6.81}$$

Where the prior is such that:

$$\mathbb{P}_X(\mathbf{x}) = \frac{1}{d} \sum_{i=1}^{d} \delta_{\mathbf{x},\mathbf{e}_i} = \frac{1}{2^N} \sum_i \delta_{\mathbf{x},\mathbf{e}_i} \tag{6.82}$$

For elementary vectors indexed by i of the form $\mathbf{e}_i$. This prior goes over all the valid choices[8] for x^*, and ignores any other vector of dimension d.
Such a prior influences the partition function deeply. First of all, for elementary vectors we intuitively have that:

$$\langle \mathbf{e}_i, \mathbf{e}_i \rangle = 1 \quad \forall i \tag{6.83}$$
$$\langle \mathbf{e}_i, \mathbf{y} \rangle = y_i \quad \forall i, \forall \mathbf{y} \in \mathbb{R}^d \tag{6.84}$$

Additionally notice that the observed signals, originating from a vector with the form of Equation are:

$$y_i = x_i^* + \sqrt{\Delta} z_i = \begin{cases} \sqrt{\Delta} z_i & i \neq i^* \\ 1 + \sqrt{\Delta} z_{i^*} & i^* \end{cases} \tag{6.85}$$

Where i^* is the index at which the vector $\mathbf{x}^*$ is non zero.
So that:

$$\begin{aligned}
\implies Z(\mathbf{y}) &= \int d\mathbf{x} \mathbb{P}_X(\mathbf{x}) exp\Big\{ - \frac{\langle \mathbf{x}, \mathbf{x}\rangle}{\frac{2\Delta}{N}} + \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\frac{\Delta}{N}} \Big\} && (6.86)\\
&= \int d\mathbf{x} \frac{1}{2^N} \sum_i \delta_{\mathbf{x}\mathbf{e}_i} exp\Big\{ - \frac{\langle \mathbf{x}, \mathbf{x}\rangle}{\frac{2\Delta}{N}} + \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\frac{\Delta}{N}} \Big\} \quad \text{Eqn. 6.82} && (6.87)\\
&= \frac{1}{2^N} \sum_{i=1}^{d} e^{-\frac{N}{2\Delta} + \frac{N y_i}{\Delta}} \quad \text{Eqns. 6.83, 6.84} && (6.88)\\
&= \frac{1}{2^N} \sum_{i \neq i^*} e^{-\frac{N}{2\Delta} + \frac{N\sqrt{\Delta} z_i}{\Delta}} + e^{-\frac{N}{2\Delta} + \frac{N}{\Delta}(1+\sqrt{\Delta} z_i)} \quad \text{Eqn. 6.85} && (6.89)\\
&= \frac{1}{2^N} \sum_{i=1}^{d} e^{-\frac{N}{2\Delta} + \frac{N\delta_{ii^*}}{\Delta} + \sqrt{\frac{N}{\Delta}} z_i} \quad \text{short form} && (6.90)
\end{aligned}$$

A closed form of Z allows us to apply Theorem 6.22 to extract the MMSE at a given variance Δ. Specifically the best overlap:

$$q = 2 \frac{\partial}{\partial \Delta^{-1}} \Phi(\Delta) \tag{6.91}$$

[7]The $\frac{1}{N}$ constant factor is added to obtain a **sharp** phase transition
[8]Vectors with only one non zero unit entry.

Will promise us that:

$$MMSE = \rho - m = \rho - q = \mathbb{E}_{X^*|Y}\Big[\underbrace{\langle \mathbf{x}^*, \mathbf{x}^* \rangle}_{=1\ Eqn.6.83}\Big] - q = 1 - q \tag{6.92}$$

Is the best achievable error under square loss. The next Theorem closes the circle, proposing a function that perfectly describes q as a function of Δ.

To clear out some passages of the proof, recall that a quite immediate identity in Gaussian integrals from Lemma A.3 is the following:

$$X \sim \mathcal{N}(0,1) \implies \mathbb{E}_X\left[e^{\kappa x}\right] = e^{\frac{1}{2}\kappa^2} \tag{6.93}$$

Theorem 6.24 (Denoising MMSE function)**.** *The function:*

$$f(\Delta) = \begin{cases} \frac{1}{2\Delta} - \log 2 & \Delta \leq \frac{1}{2\log 2} \\ 0 & \Delta > \frac{1}{2\log 2} \end{cases} \tag{6.94}$$

Is such that:

$$\Phi(\Delta) = \lim_{N\to\infty} \Phi_N(\Delta) = \lim_{N\to\infty}\left\{\frac{\mathbb{E}[\log(Z)]}{N}\right\} = f(\Delta) \tag{6.95}$$

Proof. **(Claim $\Phi_N(\Delta) \geq f(\Delta) + o(1)\forall\Delta$)**
It holds that:

$$\begin{aligned}
\Phi_N(\Delta) &= \frac{1}{N}\mathbb{E}_{X^*,Y}\left\{\log\left[\frac{1}{2^N}\sum_{i=1}^{d} e^{-\frac{N}{2\Delta}+\frac{N\delta_{ii^*}}{\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right]\right\} && \text{Eqn. 6.90} \\
&= \frac{1}{N}\mathbb{E}_{X^*,Y}\left\{-N\log(2) + \log\left[\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + e^{-\frac{N}{2\Delta}+\frac{N}{\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}}\right]\right\} \\
&\geq \frac{1}{N}\mathbb{E}_{X^*,Y}\left\{-N\log(2) + \log\left[e^{-\frac{N}{2\Delta}+\frac{N}{\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}}\right]\right\} \\
&\geq -\log(2) + \frac{1}{N}\mathbb{E}_{X^*,Y}\left\{-\frac{N}{2\Delta}+\frac{N}{\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}\right\} \\
&\geq -\log(2) - \frac{1}{2\Delta} + \frac{1}{\Delta} + \frac{1}{N}\sqrt{\frac{N}{\Delta}}\underbrace{\mathbb{E}_{X^*,Y}[z_{i^*}]}_{=0} \\
&\geq -\log(2) + \frac{1}{2\Delta}
\end{aligned}$$

(Claim $\Phi_N(\Delta) \leq f(\Delta)\forall\Delta$)
The second inequality is a slightly convoluted application of Jensen's Inequality[9]:

$$\Phi_N(\Delta) = \frac{1}{N}\mathbb{E}_{X^*,Y}\left\{ \log\left[\frac{1}{2^N}\sum_{i=1}^{d} e^{-\frac{N}{2\Delta}+\frac{N\delta_{ii^*}}{\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right]\right\} \qquad \text{Eqn. 6.90} \tag{6.96}$$

$$= \frac{1}{N}\mathbb{E}_{z_i,z_{i^*}}\left\{ \log\left[\frac{1}{2^N}\sum_{i=1}^{d} e^{-\frac{N}{2\Delta}+\frac{N\delta_{ii^*}}{\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right]\right\} \tag{6.97}$$

$$= \frac{1}{N}\mathbb{E}_{z_i,z_{i^*}}\left\{ \log\left[\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + e^{-\frac{N}{2\Delta}+\frac{N}{\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}}\right]\right\} \tag{6.98}$$

$$= \frac{1}{N}\mathbb{E}_{z_i,z_{i^*}}\left\{ \log\left[\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right]\right\} \tag{6.99}$$

$$= \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \underbrace{\mathbb{E}_{z_i}\left[\log}_{\mathbb{E}\log}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right)\right]\right\} \tag{6.100}$$

$$\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \underbrace{\log\left[\mathbb{E}_{z_i}}_{\log\mathbb{E}}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right)\right]\right\} \tag{6.101}$$

$$\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \log\left[\mathbb{E}_{z_i}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i} + \underbrace{e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}}_{\perp\!\!\!\perp z_i}\right)\right]\right\} \tag{6.102}$$

$$\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \log\left[\mathbb{E}_{z_i}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right) + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right]\right\} \tag{6.103}$$

Focus on the red term to compute the expectation:

$$\begin{aligned}
\mathbb{E}_{z_i}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right) &= \frac{1}{2^N}\sum_{i\neq i^*}\mathbb{E}_{z_i}\left(e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right) && \mathbb{E}\text{ linearity}\\
&= \frac{e^{-\frac{N}{2\Delta}}}{2^N}\sum_{i\neq i^*}\mathbb{E}_{z_i}\left(e^{\sqrt{\frac{N}{\Delta}}z_i}\right) && \text{where } z_i \sim \mathcal{N}(0,1)\\
&= \frac{e^{-\frac{N}{2\Delta}}}{2^N}\sum_{i\neq i^*} e^{\frac{N}{2\Delta}} && \text{Lem. A.3, } \kappa = \sqrt{\frac{N}{\Delta}}\\
&= \frac{1}{2^N}\sum_{i\neq i^*} 1\\
&= \frac{1}{2^N}(2^N-1) && \text{only one index is nonzero}\\
&= 1 - \frac{1}{2^N}
\end{aligned}$$

Coming back to Equation 6.103, we can further complete the inequality as:

$$\Phi_N(\Delta) \leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \log\left[\mathbb{E}_{z_i}\left(\frac{1}{2^N}\sum_{i\neq i^*} e^{-\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_i}\right) + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right]\right\} \tag{6.104}$$

$$\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \log\left[1-\frac{1}{2^N} + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}\right]\right\} \tag{6.105}$$

$$\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{ \log\left[\underbrace{1 + e^{\frac{N}{2\Delta}+\sqrt{\frac{N}{\Delta}}z_{i^*}-N\log(2)}}_{=\eta(z_{i^*})}\right]\right\} \tag{6.106}$$

[9] $\mathbb{E}\log \leq \log\mathbb{E}$

Where in the last passage we remove $-\dfrac{1}{2^N} < 0$ since the log function is increasing.
At this point, we recognize that:

$$\eta(z_{i^*}) = 1 + e^{\frac{N}{2\Delta} + \sqrt{\frac{N}{\Delta}} z_{i^*} - N\log(2)} \leq \eta(|z_{i^*}|) \tag{6.107}$$

Where further:

$$\begin{aligned}
\log[|\eta(z_{i^*})|] &= \log[\eta(0)] + (|z_{i^*}| - 0)\frac{\partial}{\partial z}\Big(\log[\eta(z)]\Big)\Big|_{z=z_{i^*}} && \text{Taylor} \quad (6.108)\\
&= \log[\eta(0)] + |z_{i^*}|\frac{\eta'(z)}{\eta(z)}\Big|_{z=z_{i^*}} && (6.109)\\
&\leq \log[\eta(0)] + |z_{i^*}|\max_z\left\{\frac{\eta'(z)}{\eta(z)}\right\} && (6.110)\\
&\leq \log[\eta(0)] + |z_{i^*}|\max_z\left\{\frac{\sqrt{\frac{N}{\Delta}}e^{\frac{N}{2\Delta} + \sqrt{\frac{N}{\Delta}} z_{i^*} - N\log(2)}}{1 + e^{\frac{N}{2\Delta} + \sqrt{\frac{N}{\Delta}} z_{i^*} - N\log(2)}}\right\} && (6.111)\\
&\leq \log[\eta(0)] + |z_{i^*}|\sqrt{\frac{N}{\Delta}}\max_z\left\{\frac{e^{f(z)}}{1 + e^{f(z)}}\right\} && (6.112)\\
&\leq \log[1 + e^{\frac{N}{2\Delta} - N\log(2)}] + |z_{i^*}|\sqrt{\frac{N}{\Delta}} && \text{max is 1} \quad (6.113)
\end{aligned}$$

Coming back to the calculation of the free entropy, we can eventually say that:

$$\begin{aligned}
\Phi_N(\Delta) &\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{\log\Big[\eta(z_{i^*})\Big]\right\} && \text{Eqn. 6.106} \quad (6.114)\\
&\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{\log\Big[\eta(|z_{i^*}|)\Big]\right\} && \text{Eqn. 6.107} \quad (6.115)\\
&\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{\log[1 + e^{\frac{N}{2\Delta} - N\log(2)}] + |z_{i^*}|\sqrt{\frac{N}{\Delta}}\right\} && \text{Eqn. 6.113} \quad (6.116)\\
&\leq \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{\log[1 + e^{\frac{N}{2\Delta} - N\log(2)}]\right\} + \frac{1}{N}\mathbb{E}_{z_{i^*}}\left\{|z_{i^*}|\sqrt{\frac{N}{\Delta}}\right\} && (6.117)\\
&\leq \frac{1}{N}\log[1 + e^{N(\frac{1}{2\Delta} - \log(2))}] + \underbrace{\sqrt{\frac{1}{N\Delta}}\mathbb{E}_{z_{i^*}}\left\{|z_{i^*}|\right\}}_{=o(1)\ as\ N\to\infty} && (6.118)\\
&\leq \frac{1}{N}\log\left[(e^{N\vartheta})\Big(\frac{1}{e^{N\vartheta}} + 1\Big)\right] + o(1) && \vartheta = \frac{1}{2\Delta} - \log(2) \quad (6.119)\\
&\leq \frac{1}{N}\log[e^{N\vartheta}] + \frac{1}{N}\log\left[\frac{1}{e^{N\vartheta}} + 1\right] + o(1) && (6.120)\\
&\leq \vartheta + \underbrace{\frac{1}{N}\frac{1}{e^{N\vartheta}}}_{=o(1)\ as\ N\to\infty,\vartheta\geq 0} + o(1) && \log(1+x) \leq x\forall x \quad (6.121)
\end{aligned}$$

Where, taking the limit we approach $f(\Delta|\vartheta \geq 0)$ from below meaning:

$$\Phi(\Delta) = \lim_{N\to\infty,\theta\geq 0}\left\{\Phi_N(\Delta)\right\} = \lim_{N\to\infty}\left\{\frac{1}{2\Delta} - \log(2) + o(1)\right\} = f(\Delta) \tag{6.122}$$

If $\vartheta < 0$ instead, we get that $\Phi_N(\Delta) \leq 0$, again recovering the second case of the claimed form of $f(\Delta)$. □

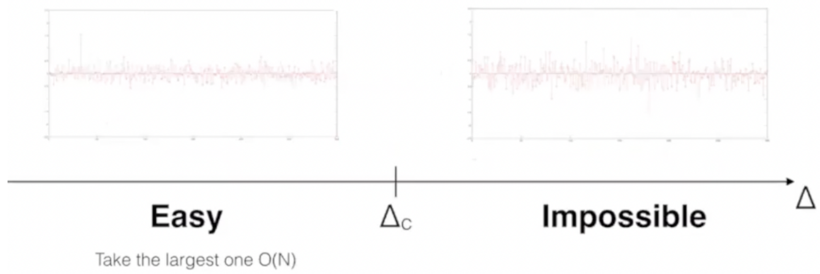


Figure 6.3: Tractability map as Δ increases, $N = 300$

The result of Theorem 6.24 allows us to apply Theorem 6.22 to recover the MMSE for denoising a d dimensional vector:

$$\begin{aligned}
MMSE &= 1 - q && \text{Eqn. 6.92} && (6.123)\\
&= 1 - 2\frac{\partial}{\partial \Delta^{-1}}\Phi(\Delta) && \text{inverse Thm. 6.22} && (6.124)\\
&= \begin{cases} 1 - 2\frac{1}{2} & \Delta \leq \frac{1}{2\log 2} \\ 1 - 0 & \Delta > \frac{1}{2\log 2} \end{cases} && && (6.125)\\
&= \begin{cases} 0 & \Delta \leq \frac{1}{2\log 2} \\ 1 & \Delta > \frac{1}{2\log 2} \end{cases} && && (6.126)
\end{aligned}$$

Where we recognize a **sharp** phase transition from impossible to possible at the critical variance:

$$\Delta_c = \frac{1}{2\log 2} \tag{6.127}$$

In the event in which $\Delta > \Delta_c$ the MMSE is as good as random guessing. On the contrary, if the variance of the noise is low enough, there should be a method to find 0 MMSE (i.e. an algorithm which is correct in expectation). The *tractability* map can be visualized in Figure 6.3

In the original lecture notes, under the condition that $\Delta < \Delta_c$, the authors explain how a very much trivial algorithm can be derived [KZ21b].

6.4 Visualizing an Example and Further References

This final section is devoted to presenting quickly a real example and reporting the paragraph present in the original lecture notes [KZ21b], as to give further directions to delve into the topic.

Example 6.25 (Denoising Graphically)**.** Consider a sparse vector of size $d = 300$. For various examples of Δ, the situation is different. In Figures 6.4, 6.5 we can see that the problem is almost impossible, while getting closer to the critical value of $\Delta_c = 0.08$, it becomes more and more *trivial.*

After having surpassed the phase transition at Figure 6.6, the easiest algorithm has a guaranteed $MMSE$ of 0 as in Figure 6.7.

Below, the Bibliography paragraph is reported for the sake of completeness

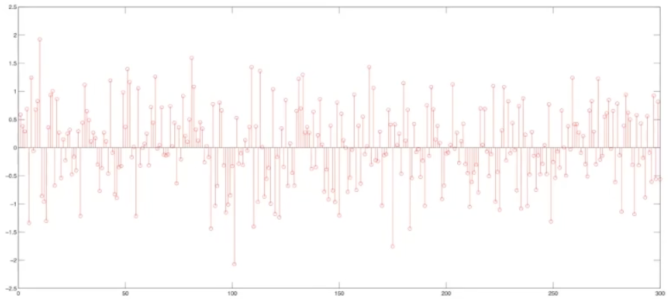


Figure 6.4: Noisy vector $\Delta = 0.5, N = 300$

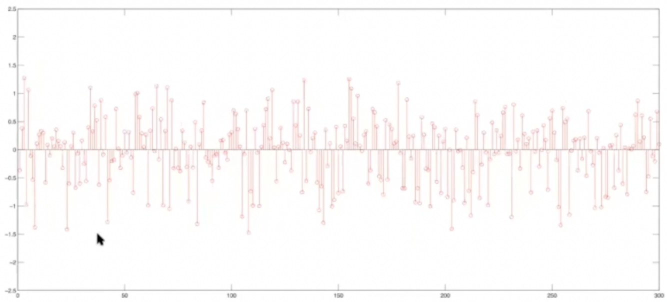


Figure 6.5: Noisy vector $\Delta = 0.3, N = 300$

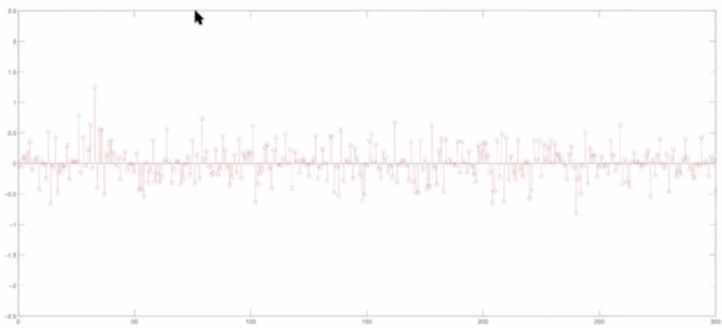


Figure 6.6: Noisy vector $\Delta = 0.08, N = 300$

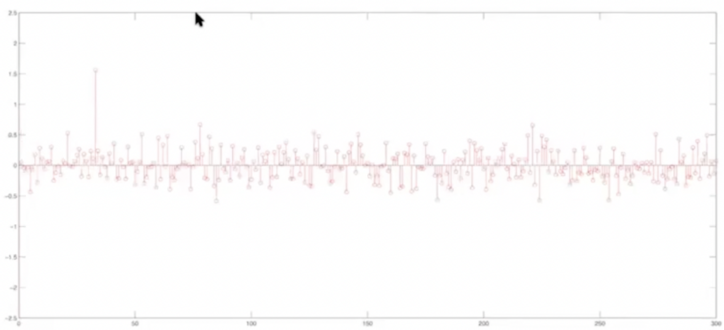


Figure 6.7: Noisy vector $\Delta = 0.05, N = 300$

The legacy of the Bayes theorem, and the fundamental role of Laplace in the invention of "inverse probabilities" is well discussed in [McG11]. Bayesian estimation is a fundamental field at the frontier between information theory and statistics, and is discussed in many references such as [CT06]. The I-MMSE theorem was introduced by [GSV04]. Nishimori symmetries were introduced in physics in [Nis80] and soon realized to have deep connection to information theory [Nis93] and Bayesian inference [Iba99]. The model of denoising a sparse vector was discussed in Donoho et al. (1998). This problem has deep relation to Shannon's Random codes [Sha48] and the Random energy model in statistical physics [Der81].

Chapter 7

Matrix Factorization

We now move to a more complicated setting in which the tools studied in Chapter 2 can be applied. This Chapter, we will specifically implement the Replica Method of Section 2.4.2.

7.1 Clustering for Variable *Alignment*

A common problem in data analysis is finding patterns in data. In case of correlated features, understanding *clusters* of variables that move together is of pivotal importance. An example of satisfactory result is Figure 7.1.

However, making sense of a correlation matrix by grouping variables is not an easy task, and plotting patterns ends up being not straightforward. While a desired result would be that of Figure 7.2, we recognize that in reality data does not present itself with this visualization advantage.

There are two main issues in how information actually presents itself:

- **noise:** correlation is a statistical phenomenon, and is different at different realizations. See Figure 7.3 for an example of noisy
- **absence of order and association:** correlated features are not adjacent in the columns and rows of the correlation matrix in principle. Reality has no implied order of factors, nor it naturally presents which features should be grouped under a distinctive color. Eventually, data appears most of the times as in Figure 7.4, where permutation and noise make grouping non trivial.

To understand the nature of this task, we will recover an **idealized version** of it.

Definition 7.1 (Spike-Wigner model)**.** Consider a set of variables

$$\mathbf{x}^* = \{x_i^*\}_{i=1}^N \in \mathbb{R}^N \quad | \quad x_i^* \overset{i.i.d.}{\sim} \mathbb{P}_X(x) \tag{7.1}$$

Such that its correlation matrix is perturbed by Gaussian White noise as:

$$\mathbb{R}^N \times \mathbb{R}^N \ni Y = \frac{\mathbf{x}^*(\mathbf{x}^*)^T}{\sqrt{N}} + \sqrt{\Delta}\boldsymbol{\xi} \mid \xi_{ij} \sim \mathcal{N}(0,1),\ \xi_{ij} = \xi_{ji} \forall i < j \tag{7.2}$$

If the matrix Y was only Gaussian noise it would be a wigner matrix as in Definition B.2, but we are adding a rank one[1] matrix, which has one eigenvalue (i.e. a spike), hence the name Spike-Wigner.

[1]Notice the $\frac{1}{\sqrt{N}}$ normalization

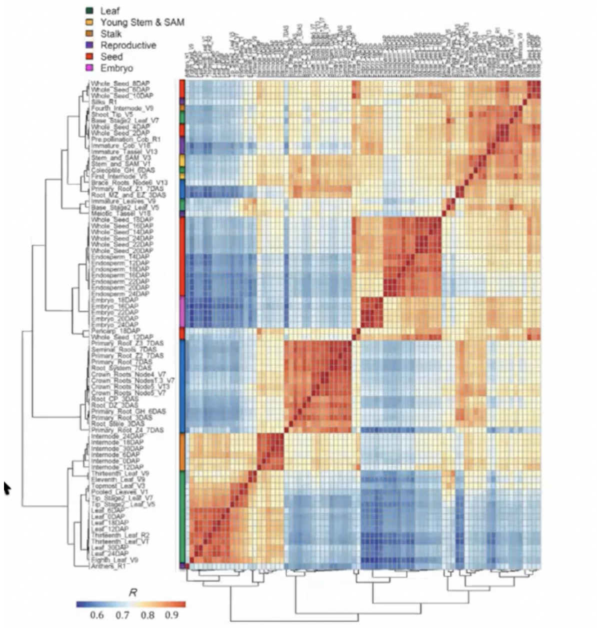


Figure 7.1: Gene Expression of corn tissues

Figure 7.2: Ordered correlation matrix

Figure 7.3: Ordered noisy correlation matrix

Figure 7.4: Permuted noisy correlation matrix

From Definition 7.1, we draw our starting point, with the task of recovering the patterns in the **original** correlation matrix $\mathbf{x}^*(\mathbf{x}^*)^T$

Definition 7.2 (Signal to noise ratio, SNR λ)**.** In Statistics and Signal Processing it is more natural to choose as variance parameter:

$$\lambda := \Delta^{-1} \tag{7.3}$$

Readjusting Equation 7.2 to the more intuitive form:

$$Y = \sqrt{\frac{\lambda}{N}}\mathbf{x}^*(\mathbf{x}^*)^T + \boldsymbol{\xi} \tag{7.4}$$

With equivalence up to rescaling.

Observation 7.3 (SNR role)**.** *It is clear that:*

$$\begin{aligned} \lambda \to 0 &\implies \textit{only noise} \\ \lambda \to \infty &\implies \textit{strong signal} \end{aligned}$$

The question we will answer in this Chapter can be summarized as follows:

What is the best possible error on inferring $\boldsymbol{x}^*$ *upon knowledge of* $Y, \mathbb{P}_X, \mathbb{P}_\Xi$?

Assumption 7.4 (Choice of error function)**.** Throughout the discussion, we will impose as error function the MMSE, with the formalism of Definition 6.8, found in Example 6.9. In this case, it will explicitly depend on the SNR with the form $MMSE(\lambda)$.

In the next chapter, we will formalize how to find a good estimator $\widehat{x}$ in practice as an algorithm. For the moment, we focus on the error function discussion.

For such a measure of correctness, the best overall is obviously zero, and the worst is collapsing to random guessing performance. When random guessing, the error will fall inside the interval $MMSE_{random} \in [1, 2]$ **TODO check**.

7.2 Bayes Formalism and Replica Method Approach

Thanks to the discussion carried out in Chapter 6, we already have all the tools to draw meaningful conclusions with:

- bayesian probability
- statistical physics interpretation
- free entropy
- thermodynamic limit $N \to \infty$

Using Bayes Rule we have:

$$\mathbb{P}_{X|Y}(\mathbf{x}|Y) = \frac{\mathbb{P}_{Y|X}(Y|\mathbf{x})\mathbb{P}_X(\mathbf{x})}{\mathbb{P}_Y(Y)} \tag{7.5}$$

$$\propto \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_{ij} - \sqrt{\frac{\lambda}{N}} x_i x_j)^2} \qquad Y_{ij}|\mathbf{x} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}} x_i x_j, 1\right) \ \textbf{TOD} \tag{7.6}$$

$$\propto \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_{ij}^2 - 2\sqrt{\frac{\lambda}{N}} x_i x_j y_{ij} + \frac{\lambda}{N} x_i^2 x_j^2)} \qquad \text{expand square} \tag{7.7}$$

$$\propto \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} e^{\frac{1}{2}(+2\sqrt{\frac{\lambda}{N}} x_i x_j y_{ij} - \frac{\lambda}{N} x_i^2 x_j^2)} \qquad \text{remove constant w.r.t } x \text{ factors} \tag{7.8}$$

$$\implies \mathbb{P}_{X|Y}(\mathbf{x}|Y) = \frac{1}{Z(y)} \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} e^{-\frac{\lambda}{2N} x_i^2 x_j^2 + \sqrt{\frac{\lambda}{N}} y_{ij} x_i x_j} \qquad \text{letting } \beta = 1 \tag{7.9}$$

$$= \frac{1}{Z(y)} \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} e^{-\frac{\lambda}{2N} x_i^2 x_j^2 + \sqrt{\frac{\lambda}{N}} (\sqrt{\frac{\lambda}{N}} x_i^* x_j^* + \xi_{ij}) x_i x_j} \tag{7.10}$$

$$= \frac{1}{Z(y)} \prod_i \mathbb{P}_X(x_i) \prod_{i \leq j} e^{-\frac{\lambda}{2N} x_i^2 x_j^2 + \frac{\lambda}{N} x_i x_j x_i^* x_j^* + \sqrt{\frac{\lambda}{N}} \xi_{ij} x_i x_j} \tag{7.11}$$

Where the MMSE, from Proposition 6.11 will be of the form:

$$\mathbb{R}^N \ni \widehat{x}_{MMSE}(Y) = \langle x \rangle = \begin{bmatrix} \widehat{x}_{MMSE,1} \\ \cdots \\ \widehat{x}_{MMSE,N} \end{bmatrix} = \begin{bmatrix} \int d\mathbf{x} \mathbb{P}_{X|Y}(\mathbf{x}|Y) x_1 \\ \cdots \\ \int d\mathbf{x} \mathbb{P}_{X|Y}(\mathbf{x}|Y) x_n \end{bmatrix} \tag{7.12}$$

The discussion carried out in Section 2.3 and Theorem 6.22 suggest inspecting the free entropy at the thermodynamic limit:

$$\Phi(\lambda) = \lim_{N \to \infty} \frac{\Phi_N(\lambda)}{N} = \lim_{N \to \infty} \frac{\mathbb{E}_Y[\log[Z(Y)]}{N} = \lim_{N \to \infty} \frac{\mathbb{E}_{\mathbf{x}^*, \boldsymbol{\xi}} \log[Z(Y)]}{N} \tag{7.13}$$

However, to evaluate this expression, we would need to compute an integral over infinite dimensions as $N \to \infty$. While this is not feasible, we can exploit the replica method from Section 2.4.2 and start from the expectation of $Z^n : n \to 0$. To do so, we would need to apply Nishimori identity again (Thm. 6.19), where we state that $q = m$ and replica symmetry holds. Another important assumption we make is concentration, which in this case is equivalent to asserting that the magnetization variance is nullified at the thermodynamic limit:

$$\mathbb{E}_{\mathbf{x}^*, \boldsymbol{\xi}} \left[\left\langle \left(\sum_i \frac{x_i x_i^*}{N} - m \right)^2 \right\rangle \right] \underset{N \to \infty}{\to} 0 \tag{7.14}$$

Which by Nishimori identity holds for two copies and their overlaps as well:

$$\mathbb{E}_{\mathbf{x}^*, \boldsymbol{\xi}} \left[\left\langle \left(\sum_i \frac{x_i^{(\alpha)} x_i^{(\beta)}}{N} - q \right)^2 \right\rangle \right] \underset{N \to \infty}{\to} 0 \tag{7.15}$$

This fact holds since if we imagine the generating process detailed in Equation 7.4 paired with a problem of the easier form:

$$\widetilde{\mathbf{x}} = \sqrt{\lambda'} \mathbf{x}^* + \boldsymbol{\xi} \tag{7.16}$$

We can see that this augmented problem is solved by applying Theorem 6.22 for one observation of vector with noise as:

$$\frac{\partial \Phi}{\partial \lambda} = \frac{\partial (\lim_{N\to\infty} \frac{\log[Z]}{N})}{\partial \lambda} = \frac{1}{2}q = \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}[\langle x\rangle^2] = \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\left[\frac{1}{Z(Y)}\int d\mathbf{x}\mathbb{P}_X(\mathbf{x})\prod_{i\leq j} e^{-\frac{\lambda}{2N}x_i^2x_j^2+\frac{\lambda}{N}x_ix_jx_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_ix_j}x\right]^2\right\} \tag{7.17}$$

Which in turn means that the second derivative, or the curvature easily collapses to being **TODO maybe rewatch video**:

$$\frac{\partial^2 \Phi}{\partial \lambda^2} = \frac{N}{4}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\left\langle\left(\sum_i \frac{x_i^{(\alpha)}x_i^{(\beta)}}{N} - q\right)^2\right\rangle\right] \tag{7.18}$$

$$= \frac{N}{4}Var[q] \tag{7.19}$$

Which implies that:

$$\frac{1}{4}\int_{\lambda}^{\lambda'} NVar[q]d\lambda = \int_{\lambda}^{\lambda'} \frac{\partial^2\Phi}{\partial\lambda^2} = \frac{1}{2}\Big[q(\lambda_2) - q(\lambda_1)\Big] \tag{7.20}$$

And, since q is a constant number, it is bounded and we can say that

$$\int Var[q] \leq \frac{c}{N} \implies Var[q] \underset{N\to\infty}{\to} 0 \quad for\ a.e.\ \lambda \in (\lambda_1, \lambda_2) \tag{7.21}$$

This concept will be precisely exploited in the replica computations **TODO check exacly where and make it a theorem**.

Coming back to out original computation, we evaluate the expectation over the noise of the partition function to the power n:

$$\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[Z^n\right] = \mathbb{E}_{\mathbf{x}^*,\xi}\left\{\left[\int d\mathbf{x}\mathbb{P}_X(\mathbf{x})\prod_{i\leq j} e^{-\frac{\lambda}{2N}x_i^2x_j^2+\frac{\lambda}{N}x_ix_jx_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_ix_j}\right]^n\right\} \tag{7.22}$$

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\prod_{\alpha=1}^{n}\int d\mathbf{x}^{(\alpha)}\prod_{i=1}^{N}\mathbb{P}_X(\mathbf{x}_i^{(\alpha)})\right. \tag{7.23}$$

$$\left.\times\prod_{i\leq j} exp\left\{-\frac{\lambda}{2N}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2+\frac{\lambda}{N}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_i^{(\alpha)}x_j^{(\alpha)}\right\}\right\}$$

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\int\prod_{\alpha,i}dx_i^{(\alpha)}\mathbb{P}_X(x_i^{(\alpha)})\right. \tag{7.24}$$

$$\left.\times\, exp\left\{\sum_{i\leq j}-\frac{\lambda}{2N}\sum_{\alpha}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2+\frac{\lambda}{N}\sum_{\alpha}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}\sum_{\alpha}x_i^{(\alpha)}x_j^{(\alpha)}\right\}\right\}$$

exchange pro

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\int\prod_{\alpha,i}dx_i^{(\alpha)}\mathbb{P}_X(x_i^{(\alpha)})exp\{\sum_{i\leq j}-\frac{\lambda}{2N}\sum_{\alpha}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2+\frac{\lambda}{N}\sum_{\alpha}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*\}\right. \tag{7.25}$$

$$\left.\times\prod_{i\leq j}\underbrace{\mathbb{E}_{\xi_{ij}}\left[exp\left\{\xi_{ij}\left(\sqrt{\frac{\lambda}{N}}\sum_{\alpha}x_i^{(\alpha)}x_j^{(\alpha)}\right)\right\}\right]}_{Lem\ A.3,\ for\ \xi_{ij}\sim\mathcal{N}(0,1)}\right\}$$

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\int\prod_{\alpha,i}dx_i^{(\alpha)}\mathbb{P}_X(x_i^{(\alpha)})exp\{\sum_{i\leq j}-\frac{\lambda}{2N}\sum_{\alpha}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2+\frac{\lambda}{N}\sum_{\alpha}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*\}\right. \tag{7.26}$$

$$\left.\times\prod_{i\leq j}exp\left\{\frac{\lambda}{2N}\sum_{\alpha,\beta}x_i^{(\alpha)}x_j^{(\alpha)}x_i^{(\beta)}x_j^{(\beta)}\right\}\right\}$$

copies indexe

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\int\prod_{\alpha,i}dx_i^{(\alpha)}\mathbb{P}_X(x_i^{(\alpha)})exp\{\sum_{i\leq j}\sum_{\alpha}-\frac{\lambda}{2N}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2+\sum_{i\leq j}\sum_{\alpha}\frac{\lambda}{N}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*\}\right. \tag{7.27}$$

$$\left.\times\, exp\left\{\frac{\lambda}{2N}\sum_{i\leq j}\sum_{\alpha,\beta}x_i^{(\alpha)}x_j^{(\alpha)}x_i^{(\beta)}x_j^{(\beta)}\right\}\right\}$$

Notice now that for the *red*, *blue*, and *orange* terms we can apply the identity:

$$\sum_{i\leq j}\frac{a_i}{N}\frac{a_j}{N} = \frac{1}{2}\left(\sum_i\frac{a_i}{N}\right)^2+\underbrace{\frac{1}{2}\sum_i\left(\frac{a_i^2}{N^2}\right)^2}_{\in O(N^{-1})}\underset{N\to\infty}{\approx}\frac{1}{2}\left(\sum_i\frac{a_i}{N}\right)^2 \tag{7.28}$$

Which transferred to our three terms, by adjusting the $\frac{1}{N}$ factor inside the sum, makes them:

$$\sum_{i\leq j}\sum_{\alpha}-\frac{\lambda}{2N}(x_i^{(\alpha)})^2(x_j^{(\alpha)})^2 = -\frac{\lambda N}{2}\sum_{\alpha}\frac{1}{2}\left(\sum_i\frac{(x_i^{(\alpha)})^2}{N}\right)^2 = -\frac{\lambda N}{4}\sum_{\alpha}\left(\sum_i\frac{(x_i^{(\alpha)})^2}{N}\right)^2$$

$$\sum_{i\leq j}\sum_{\alpha}\frac{\lambda}{N}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^* = \lambda N\sum_{\alpha}\frac{1}{2}\sum_i\left(\frac{x_i^{(\alpha)}x_i^*}{N}\right)^2 = \frac{\lambda N}{2}\sum_{\alpha}\sum_i\left(\frac{x_i^{(\alpha)}x_i^*}{N}\right)^2$$

$$\frac{\lambda}{2N}\sum_{i\leq j}\sum_{\alpha,\beta}x_i^{(\alpha)}x_j^{(\alpha)}x_i^{(\beta)}x_j^{(\beta)} = \frac{\lambda N}{2}\sum_{\alpha,\beta}\frac{1}{2}\sum_i\left(\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2 = \frac{\lambda N}{4}\sum_{\alpha,\beta}\sum_i\left(\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2$$

Where the first and the third have the same coefficient in front and obey the identity **TODO maybe state in its general form**:

$$2\sum_{\alpha<\beta}\sum_i\left(\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2=\sum_{\alpha,\beta}\sum_i\left(\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2-\sum_\alpha\left(\sum_i\frac{(x_i^{(\alpha)})^2}{N}\right)^2$$

Going back to Equation 7.27, we get that:

$$\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[Z^n\right]=\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left\{\int\prod_{\alpha,i}dx_i^{(\alpha)}\mathbb{P}_X(x_i^{(\alpha)})exp\left\{\underbrace{\frac{\lambda N}{2}\sum_\alpha\left(\sum_i\frac{x_i^{(\alpha)}x_i^*}{N}\right)^2}_{=blue}+\underbrace{\frac{\lambda N}{2}\sum_{\alpha<\beta}\left(\sum_i\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2}_{=orange+red}\right\}\right\} \tag{7.29}$$

$$=TODOfinish \tag{7.30}$$

Eventually, the free entropy under the replica symmetry ansatz reads:

$$\Phi_{RS}(m)=-\frac{\lambda}{4}m^2+\mathbb{E}_{x^*,z}\left[\log\left(\int\mathbb{P}_X(x)dx\; e^{-\frac{\lambda m}{2}x^2+(\lambda m x^*+\sqrt{\lambda m}z)x}\right)\right] \tag{7.31}$$

And the self consistent equation obtained by taking the maximum is:

$$\left.\frac{\partial\Phi_{RS}(m)}{\partial m}\right|_{m^*}=0\implies m^*=\mathbb{E}_{x^*,z}\left[\frac{\int\mathbb{P}_X(x)dx\; e^{-\frac{\lambda m}{2}x^2+(\lambda m x^*+\sqrt{\lambda m}z)x}\; x}{\int\mathbb{P}_X(x)dx\; e^{-\frac{\lambda m}{2}x^2+(\lambda m x^*+\sqrt{\lambda m}z)x}}\right] \tag{7.32}$$

Requires the evaluation of three integrals, a generally feasible task, either analytically or by numerical integration.

Remark. Remember that we started with ∞ integrals! Even if the formula looks ugly, it is far easier than what it was before.

TODO HW 7.1, 7.2

Observation 7.5 (On the maximizer m^*). *An interesting fact is that the maximizer is equal to the overlap:*

$$m^*=q=\mathbb{E}\left[\frac{\sum_i\langle x_i\rangle x_i^*}{N}\right]=\mathbb{E}\left[\frac{\sum_i\langle x_i\rangle^2}{N}\right] \tag{7.33}$$

*Which also means that the MMSE is **TODO maybe recall equation**:*

$$MMSE=\mathbb{E}[\langle x^2\rangle]-m^* \tag{7.34}$$

7.3 Rigorous Matrix Factorization by the Interpolation Method

Following the approach of Section 3.2 with the method designed by Guerra **TODO cite**, we aim to prove rigourously that Equation 7.32 is correct.

To do so, we consider the easier problem, labeled by the letter **A**, as a parallelized denoising[2] problem of the form:

$$\widetilde{\mathbf{y}} = \sqrt{\lambda m}\mathbf{x}^* + \boldsymbol{\omega} \tag{7.35}$$

$$\implies \mathbb{P}_{X|Y}(\mathbf{x}|\mathbf{y}) \propto exp\Big\{\sum_i \log\Big[\mathbb{P}_X(x_i)\Big] + \sum_i \Big[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\Big]\Big\} \tag{7.36}$$

$$\mathscr{H}_{\mathbf{A}}(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}; m) = \sum_i -\log\Big[\mathbb{P}_X(x_i)\Big] - \sum_i \Big[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\Big] \tag{7.37}$$

$$\beta = 1 \tag{7.38}$$

For such a formulation, we have that the partition function and the free entropy take the form:

$$Z_{\mathbf{A}}(\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m) = \int \prod_i dx_i e^{-\beta\mathscr{H}_{\mathbf{A}}} \tag{7.39}$$

$$= \int \prod_i dx_i exp\Big\{\sum_i \log\Big[\mathbb{P}_X(x_i)\Big] + \sum_i \Big[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\Big]\Big\} \tag{7.40}$$

$$= \int \prod_i dx_i \mathbb{P}_X(x_i) exp\Big\{\sum_i \Big[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\Big]\Big\} \tag{7.41}$$

$$\implies \Phi_{\mathbf{A},N} = \frac{1}{N}\sum_i \mathbb{E}_{x_i^*,\omega_i}\Big[\log\Big\{\int dx_i \mathbb{P}_X(x_i) exp\Big\{\sum_i \Big[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\Big]\Big\} \tag{7.42}$$

$$= \mathbb{E}_{x_1^*,\omega_1}\Big[\log\Big\{\int dx_1 \mathbb{P}_X(x_1) exp\Big\{\sum_1 \Big[-\frac{\lambda m}{2}x_1^2 + \lambda m x_1^* x_i + \sqrt{\lambda m}\omega_1 x_1\Big]\Big\} \tag{7.43}$$

$$= \Phi_{denoising}(\lambda m) \tag{7.44}$$

Where we have exploited the parallel formalism (i.e. independence) and recognized that the final result is just the free entropy from Chapter 6 with a $\lambda m = \Delta^- 1m$ argument inside.

Additionally, we recall the focus points of the target problem, labelled by the letter **B**, which is stated in Equation 7.4, has posterior described in Equation 7.8 and Hamiltonian:

$$\mathscr{H}_{\mathbf{B}}(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\xi}) = -\sum_i \log(\mathbb{P}_X(x_i)) - \sum_{i \leq j} \frac{\lambda}{2N}x_i^2 x_j^2 + \frac{\lambda}{N}x_i x_j x_i^* x_j^* + \sqrt{\frac{\lambda}{N}}\xi_{ij}x_i x_j \tag{7.45}$$

Of which we specifically look for its partition function $Z_{\mathbf{B}}(\lambda, \mathbf{x}^*, \boldsymbol{\xi})$.

[2]For more see Chaper 6 adapting it to a vector in $\mathbb{R}^N$

We eventually implement a time interpolation of Equations 7.37, 7.45 that reaches at $t = 1$ the target form for our partition function $Z_{\mathbf{B}}$:

$$\mathscr{H}_t(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m) := \mathscr{H}_{\mathbf{A}}(\mathbf{x}, (1-t)\lambda, \mathbf{x}^*, \boldsymbol{\xi}; m) + \mathscr{H}_{\mathbf{B}}(\mathbf{x}, t\lambda, \mathbf{x}^*, \boldsymbol{\xi}) \tag{7.46}$$

$$= -\sum_i \left[-\frac{(1-t)\lambda m}{2} x_i^2 + (1-t)\lambda m x_i^* x_i + \sqrt{(1-t)\lambda m}\,\omega_i x_i \right] \tag{7.47}$$

$$- \sum_{i \le j} \frac{t\lambda}{2N} x_i^2 x_j^2 + \frac{t\lambda}{N} x_i x_j x_i^* x_j^* + \sqrt{\frac{t\lambda}{N}} \xi_{ij} x_i x_j$$

Notice that we model the strength of the SNR in each of the two energy configurations, where however λ is shared across. Basically, we choose depending on time which model to completely make silent.
Denoting the parameters by the shorthand $\boldsymbol{\vartheta} = \{\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}\}$ we precompute some quantities of interest such as:

$$\frac{\partial}{\partial t}\mathscr{H}_t(\mathbf{x}, \boldsymbol{\vartheta}; m) = -\sum_i \left[\frac{\lambda m}{2} x_i^2 - \lambda m x_i x_i^* - \frac{\sqrt{\lambda m}}{2\sqrt{1-t}} \omega_i x_i \right] - \sum_{i \le j} \frac{\lambda}{2N} x_i^2 x_j^2 + \frac{\lambda}{N} x_i x_j x_i^* x_j^* + \frac{\sqrt{\frac{\lambda}{N}}}{2\sqrt{t}} \xi_{ij} x_i x_j \tag{7.48}$$

$$\frac{\partial}{\partial t}\frac{\log[Z_t(\boldsymbol{\vartheta}; m)]}{N} = \frac{1}{N}\frac{1}{Z_t(\boldsymbol{\vartheta}; m)}\frac{\partial Z_t(\boldsymbol{\vartheta}; m)}{\partial t} \tag{7.49}$$

$$= \frac{1}{N}\frac{1}{Z_t(\boldsymbol{\vartheta}; m)}\frac{\partial}{\partial t}\int d\mathbf{x} e^{-\mathscr{H}_t(\mathbf{x}, \boldsymbol{\vartheta}; m)} \tag{7.50}$$

$$= \frac{1}{N}\int d\mathbf{x} \underbrace{\frac{e^{-\mathscr{H}_t(\mathbf{x}, \boldsymbol{\vartheta}; m)}}{Z_t(\boldsymbol{\vartheta}; m)}}_{= Boltz\ weight\ t, \boldsymbol{\vartheta}; m} \frac{\partial}{\partial t}\Big(-\mathscr{H}_t(\mathbf{x}, \boldsymbol{\vartheta}; m)\Big) \tag{7.51}$$

$$= -\frac{1}{N}\left\langle \frac{\partial}{\partial t}\mathscr{H}_t(\mathbf{x}, \boldsymbol{\vartheta}; m)\right\rangle_{t, \boldsymbol{\vartheta}; m} \tag{7.52}$$

$$= -\frac{1}{N}\left\langle -\sum_i \left[\frac{\lambda m}{2} x_i^2 - \lambda m x_i x_i^* - \frac{\sqrt{\lambda m}}{2\sqrt{1-t}} \omega_i x_i \right] \right. \tag{7.53}$$

$$\left. - \sum_{i \le j} \frac{\lambda}{2N} x_i^2 x_j^2 + \frac{\lambda}{N} x_i x_j x_i^* x_j^* + \frac{\sqrt{\frac{\lambda}{N}}}{2\sqrt{t}} \xi_{ij} x_i x_j \right\rangle_{t, \boldsymbol{\vartheta}; m}$$

$$= -\frac{1}{N}\left\{ -\sum_i \left[\frac{\lambda m}{2}\langle x_i^2\rangle_{t, \boldsymbol{\vartheta}; m} - \lambda m \langle x_i x_i^*\rangle_{t, \boldsymbol{\vartheta}; m} - \frac{\sqrt{\lambda m}}{2\sqrt{1-t}} \omega_i \langle x_i\rangle_{t, \boldsymbol{\vartheta}; m} \right] \right. \tag{7.54}$$

$$\left. - \sum_{i \le j} \frac{\lambda}{2N} \langle x_i^2 x_j^2\rangle_{t, \boldsymbol{\vartheta}; m} + \frac{\lambda}{N} \langle x_i x_j x_i^* x_j^*\rangle_{t, \boldsymbol{\vartheta}; m} + \frac{\sqrt{\frac{\lambda}{N}}}{2\sqrt{t}} \xi_{ij} \langle x_i x_j\rangle_{t, \boldsymbol{\vartheta}; m} \right\}$$

$$= -\frac{1}{N}\left\{ \frac{\lambda}{N}\sum_{i \le j}\left[\frac{\langle x_i^2 x_j^2\rangle_{t, \boldsymbol{\vartheta}; m}}{2} - \langle x_i^* x_j^* x_i x_j\rangle_{t, \boldsymbol{\vartheta}; m}\right] - \lambda m \sum_i \left[\frac{\langle x_i^2\rangle_{t, \boldsymbol{\vartheta}; m}}{2} - \langle x_i^* x_i\rangle_{t, \boldsymbol{\vartheta}; m} \right] \right. \tag{7.55}$$

$$\left. - \frac{\sqrt{\frac{\lambda}{N}}}{2\sqrt{t}} \sum_{i \le j} \xi_{ij} \langle x_i x_j\rangle_{t, \boldsymbol{\vartheta}; m} + \frac{\sqrt{\lambda m}}{2\sqrt{1-t}} \sum_i \omega_i \langle x - i\rangle_{t, \boldsymbol{\vartheta}; m} \right\}$$

As argued before, by simple integration one can conclude that the free entropy for the matrix factorization problem, denoted as $\Phi_{MF}(\lambda)$ takes the form:

$$\Phi_{MF}(\lambda) = \lim_{N\to\infty}\left\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\frac{\log(Z_B(\lambda,\mathbf{x}^*,\boldsymbol{\xi})}{N}\right]\right\} \tag{7.56}$$

$$= \lim_{N\to\infty}\left\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\frac{\log(Z_A(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m)}{N} + \int_0^1 d\tau\,\frac{\partial}{\partial t}\frac{\log(Z_t(\boldsymbol{\vartheta};m)}{N}\bigg|_{t=\tau}\right]\right\} \tag{7.57}$$

Tedious **TODO write them** calculations eventually lead to an upper bound result:

$$\Phi_{MF}(\lambda) \geq \Phi_{denoising}(\lambda m) - \frac{\lambda m^2}{4}\forall m \implies \Phi_{MF}(\lambda) = \max_m\left\{\Phi_{denoising}(\lambda m) - \frac{\lambda m^2}{4}\right\} \tag{7.58}$$

Observations to put: we started with overlaps, now only $\langle x_i^2\rangle, \langle x_i^2 x_j^2\rangle$. and we will get rid of them.
$\langle x_i\rangle z_i$ and $\langle x_i x_j\rangle \xi_{ij}$ are hard to interpret, we solve them using stein's lemma.

With a lower bound, we have one direction of the proof of equality. We now move to the creation of an upper bound by fixed magnetization.
Readjusting problem **A** into a fixed magnetization problem, we again find that by interpolating **TODO continue**:

$$\Phi_{MF}^{fixed}(\lambda;M) \leq \Phi_{denoising}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(m-M)^2 - \frac{\lambda m^2}{2} \qquad \forall m,q \tag{7.59}$$

$$\implies \Phi_{MF}^{fixed}(\lambda;M) \leq \min_{m,q}\left\{\Phi_{denoising}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(m-M)^2 - \frac{\lambda m^2}{2}\right\} \tag{7.60}$$

To exploit these bounds, observe that by the Laplace transform in the thermodynamic limit, the configurations of magnetization M will dominate the landscape **TODO check explain better**:

$$\Phi_{MF}(\lambda) = \max_M\left\{\Phi_{MF}^{fixed}(\lambda;M)\right\} \tag{7.61}$$

$$\leq \max_M\left\{\min_{m,q}\left[\Phi_{denoising}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(m-M)^2 - \frac{\lambda m^2}{2}\right]\right\} \tag{7.62}$$

$$\leq \max_M\left\{\Phi_{denoising}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(m-M)^2 - \frac{\lambda m^2}{2}\bigg|_{m=M,q=M}\right\} \tag{7.63}$$

$$\leq \max_M\left\{\Phi_{denoising}(\lambda m) - \frac{\lambda M^2}{4}\right\} \tag{7.64}$$

Combining Equation 7.58 and 7.64 we reach the same conclusion of the replica method (Equation 7.31).

Observation 7.6 (Lower bound and Replica symmetry)**.** *The lower bound obtained in Equation 7.58 is part of the proof of the large deviation* ***TODO check which one****, but is not true under general conditions. Indeed, if the problem is not convex, outside of the maximum it does not hold and we get* ***replica symmetry breaking****, a concept that will be explored in later Chapters.*

In the next chapter, we will explore a solution by the cavity method by assuming that concentration principle holds.

Below, the Bibliography paragraph is reported for the sake of completeness

Nishimori demonstrated that his symmetry implied replica-symmetry in Nishimori (1980). The modern approach in terms of perturbations is discussed in Korada and Macris (2009); Abbe and Montanari (2013); Coja-Oghlan et al. (2018). The spikedWigner model has been studied in great detail over the last decades, and has been the topics of many fundamental papers Johnstone (2001); Baik et al. (2005). The replica approach to this problem is reviewed in details in Lesieur et al. (2017). The results was first proved in Barbier et al. (2016). We presented here the alternative later proof of El Alaoui and Krzakala (2018). Thanks to an universality theorem Krzakala et al. (2016), many problems can be reduced to variants of such low-rank factorization problems and generic formula have been proven for these Lelarge and Miolane (2019); Miolane (2017); Barbier and Macris (2019). Extensions can also be made for almost arbitrary priors, including neural networks generating models Aubin et al. (2020).

Appendix A

Mathematical Facts

The purpose of this collection is providing the *less* experienced reader with noteworthy mathematical results. To avoid interrupting the discussion of the reasoning, they are reported here for coherence.

The starting triplet of results is directly derivable from the first, or *indirectly*, using the fact that a normal density sums to one, and reworking the coefficients.

Fact A.1 (Easy gaussian integrals)**.** It holds that:

1. $$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \tag{A.1}$$

2. $$a > 0 \implies \int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \tag{A.2}$$

3. $$\int_{-\infty}^{\infty} e^{-ax^2+bx} dx = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}} \tag{A.3}$$

Proof. **(Claim** 1) Let $I = \int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} e^{-y^2} dy$ since the summand is a *dummy index*. Then:

$$\begin{aligned}
I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) && && \text{(A.4)}\\
&= \int_{-\infty}^{\infty} e^{-x^2} \left(\int_{-\infty}^{\infty} e^{-y^2}\right) dydx && && \text{(A.5)}\\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\right) dydx && \text{bring inside } x \text{ argument as } x \perp\!\!\!\perp y && \text{(A.6)}\\
&= \int_{0}^{2\pi} \int_{0}^{\infty} e^{-r^2} r\, drd\vartheta && \text{change to polar coordinates } (\star) && \text{(A.7)}\\
&= 2\pi \int_{-\infty}^{0} \frac{1}{2} e^{s}\, ds && \text{substituting } s = -r^2,\ ds = -2rdr && \text{(A.8)}\\
&= \pi \int_{-\infty}^{0} e^{s}\, ds && && \text{(A.9)}\\
&= \pi(e^0 - e^{-\infty}) && && \text{(A.10)}\\
&= \pi \implies \sqrt{I^2} = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} && && \text{(A.11)}
\end{aligned}$$

Warning: we have **overlooked** some improper integrals to give a sketch of the proof, a more formal treatment is proposed in the Wikipedia page *"Gaussian Integral"*[1].
Where in ($\star$) we mean:

$$\begin{cases} r = \sqrt{x^2 + y^2} \\ \vartheta = \tan^{-1}(\frac{y}{x}) \\ |J| = r & \textit{Jacobian determinant} \end{cases} \tag{A.12}$$

(Claim 2**)** Rework the integral as follows:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \int_{-\infty}^{\infty} e^{-u^2} \frac{1}{\sqrt{a}}\, du \qquad \text{substitute } u = \sqrt{a}x,\ du = \sqrt{a}dx \tag{A.13}$$

$$= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-u^2}\, du \tag{A.14}$$

$$= \frac{1}{\sqrt{a}} \sqrt{\pi} \qquad \text{Eqn. A.1} \tag{A.15}$$

(Claim 3**)** we prove this in Lemma A.3 which is equivalent. □

Fact A.2 (Easy identity)**.**

$$\mathbf{x} \sim \mathcal{N}^N(\mathbf{0}, M^{-1}) \implies \sqrt{det(M)} \int_{\mathbb{R}^N} \frac{d\mathbf{x}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}\mathbf{x}^T M \mathbf{x}} = 1 \tag{A.16}$$

Proof. Notice this is just the definition of density function for a multivariate normal, which integrates to 1. M is the inverse of the variance covariance matrix. Whenever this integral is proposed with M invertible, it is possible to state the equivalence to 1. Observe also that it is the multidimensional equivalent of Equation A.1. □

Lemma A.3 (Another perspective for Equation A.3)**.** *Let $X \sim \mathcal{N}(0,1)$. Then:*

$$\mathbb{E}_X \left[e^{\kappa x} \right] = e^{\frac{1}{2}\kappa^2} \tag{A.17}$$

Proof. Expanding the expectation:

$$\mathbb{E}_X \left[e^{\kappa x} \right] = \int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{\kappa x} \tag{A.18}$$

$$= \int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + \kappa x} \tag{A.19}$$

$$= \int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2\kappa x)} \tag{A.20}$$

$$= \int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2 - \kappa^2]} \qquad \text{completing the square} \tag{A.21}$$

$$= \int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2]} e^{-\frac{1}{2}(-\kappa^2)} \tag{A.22}$$

$$= e^{\frac{1}{2}\kappa^2} \underbrace{\int dx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\kappa)^2]}}_{x \sim \mathcal{N}(\kappa, 1)\ density} \tag{A.23}$$

$$= e^{\frac{1}{2}\kappa^2} \tag{A.24}$$

□

[1] wiki/gaussian_integral

Lemma A.4 (Gaussian Identity). *The following equality holds:*

$$\int dx(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\sigma^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{A.25}$$

Proof. Proceed by substitution, letting:

$$u = -\frac{(x-\mu)^2}{2\sigma^2} \implies du = -\frac{x-\mu}{\sigma^2}dx \implies -\sigma^2 du = (x-\mu)dx$$

So that:

$$\begin{aligned}
\int \underbrace{dx(x-\mu)}_{=du} exp\Big\{\underbrace{-\frac{(x-\mu)^2}{2\sigma^2}}_{=u}\Big\} &= -\sigma^2 \int due^u \\
&= -\sigma^2 e^u \\
&= -\sigma^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\end{aligned}$$

□

Appendix B

Random Matrices, Replica & Cavity

The following is a nice and rigorous application of the Replica Method presented in Chapter 3 and of the Cavity Method introduced in Chapter 2.

Definition B.1 (Random Matrix X). $X \in \mathbb{R}^{N\times N}$ is a matrix such that:

$$X_{ij} \sim \nu(\vartheta) \quad \forall i,j \in \{1,\dots,N\} \tag{B.1}$$

An interesting random matrix was used by Wigner to model the nuclei of heavy atoms [WF60].

Definition B.2 (Wigner Matrix A_N). A Wigner matrix is a random matrix where:

$$A_N = \frac{1}{\sqrt{2N}}(G+G^T) \quad \in \mathbb{R}^{N\times N} \tag{B.2}$$

$$G_{ij} \sim \mathcal{N}(0,1) \quad \forall i,j \tag{B.3}$$

Sometimes, when it is clear that the size is N, the subscript will be omitted

Proposition B.3 (Properties of A_N). *For a Wigner matrix as in Definition B.2 it holds that:*

1. $A_{ij} = A_{ji} \quad \forall i,j$
2. $\lambda_i \in \mathbb{R} \quad \forall i$ *eigenvalues*

Proof. **(Claim** 1**)** Almost definitional, as $G+G^T = G^T+G$
(Claim 2**)** By direct implication of Claim 1, a symmetric matrix has real eigenvalues. Let $\dagger$ denote the complex conjugate. For an eigenpair (x,λ) we have:

$$Ax = \lambda x \implies x^\dagger Ax - (x^\dagger Ax)^\dagger = 0 \qquad \text{symmetry of } A \tag{B.4}$$

$$x^\dagger \lambda x - (x^\dagger \lambda x)^\dagger = 0 \qquad \text{eigenpair} \tag{B.5}$$

$$(\lambda - \lambda^\dagger)(x^\dagger x) = 0 \tag{B.6}$$

$$\iff \lambda - \lambda^\dagger = 0 \iff \lambda = \lambda^\dagger \iff \lambda \in \mathbb{R} \tag{B.7}$$

Which holds for any eigenpair. □

An informative question to ask is what is the spectrum of a Wigner matrix? In other terms:

what is the distribution of its eigenvalues?

It turns out that both the Replica and the Cavity method can be used to determine the (limiting) distribution of the eigenvalues $\nu_{A_N}(\lambda)$ *as* $N \to \infty$. We will show that it tends to the **semicircle law** distribution:

$$\nu_{A_N}(\lambda) \to \begin{cases} \frac{1}{2\pi}\sqrt{4-\lambda^2} & \lambda \in [-2,2] \\ 0 & otherwise \end{cases} \tag{B.8}$$

Most of random matrix theory is based on the Stieltjes Transform. Its starting point is a very important identity for the dirac delta function.

Lemma B.4 (Dirac Delta identity). *Let $\Im$ denote the imaginary part of a complex valued number. Then:*

$$\delta(x-x_0) = -\lim_{\epsilon\to 0}\frac{1}{\pi}\Im\left[\frac{1}{x-x_0+i\epsilon}\right] \tag{B.9}$$

Definition B.5 (Stieltjes Transform $S_A(\cdot)$). Given a random matrix A, define the Stieltjes Transform as:

$$S_A(\lambda) = -\frac{1}{N}\sum_i \frac{1}{\lambda-\lambda_i} = -\frac{1}{N}\partial_\lambda \log[det(A-\lambda\mathbf{I})] \tag{B.10}$$

Where the second equality is proved for a similar case in Theorem B.6.

Theorem B.6 (Stieltjes Transform Properties). *The distribution of the eigenvalues of a Wigner matrix is of the form:*

$$\nu_A(\lambda) = \frac{1}{\pi}\lim_{\epsilon\to 0}\Im\left\{S_A(\lambda+i\epsilon)\right\} \tag{B.11}$$

Proof. Consider a uniform distribution $\nu(x)$ that can take $\{x_i\}_{i=1}^N$ values with uniform measure, using Lemma B.4 we have that:

$$\nu(x) = \frac{1}{N}\sum_i \delta(x-x_i) = -\lim_{\epsilon\to 0}\frac{1}{N\pi}\sum_i \Im\left[\frac{1}{x-x_i+i\epsilon}\right] \tag{B.12}$$

It can be noticed that a sum of fractions can be expressed as the derivative of a logarithm:

$$\frac{1}{N}\sum_i \frac{1}{x-x_i} = \frac{1}{N}\frac{d}{dx}\sum_i \log(x-x_i) = \frac{1}{N}\frac{d}{dx}\log\left[\prod_i (x-x_i)\right] \tag{B.13}$$

So that for A_N and its eigenvalues $\lambda_1,\ldots,\lambda_N$:

$$\begin{aligned}
\nu_A(\lambda) &= \frac{1}{N}\sum_i \delta(\lambda-\lambda_i) && && \text{(B.14)}\\
&= -\lim_{\epsilon\to 0}\frac{1}{N\pi}\sum_i \Im\left[\frac{1}{\lambda-\lambda_i+i\epsilon}\right] && \text{Lemma B.4} && \text{(B.15)}\\
&= -\frac{1}{N\pi}\Im\left\{\lim_{\epsilon\to 0}\frac{d}{d\lambda}\log\left[\prod_{i=1}^N \lambda-\lambda_i+i\epsilon\right]\right\} && \text{Eq B.13, linearity of } \Im && \text{(B.16)}\\
&= -\frac{1}{N\pi}\Im\left\{\lim_{\epsilon\to 0}\partial_\lambda \log\left[\prod_i (\lambda+i\epsilon)-\lambda_i\right]\right\} && && \text{(B.17)}\\
&= -\frac{1}{N\pi}\Im\left\{\lim_{\epsilon\to 0}\partial_\lambda \log\left[det[A-(\lambda+i\epsilon)\mathbf{I}]\right]\right\} && det \text{ definition} && \text{(B.18)}\\
&= \frac{1}{\pi}\lim_{\epsilon\to 0}\Im\left\{S_A(\lambda+i\epsilon)\right\} && \text{Def. B.5} && \text{(B.19)}
\end{aligned}$$

□

Being able to compute exactly the Stieltjes transform of Definition B.5 is sufficient to know the distribution of eigenvalues of a function. As $N \to \infty$, assume further that the density of eigenvalues is self-averaging with a result similar to that of Theorem 3.4.

We wish to compute:

$$\lim_{N\to\infty} \mathbb{E}[S_{A_N}(\lambda)] = -\partial_\lambda \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Big[\log[det(A_N - \lambda\mathbf{I})]\Big] \tag{B.20}$$

Where to simplify calculations we use the following for for the logarithm and expectation:

$$\mathbb{E}\Big[\log[det(A_N - \lambda\mathbf{I})]\Big] = -2\mathbb{E}\Big[\log[det(A_N - \lambda\mathbf{I})^{-\frac{1}{2}}]\Big] \tag{B.21}$$

B.1 Replica Method

Following the approach of Section 3.1, we avoid computing $\log[det(\ldots)^{-\frac{1}{2}}]$ and opt for a small n approximation through powers of the argument $det[\ldots]^{-\frac{1}{2}n}$.

$$\mathbb{E}\Big[-2\log[det(A - \lambda\mathbf{I})^{-\frac{1}{2}}]\Big] \approx -2\lim_{n\to 0}\left[\frac{\mathbb{E}[det(A-\lambda\mathbf{I})^{-\frac{1}{2}n}] - 1}{n}\right] \tag{B.22}$$

In this context, the formulation of Equation B.21 allows us to exploit the Gaussian Integral identity from Proposition A.2:

$$\sqrt{det(M)}\int_{\mathbb{R}^N} \frac{d\mathbf{x}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}\mathbf{x}^T M\mathbf{x}} = 1 \tag{B.23}$$

Indeed:

$$\mathbb{E}\Big[det[A - \lambda\mathbf{I}]^{-\frac{1}{2}n}\Big] = \mathbb{E}\left[\left(\frac{1}{\sqrt{det[A-\lambda\mathbf{I}]}}\right)^n\right] \tag{B.24}$$

$$= \mathbb{E}\left[\int_{\mathbb{R}^N} \frac{d\mathbf{x}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}\mathbf{x}^T(A-\lambda\mathbf{I})\mathbf{x}}\right] \qquad \text{Eq. B.23} \tag{B.25}$$

$$= \mathbb{E}\left[\prod_{\alpha=1}^n \int_{\mathbb{R}^N} \frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}(A-\lambda\mathbf{I})\mathbf{x}^{(\alpha)}}\right] \tag{B.26}$$

$$= \mathbb{E}\left[\int_{\mathbb{R}^N} \prod_{\alpha=1}^n \frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}(A-\lambda\mathbf{I})\mathbf{x}^{(\alpha)}}\right] \qquad \text{copies are } \perp\!\!\!\perp \tag{B.27}$$

$$= \mathbb{E}\left[\int_{\mathbb{R}^N} \prod_{\alpha=1}^n \left[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}}\right] \prod_{\alpha=1}^n e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}} e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}(-\lambda\mathbf{I})\mathbf{x}^{(\alpha)}}\right] \tag{B.28}$$

$$= \int_{\mathbb{R}^N} \prod_{\alpha=1}^n \left[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}(-\lambda\mathbf{I})\mathbf{x}^{(\alpha)}}\right] \mathbb{E}\left[\prod_{\alpha=1}^n e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}}\right] \qquad \text{linearity} \tag{B.29}$$

$$= \int_{\mathbb{R}^N} \prod_{\alpha=1}^n \left[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}} e^{\frac{\lambda}{2}\|\mathbf{x}^{(\alpha)}\|_2^2}\right] \mathbb{E}\left[\prod_{\alpha=1}^n e^{-\frac{1}{2}(\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}}\right] \qquad (\mathbf{x}^T)^{(\alpha)}\mathbf{I}\mathbf{x}^{(\alpha)} = \|\mathbf{x}^{(\alpha)}\|_2^2 \tag{B.30}$$

$$= \int_{\mathbb{R}^N} \prod_{\alpha=1}^n \left[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}} e^{\frac{\lambda}{2}\|\mathbf{x}^{(\alpha)}\|_2^2}\right] \mathbb{E}\left[e^{-\frac{1}{2}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}}\right] \tag{B.31}$$

Clearly randomness is in the matrix A, which is a Wigner matrix (Definition B.2). At this moment, its form becomes useful since the expectation inside the integral takes the form:

$$\mathbb{E}_A\left[e^{-\frac{1}{2}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}}\right] = \mathbb{E}_G\left[e^{-\frac{1}{2}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}[\frac{1}{\sqrt{2N}}(G+G^T)]\mathbf{x}^{(\alpha)}}\right] \tag{B.32}$$

$$= \mathbb{E}_G\left[e^{-\frac{1}{2\sqrt{2N}}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}G\mathbf{x}^{(\alpha)}+(\mathbf{x}^T)^{(\alpha)}G^T\mathbf{x}^{(\alpha)}}\right] \tag{B.33}$$

$$= \mathbb{E}_G\left[e^{-\frac{\cancel{2}}{\cancel{2}\sqrt{2N}}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}G\mathbf{x}^{(\alpha)}}\right] \qquad G, G^T \text{ equal in } \mathbb{E} \tag{B.34}$$

$$= \mathbb{E}_G\left[e^{-\frac{1}{\sqrt{2N}}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}G\mathbf{x}^{(\alpha)}}\right] \tag{B.35}$$

$$= \prod_{ij}\mathbb{E}_{G_{ij}}\left[e^{-\frac{G_{ij}}{\sqrt{2N}}\sum_\alpha x_i^{(\alpha)}x_j^{(\alpha)}}\right] \qquad G_{ij}\ i.i.d. \tag{B.36}$$

We recognize a well known Gaussian integral of the form of Proposition A.1

$$\int e^{-ax^2+bx}dx = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}} \quad or \quad \mathbb{E}_x[e^{bx}] = e^{\frac{b^2}{2a}} \tag{B.37}$$

Where for the latter, we set $b = -\dfrac{1}{\sqrt{2N}}\sum_\alpha x_i^{(\alpha)}x_j^{(\alpha)}$ and $a = 1$ to get:

$$\mathbb{E}_A\left[e^{-\frac{1}{2}\sum_\alpha (\mathbf{x}^T)^{(\alpha)}A\mathbf{x}^{(\alpha)}}\right] = \prod_{ij} e^{\frac{1}{4N}[\sum_\alpha x_i^{(\alpha)}x_j^{(\alpha)}]^2} \tag{B.38}$$

$$= \prod_{ij} e^{\frac{N}{4}\sum_\alpha\sum_\beta(\frac{x_i^{(\alpha)}x_j^{(\alpha)}x_i^{(\beta)}x_j^{(\beta)}}{N^2})} \qquad \text{recollect } N\text{, reindex square by } \alpha, \beta \tag{B.39}$$

$$= e^{\frac{N}{4}\sum_{\alpha,\beta}(\frac{\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}{N})^2} \tag{B.40}$$

Putting together Equations B.31 and B.40 we conclude that:

$$\mathbb{E}\left[det[A-\lambda\mathbf{I}]^{-\frac{1}{2}n}\right] = \int_{\mathbb{R}^N}\prod_{\alpha=1}^n\left[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}}e^{\frac{\lambda}{2}\|\mathbf{x}^{(\alpha)}\|_2^2}\right]e^{\frac{N}{4}\sum_{\alpha,\beta}(\frac{\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}{N})^2} \tag{B.41}$$

The integration over the disorder has coupled the independent replicas. We call this **overlap**.

Definition B.7 (Overlap $q^{(\alpha\beta)}$)**.** Define the overlap as:

$$q^{(\alpha\beta)} := \frac{1}{N}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)} \tag{B.42}$$

In order to decouple the overlap we use the same approach as that of Section 3.1 for the magnetization of the Curie-Weiss model. The delta function in this case is:

$$f\left(\frac{1}{N}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}\right) = N^n\int\prod_{1\leq\alpha\leq\beta\leq n}dq^{(\alpha\beta)}\delta(Nq^{(\alpha\beta)}-\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)})f(\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}) \tag{B.43}$$

Again, as before, we drop the N^n factor in front and use the approximation $\approx$ symbol. As $N \to \infty$ the normalized logarithm becomes:

$$\mathbb{E}\Big[det[A-\lambda\mathbf{I}]^{-\frac{1}{2}n}\Big] = \int_{\mathbb{R}^N}\prod_{\alpha=1}^{n}\Big[\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}}e^{\frac{\lambda}{2}\|\mathbf{x}^{(\alpha)}\|_2^2}\Big]e^{\frac{N}{4}\sum_{\alpha,\beta}(\frac{\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}{N})^2} \tag{B.44}$$

$$\approx \int_{\mathbb{R}^N}\prod_{1\le\alpha\le\beta\le n} dq^{(\alpha\beta)}\delta\Big(q^{(\alpha\beta)}-\frac{\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}{N}\Big)e^{N\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{N}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2} \tag{B.45}$$

Taking the Fourier representation with the change of variables $\widehat{q^{(\alpha\beta)}} = 2\pi i\lambda$ the products of delta functions become:

$$\prod_{1\le\alpha\le\beta\le n} dq^{(\alpha\beta)}\delta\Big(q^{(\alpha\beta)}-\frac{\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}{N}\Big) = \int\prod_{1\le\alpha\le\beta\le n}\frac{d\widehat{q^{(\alpha\beta)}}}{2\pi}e^{-\sum\limits_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}(Nq^{(\alpha\beta)}-\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)})} \tag{B.46}$$

Which joined with Equation B.45 returns:

$$\mathbb{E}\Big[det[A-\lambda\mathbf{I}]^{-\frac{1}{2}n}\Big] \approx \int_{\mathbb{R}^N,\mathbb{C}^N}\prod_{1\le\alpha\le\beta\le n}\frac{dq^{(\alpha\beta)}d\widehat{q^{(\alpha\beta)}}}{2\pi} \tag{B.47}$$

$$\times\, e^{-\sum\limits_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}(Nq^{(\alpha\beta)}-\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)})}e^{N\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{N}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2}$$

$$\approx \int_{\mathbb{R}^N,\mathbb{C}^N}\prod_{1\le\alpha\le\beta\le n}\frac{dq^{(\alpha\beta)}d\widehat{q^{(\alpha\beta)}}}{2\pi} \tag{B.48}$$

$$\times\, e^{-\sum\limits_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}(Nq^{(\alpha\beta)}-\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)})+N\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{N}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2}$$

$$\approx \int_{\mathbb{R}^N,\mathbb{C}^N}\prod_{1\le\alpha\le\beta\le n}\frac{dq^{(\alpha\beta)}d\widehat{q^{(\alpha\beta)}}}{2\pi} \tag{B.49}$$

$$\times\, exp\Big\{N\Big[-\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}\Big(q^{(\alpha\beta)}-\frac{1}{N}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}\Big)+\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{1}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2\Big]\Big\}$$

$$\approx \int_{\mathbb{R}^N,\mathbb{C}^N}\prod_{1\le\alpha\le\beta\le n}\frac{dq^{(\alpha\beta)}d\widehat{q^{(\alpha\beta)}}}{2\pi}e^{N\Phi(q^{(\alpha\beta)},\widehat{q^{(\alpha\beta)}})} \tag{B.50}$$

Where in the exponential:

$$\Phi(q^{(\alpha\beta)},\widehat{q^{(\alpha\beta)}}) = -\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}\Big(q^{(\alpha\beta)}-\frac{1}{N}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}\Big)+\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{1}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2 \tag{B.51}$$

$$= -\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}q^{(\alpha\beta)}+\frac{\lambda}{2}\sum_\alpha q^{(\alpha\alpha)}+\frac{1}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2+\Psi_x(\widehat{q^{(\alpha\beta)}}) \qquad \text{reordering} \tag{B.52}$$

And further, by the decoupling of $\alpha\beta$ sites, the Ψ function can be simplified to:

$$\Psi_x(\widehat{q^{(\alpha\beta)}}) = -\sum_{1\leq\alpha\leq\beta\leq n} -\widehat{q^{(\alpha\beta)}}\frac{1}{N}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)} \tag{B.53}$$

$$= \frac{1}{N}\sum_{1\leq\alpha\leq\beta\leq n}\log\left\{e^{\widehat{q^{(\alpha\beta)}}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}\right\} \tag{B.54}$$

$$= \frac{1}{N}\log\left\{\prod_{1\leq\alpha\leq\beta\leq n} e^{\widehat{q^{(\alpha\beta)}}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}\right\} \tag{B.55}$$

$$= \frac{1}{N}\log\left\{e^{\sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}}\right\} \tag{B.56}$$

$$= \frac{1}{N}\log\int\prod_{\alpha}\frac{d\mathbf{x}^{(\alpha)}}{(2\pi)^{\frac{N}{2}}}e^{\sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}\mathbf{x}^{(\alpha)}\cdot\mathbf{x}^{(\beta)}} \tag{B.57}$$

$$= \frac{1}{N}\log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{\sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}x^{(\alpha)}\cdot x^{(\beta)}}\right\}^N \tag{B.58}$$

$$= \frac{1}{N}N\log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{\sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}x^{(\alpha)}\cdot x^{(\beta)}}\right\} \tag{B.59}$$

$$= \log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{\sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}x^{(\alpha)}\cdot x^{(\beta)}}\right\} \tag{B.60}$$

The integral in Equation B.50 can be evaluated with the saddle point method being in exponential form. Thus:

$$\mathbb{E}\left[det[A-\lambda\mathbf{I}]^{-\frac{1}{2}n}\right] \approx exp\left\{N\underset{\widehat{q^{(\alpha\beta)}},q^{(\alpha\beta)}}{Extr}\left[\Phi(q^{(\alpha\beta)},\widehat{q^{(\alpha\beta)}})\right]\right\} \tag{B.61}$$

As in Section 3.1, the extremization is over a difficult space[1]. The replica symmetric ansatz in this case is stated as:

$$\begin{cases} q^{(\alpha\beta)} = \delta^{(\alpha\beta)}q \\ \widehat{q^{(\alpha\beta)}} = -\frac{1}{2}\delta^{(\alpha\beta)}\widehat{q} \end{cases} \implies \begin{cases} \sum_\alpha q^{(\alpha\alpha)} = nq \\ \sum_{\alpha,\beta}(q^{(\alpha\beta)})^2 = nq^2 \\ \sum\limits_{1\leq\alpha\leq\beta\leq n}\widehat{q^{(\alpha\beta)}}q^{(\alpha\beta)} = -\frac{n}{2}q\widehat{q} \end{cases} \tag{B.62}$$

[1]In this case, a matrix $\mathbb{R}^{n\times n}$.

This simplification has an effect on the Ψ integral:

$$\implies \Psi_x(\widehat{q^{(\alpha\beta)}}) = \log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}x^{(\alpha)}\cdot x^{(\beta)}}\right\} \tag{B.63}$$

$$= \log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}Nq^{(\alpha\beta)}}\right\} \quad \text{Definition B.7} \tag{B.64}$$

$$\Psi(\widehat{q}) = \log\left\{\int\prod_{\alpha}\frac{dx^{(\alpha)}}{\sqrt{2\pi}}e^{-\frac{Nn}{2}q\widehat{q}}\right\} \tag{B.65}$$

$$= n\log\left\{\int\frac{dx}{\sqrt{2\pi}}e^{-\frac{N}{2}q\widehat{q}}\right\} \quad \text{take out } n \tag{B.66}$$

$$= n\log\left\{\int\frac{dx}{\sqrt{2\pi}}e^{-\frac{1}{2}\widehat{q}x^2}\right\} \quad \text{Eq. B.62} \tag{B.67}$$

$$= n\log\left\{\sqrt{\frac{1}{\widehat{q}}}\right\} \quad \text{Fct. A.1} \tag{B.68}$$

$$= -\frac{n}{2}\log(\widehat{q}) \tag{B.69}$$

Where in the last passage we use the Gaussian integral of Equation A.2. We now have a closed easy form for the Φ function:

$$\Phi(q^{(\alpha\beta)},\widehat{q^{(\alpha\beta)}}) = -\sum_{1\le\alpha\le\beta\le n}\widehat{q^{(\alpha\beta)}}q^{(\alpha\beta)} + \frac{\lambda}{2}\sum_{\alpha}q^{(\alpha\alpha)} + \frac{1}{4}\sum_{\alpha,\beta}(q^{(\alpha\beta)})^2 + \Psi_x(\widehat{q^{(\alpha\beta)}}) \quad \text{Eq. B.52} \tag{B.70}$$

$$\implies \Phi(q,\widehat{q}) = \frac{n}{2}q\widehat{q} + \frac{\lambda}{2}nq + \frac{1}{4}nq^2 - \frac{n}{2}\log(\widehat{q}) \quad \text{Eqs. B.62, B.68} \tag{B.71}$$

$$= -\frac{n}{2}\Big(-q\widehat{q} - \lambda q - \frac{1}{2}q^2 + \log(\widehat{q})\Big) \tag{B.72}$$

Which plugged into the extremization and the approximation, with an added $\frac{1}{N}$ factor as in Equation B.20 to include in the limit for $N \to \infty$:

$$\mathbb{E}\Big[-\frac{2}{N}\log[det(A-\lambda\mathbf{I})^{-\frac{1}{2}}]\Big] \approx -2\lim_{n\to 0}\Big[\frac{\mathbb{E}[det(A-\lambda\mathbf{I})^{-\frac{1}{2}n}]-1}{n}\Big] \quad \text{Eq. B.22} \tag{B.73}$$

$$\approx -\frac{2}{N}\lim_{n\to 0}\frac{exp\Big\{N\underset{\widehat{q^{(\alpha\beta)}},q^{(\alpha\beta)}}{Extr}\Big[\Phi(q^{(\alpha\beta)},\widehat{q^{(\alpha\beta)}})\Big]\Big\}-1}{n} \quad \text{Eq. B.31} \tag{B.74}$$

$$\approx -\frac{2}{N}\lim_{n\to 0}\frac{exp\Big\{\frac{-Nn}{2}\underset{\widehat{q},q}{Extr}\Big[-q\widehat{q}-\lambda q-\frac{1}{2}q^2+\log(\widehat{q})\Big]\Big\}-1}{n} \quad \text{Eq. B.72} \tag{B.75}$$

$$\approx -\frac{2}{N}\lim_{n\to 0}\frac{\Big[exp\Big\{\frac{-N}{2}\underset{\widehat{q},q}{Extr}\Big[-q\widehat{q}-\lambda q-\frac{1}{2}q^2+\log(\widehat{q})\Big]\Big\}\Big]^n-1}{n} \tag{B.76}$$

$$\approx \frac{2}{N}\frac{N}{2}\underset{\widehat{q},q}{Extr}\Big[-q\widehat{q}-\lambda q-\frac{1}{2}q^2+\log(\widehat{q})\Big] \tag{B.77}$$

$$\approx \underset{\widehat{q},q}{Extr}\Big[-q\widehat{q}-\lambda q-\frac{1}{2}q^2+\log(\widehat{q})\Big] \tag{B.78}$$

Where in the last passage we use the limit:

$$\lim_{n\to 0}\frac{e^{an}-1}{n}=a \tag{B.79}$$

With $a=\frac{N}{2}\underset{\widehat{q},q}{Extr}\Big[-q\widehat{q}-\lambda q-\frac{1}{2}q^2+\log(\widehat{q})\Big]$, and simplify the equation accordingly. To solve the extremization, we impose null partial derivatives of the argument of the extremization, denoted as Q:

$$\begin{cases}\frac{\partial}{\partial q}Q=-\widehat{q}-\lambda-q=0\\ \frac{\partial}{\partial \widehat{q}}Q=-q+\frac{1}{\widehat{q}}=0\end{cases}\iff\begin{cases}\widehat{q}=\frac{1}{q}\\ \frac{1}{q}+\lambda+q=0\end{cases}\iff q^*=\frac{-\lambda\pm\sqrt{\lambda^2-4}}{2} \tag{B.80}$$

Which eventually tells us that:

$$\lim_{N\to\infty}\mathbb{E}[S_{A_N}(\lambda)]=-\partial_\lambda\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\Big[\log[det(A_N-\lambda\mathbf{I})]\Big]\qquad\text{Eq. B.20} \tag{B.81}$$

$$=q^*=\frac{-\lambda\pm\sqrt{\lambda^2-4}}{2} \tag{B.82}$$

While there are two solutions, given that the spectrum is a probability distribution, the negative one will be discarded.

Recalling Equation B.19 we have:

$$\nu_A(\lambda)=\frac{1}{\pi}\lim_{\epsilon\to 0}\Im\Big\{S_A(\lambda+i\epsilon)\Big\} \tag{B.83}$$

$$=\frac{1}{\pi}\lim_{\epsilon\to 0}\Im\Big\{\frac{-\lambda-i\epsilon\pm\sqrt{(\lambda+i\epsilon)^2-4}}{2}\Big\} \tag{B.84}$$

Concentrating on the term inside the root:

$$\sqrt{(\lambda+i\epsilon)^2-4}=\sqrt{\lambda^2+2i\epsilon-\epsilon^2-4} \tag{B.85}$$

$$=\sqrt{(\lambda^2-\epsilon^2-4)\Big(1+\frac{2i\epsilon\lambda}{\lambda^2-\epsilon^2-4}\Big)} \tag{B.86}$$

$$=\sqrt{(\lambda^2-\epsilon^2-4)}\sqrt{\Big(1+\frac{2i\epsilon\lambda}{\lambda^2-\epsilon^2-4}\Big)} \tag{B.87}$$

$$\underset{\epsilon\to 0}{\to}\sqrt{(\lambda^2-\epsilon^2-4)} \tag{B.88}$$

$$\approx\sqrt{\lambda^2-4}\sqrt{1-\frac{\epsilon^2}{\lambda^2-4}} \tag{B.89}$$

$$\approx\sqrt{\lambda^2-4}+O(\epsilon) \tag{B.90}$$

So, if $|\lambda|>2$ the square root is real and as $\epsilon\to 0$ all the imaginary parts get canceled. On the contrary, if $\lambda|<2$ then $\sqrt{\lambda^2-4}=i\sqrt{4-\lambda^2}$ and:

$$|\lambda|<2\implies\frac{1}{\pi}\lim_{\epsilon\to 0}\Im\Big\{\frac{-\lambda-i\epsilon\pm\sqrt{(\lambda+i\epsilon)^2-4}}{2}\Big\}=\frac{1}{\pi}\Im\Big\{\frac{-\lambda-i\epsilon\pm i\sqrt{4-\lambda^2}+O(\epsilon)}{2}\Big\} \tag{B.91}$$

$$=\frac{1}{\pi}\Im\Big\{\frac{-\lambda\pm\sqrt{4-\lambda^2}}{2}\Big\} \tag{B.92}$$

$$=\frac{1}{2\pi}\Big(\pm\sqrt{4-\lambda^2}\Big) \tag{B.93}$$

Where we take the + solution to obtain a valid density over eigenvalues and conclude that as claimed:

$$\nu_A(\lambda) = \begin{cases} \frac{1}{2\pi}\sqrt{4-\lambda^2} & |\lambda| < 2 \\ 0 & otherwise \end{cases} \tag{B.94}$$

In the notes, the two factor is **omitted**, but it is necessary to make the distributino integrate to 1 over its support, and is also justified by the q^* solution chosen.

The cavity method solution can be found in the original lecture notes [KZ21b].

Bibliography

[KZ21a] Florent Krzakala and Lenka Zdeborová. *Statistical Physics For Optimization and Learning*. Statistical Physics For Optimization and Learning. 2021. URL: https://sphinxteam.github.io/EPFLDoctoralLecture2021/ (visited on 10/03/2022).

[KZ21b] Florent Krzakala and Lenka Zdeborová. "Statistical Physics Methods in Optimization and Machine Learning". In: (2021), p. 225.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford graduate texts. OCLC: ocn234430714. Oxford ; New York: Oxford University Press, 2009. 569 pp. ISBN: 978-0-19-857083-7.

[ZK16] Lenka Zdeborová and Florent Krzakala. "Statistical physics of inference: Thresholds and algorithms". In: *Advances in Physics* 65.5 (Sept. 2, 2016), pp. 453–552. ISSN: 0001-8732, 1460-6976. DOI: 10.1080/00018732.2016.1211393. arXiv: 1511.02476[cond-mat,stat]. URL: http://arxiv.org/abs/1511.02476 (visited on 09/27/2022).

[AH76] K. Appel and W. Haken. "The existence of unavoidable sets of geographically good configurations". In: *Illinois Journal of Mathematics* 20.2 (June 1976). Publisher: Duke University Press, pp. 218–297. ISSN: 0019-2082, 1945-6581. DOI: 10.1215/ijm/1256049898. URL: https://projecteuclid.org/journals/illinois-journal-of-mathematics/volume-20/issue-2/The-existence-of-unavoidable-sets-of-geographically-good-configurations/10.1215/ijm/1256049898.full (visited on 09/27/2022).

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8. URL: https://linkinghub.elsevier.com/retrieve/pii/0893608089900208 (visited on 07/11/2022).

[Cyb89] G Cybenkot. "Approximation by superpositions of a sigmoidal function". In: (1989), p. 12.

[Bar93] A.R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information Theory* 39.3 (May 1993), pp. 930–945. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/18.256500. URL: https://ieeexplore.ieee.org/document/256500/ (visited on 07/11/2022).

[Bar94] Andrew R. Barron. "Approximation and estimation bounds for artificial neural networks". In: *Machine Learning* 14.1 (Jan. 1, 1994), pp. 115–133. ISSN: 1573-0565. DOI: 10.1007/BF00993164. URL: https://doi.org/10.1007/BF00993164 (visited on 07/11/2022).

[Deb09] P. Debye. "Näherungsformeln für die Zylinderfunktionen für große Werte des Arguments und unbeschränkt veränderliche Werte des Index". In: *Mathematische Annalen* 67.4 (Dec. 1, 1909), pp. 535–558. ISSN: 1432-1807. DOI: 10.1007/BF01450097. URL: https://doi.org/10.1007/BF01450097 (visited on 09/27/2022).

[RWD13] "Sullo svolgimento del quoziente di due serie ipergeometriche in frazione continua infinita. (1863.)" In: *Bernard Riemann's gesammelte mathematische Werke und wissenschaftlicher Nachlass*. Ed. by Bernhard Riemann, Heinrich Martin Weber, and Richard Dedekind. Cambridge Library Collection - Mathematics. Cambridge: Cambridge University Press, 2013, pp. 400–406. ISBN: 978-1-108-05935-0. DOI: 10.1017/CBO9781139568050.024. URL: https://www.cambridge.org/core/books/bernard-riemanns-gesammelte-mathematische-werke-und-wissenschaftlicher-nachlass/sullo-svolgimento-del-quoziente-di-due-serie-ipergeometriche-in-frazione-continua-infinita-1863/81713E59B896BE28BEF854361 (visited on 09/27/2022).

[MPV86] M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Lecture Notes in Physics. WORLD SCIENTIFIC, Nov. 1986. ISBN: 978-9971-5-0116-7 978-981-279-937-1. DOI: 10.1142/0271. URL: https://www.worldscientific.com/worldscibooks/10.1142/0271 (visited on 09/30/2022).

[Gue03] Francesco Guerra. "Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model". In: *Communications in Mathematical Physics* 233.1 (Feb. 1, 2003), pp. 1–12. ISSN: 0010-3616, 1432-0916. DOI: 10.1007/s00220-002-0773-5. arXiv: cond-mat/0205123. URL: http://arxiv.org/abs/cond-mat/0205123 (visited on 09/30/2022).

[Ons44] Lars Onsager. "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition". In: *Physical Review* 65.3 (Feb. 1, 1944). Publisher: American Physical Society, pp. 117–149. DOI: 10.1103/PhysRev.65.117. URL: https://link.aps.org/doi/10.1103/PhysRev.65.117 (visited on 10/17/2022).

[DM09] Amir Dembo and Andrea Montanari. *Gibbs Measures and Phase Transitions on Sparse Random Graphs*. Oct. 28, 2009. DOI: 10.48550/arXiv.0910.5460. arXiv: 0910.5460[cond-mat,physics:math-ph]. URL: http://arxiv.org/abs/0910.5460 (visited on 10/17/2022).

[Dem+12] Amir Dembo et al. *The replica symmetric solution for Potts models on d-regular graphs*. July 23, 2012. DOI: 10.48550/arXiv.1207.5500. arXiv: 1207.5500[cond-mat]. URL: http://arxiv.org/abs/1207.5500 (visited on 10/17/2022).

[CV14] Amin Coja-Oghlan and Dan Vilenchik. *Chasing the k-colorability threshold*. Apr. 29, 2014. DOI: 10.1093/imr/rnv333. arXiv: 1304.1063[cs,math]. URL: http://arxiv.org/abs/1304.1063 (visited on 10/17/2022).

[KS66] H. Kesten and B. P. Stigum. "A Limit Theorem for Multidimensional Galton-Watson Processes". In: *The Annals of Mathematical Statistics* 37.5 (1966). Publisher: Institute of Mathematical Statistics, pp. 1211–1223. ISSN: 0003-4851. URL: https://www.jstor.org/stable/2239076 (visited on 10/17/2022).

[AT78] J. R. L. de Almeida and D. J. Thouless. "Stability of the Sherrington-Kirkpatrick solution of a spin glass model". In: *Journal of Physics A: Mathematical and General* 11.5 (May 1978). Publisher: IOP Publishing, pp. 983–990. ISSN: 0305-4470. DOI: 10.1088/0305-4470/11/5/028. URL: https://doi.org/10.1088/0305-4470/11/5/028 (visited on 10/17/2022).

[McG11] Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. OCLC: ocn670481486. New Haven [Conn.]: Yale University Press, 2011. 320 pp. ISBN: 978-0-300-16969-0.

[CT06] T. M. Cover and Joy A. Thomas. *Elements of information theory*. 2nd ed. OCLC: ocm59879802. Hoboken, N.J: Wiley-Interscience, 2006. 748 pp. ISBN: 978-0-471-24195-9.

[GSV04] Dongning Guo, Shlomo Shamai, and Sergio Verdu. *Mutual Information and Minimum Mean-square Error in Gaussian Channels*. Dec. 23, 2004. arXiv: cs/0412108. URL: http://arxiv.org/abs/cs/0412108 (visited on 10/18/2022).

[Nis80] H. Nishimori. "Exact results and critical properties of the Ising model with competing interactions". In: *Journal of Physics C: Solid State Physics* 13.21 (July 1980). Publisher: IOP Publishing, pp. 4071–4076. ISSN: 0022-3719. DOI: 10.1088/0022-3719/13/21/012. URL: https://doi.org/10.1088/0022-3719/13/21/012 (visited on 10/18/2022).

[Nis93] Hidetoshi Nishimori. "Optimum Decoding Temperature for Error-Correcting Codes". In: *Journal of the Physical Society of Japan* 62.9 (Sept. 15, 1993). Publisher: The Physical Society of Japan, pp. 2973–2975. ISSN: 0031-9015. DOI: 10.1143/JPSJ.62.2973. URL: https://journals.jps.jp/doi/10.1143/JPSJ.62.2973 (visited on 10/18/2022).

[Iba99] Yukito Iba. "The Nishimori line and Bayesian Statistics". In: *Journal of Physics A: Mathematical and General* 32.21 (May 28, 1999), pp. 3875–3888. ISSN: 0305-4470, 1361-6447. DOI: 10.1088/0305-4470/32/21/302. arXiv: cond-mat/9809190. URL: http://arxiv.org/abs/cond-mat/9809190 (visited on 10/18/2022).

[Sha48] C. E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 00058580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024 (visited on 10/18/2022).

[Der81] Bernard Derrida. "Random-energy model: An exactly solvable model of disordered systems". In: *Physical Review B* 24.5 (Sept. 1, 1981). Publisher: American Physical Society, pp. 2613–2626. DOI: 10.1103/PhysRevB.24.2613. URL: https://link.aps.org/doi/10.1103/PhysRevB.24.2613 (visited on 10/18/2022).

[WF60] E. P. Wigner and U. Fano. "Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra". In: *American Journal of Physics* 28.4 (Apr. 1960), pp. 408–409. ISSN: 0002-9505, 1943-2909. DOI: 10.1119/1.1935822. URL: http://aapt.scitation.org/doi/10.1119/1.1935822 (visited on 09/30/2022).