Elements of Statistical Mechanics and Large Deviation Theory

Contents

Chapter 1. Introduction . . . 1
1. Equilibrium Statistical Mechanics . . . 1
2. Large deviations . . . 5
3. Spin systems . . . 8

Chapter 2. The Maximum Entropy Principle . . . 9
1. The Gibbs Distribution . . . 13
2. Uniqueness of Entropy . . . 15

Chapter 3. Large Deviations for finite alphabets . . . 19
1. The Theorem of Sanov . . . 19
2. Applications . . . 25
3. The Theorem of Sanov for Pairs . . . 31

Chapter 4. The Ideal Gas . . . 35
1. The Microcanonical Distribution . . . 36
2. The equilibrium value of the kinetic energy . . . 38
3. The equilibrium value of other observables . . . 40

Chapter 5. The Large Deviation Principle . . . 41
1. Definition of the LDP . . . 41
2. The Varadhan Lemma . . . 45
3. LDP for Tilted Measures . . . 46
4. The Contraction Principle . . . 47

Chapter 6. The Curie-Weiss Model . . . 49
1. The LDP for the magnetization . . . 50
2. The free energy . . . 52

Chapter 7. The Theorem of Cramér . . . 53
1. The logarithmic moment generating function . . . 53
2. Main Theorem . . . 54
3. The Theorem of Cramér in $\mathbb{R}^d$ . . . 58

Chapter 8. The Ising Model (COMPLETER) . . . 59
1. The FKG Inequality . . . 60
2. The thermodynamic limit . . . 60
3. Proofs . . . 62

Chapter 9. The DLR Formalism . . . 65
1. Random Fields via Kolmogorov's Extension Theorem . . . 65

2. Random Fields via Specifications 66
3. The Ising model, again 71
4. An Inhomogeneous Ising chain on ℕ 71
5. Uniqueness; Dobrushin's Condition of Weak Dependence 72

Chapter 10. The Variational Principle 81
1. Introduction 81
2. The Entropy of an Invariant Random Field 83
3. Gibbs measures as reference measures 86

Chapter 11. Gibbs Measures and Large Deviations 93
1. Introduction 93
2. The Free Gibbs Measure 94
3. The LDP for L_n under the product measure 95
4. The LDP for L_n under the Free Gibbs measure 105

CHAPTER 1

Introduction

These notes aim at presenting some aspects of two intimately related areas, namely Equilibirum Statistical Mechanics (ESM) and Large Deviation Theory (LDT). On one side, ESM defines and studies the probability measures associated to large systems of particles. On the other, LDT is a classical chapter of probability theory which can be loosely described as a refinement of the Law of Large Numbers. I will try to introduce concepts from both sides in the most natural way, and show what are their common features. The text does not aim at presenting the most general results, but rather at going deeper into the richness of a few examples, such as random variables with values in a finite alphabets, which, in the statistical mechanics language, amounts to restrict to lattice spin systems.

The material presented in these notes is taken from a series of standard texts on the subject. There are two references that cover both LDT and ESM: [**?**, **?**]. For LDT, the main references are [**?**, **?**, **?**, **?**, **?**]. A non-technical reference, which covers a few aspects of the material in the simplest way, is [**?**]. Concerning ESM, some basic references are [**?**, **?**, **?**, **?**, **?**]. Finally, a series of papers on equivalence of ensembles will be exposed: [**?**, **?**]. [**?**, **?**, **?**] In this introduction I briefly describe what will be the main lines followed in the notes.

1. Equilibrium Statistical Mechanics

Consider for example a gas [1], composed of a large number of identical particles, say $n = 10^{25}$. Statistical mechanics is concerned with giving a reasonable description of such a large system. By a *reasonable description*, we mean a theory capable of making prediction about some properties of the system, like how the gas will react to external forces or thermodynamic changes. Since the gas is made of particles, knowing the state of each particle is equivalent to knowing the state of the gas. Therefore, we can assume that a perfect knowledge of the system at a given time t is a vector

$$x(t) = (q_1(t), p_1(t), \dots, q_n(t), p_n(t)) \in (\mathbb{R}^3 \times \mathbb{R}^3)^n \simeq \mathbb{R}^{6n},$$

giving the position $q_i(t) \in \mathbb{R}^3$ and a momentum $p_i(t) \in \mathbb{R}^3$ of each particle. If an initial confition $x_0 \in \mathbb{R}^{6n}$ is fixed, i.e. $x_0 = x(t = 0)$, Classical Newtonian mechanics gives, in principle, a way of knowing $x(t)$ for each $t > 0$. If the interaction

[1] We will give a detailed analysis of the model presented here in Chapter 4. The informal discussion of this introduction also applies to other systems, such as ferromagnets, which will also be considered later.

potential among the particles is given (and not too singular), then $x(t)$ is solution of a system of $\mathbb{R}^{6n}$ first order differential equations (Newton Equations). Unless the potential is trivial, or if the initial condition has pathological features, solving this differential system and obtaining explicit information on the time evolution seems a rather tough analytic problem.

But before even opening a textbook on ordinary differential equations, the reader must agree that solving exactly the above system is not what one wants to do, for the following reasons.

First, the exact microscopic position of each atom in a gas is not a particularly exciting piece of information, since our aim is to describe more relevant observables related to the *global* behaviour of the system (see U_n hereafter). In order to solve the above system of differential equations, one must also determine the initial condition x_0 which in itself should be considered as impossible, at least experimentally.

Second, we can start by restricting our attention to particular conditions. A simple but interesting one is the one which consists in waiting for the gas to have reached *equilibrium*. This amounts to study the limit $t \to \infty$. From the analytical point of view mentionned before, this limit seems to be even more difficult (existence of solutions to differential equations are typically guaranteed over finite time intervals), but the system seems nevertheless simpler to describe once a certain equilibrium has been reached.

To illustrate these ideas, consider a typical thermodynamic quantity like the average kinetic energy (we assume all particles have equal mass $m = 1$)

$$U_n(t) = \sum_{i=1}^{n} \frac{p_i(t)^2}{2} .$$

When n is large, $U_n(t)$ must be of order n. Common sense leads us to think that at equilibrium, $U_n(t)/n$ depends weakly on time. One sees here how the information contained in $x(t)$ is redundant: at a microscopic scale, the individual $p_i(t)$s certainly suffer dramatic changes over short intervals of times (Brownian motion), although interesting macroscopic quantities like $U_n(t)/n$ remain essentially constant. Common sense is thus lead to believe that there must exist some method which allows to compute $U_n(t)/n$ at least up to a certain precision, without necessarily knowing exactly $x(t)$. Therefore, it is reasonnable to abandon the search for the perfect knowledge of the state of the system, and to seek for an alternate way of computing observables, at least within a certain precision.

Accepting that one does not have access to the perfect knowledge of the state of the system is equivalent to describing the system using *probability theory*. Quoting Jaynes [**?**],

> *The purpose of probability theory is to help us forming plausible conclusions in cases where there is not enough information to lead to a certain information.*

If one assumes that one does not have access to the exact microscopic state at time t, one is lead to look for a probability distribution P_t on $\mathbb{R}^{6n}$. Assuming the system is at equilibrium, one can assume that the distribution P_t does not depend on time, and simply denote it by P. Then, a reasonable statement about the average kinetic energy is that any measurement of U_n/n will result, with overwhelming P-probability, in a number lying close to some ideal value $\bar{u}$. More precisely, there exists a small interval $[\bar{u}-\delta, \bar{u}+\delta] \subset \mathbb{R}$ and an $\epsilon > 0$ such that

$$P\Big(\frac{U_n}{n} \in [\bar{u}-\delta, \bar{u}+\delta]\Big) \geq 1-\epsilon\,. \tag{1}$$

Clearly, δ and ϵ must be small enough in order to provide an interesting information. We also expect that ϵ can be taken arbitrarily small when n is large, i.e.

$$\lim_{n\to\infty} P\Big(\frac{U_n}{n} \in [\bar{u}-\delta, \bar{u}+\delta]\Big) = 1\,. \tag{2}$$

(2) is nothing but a LLN-like statement and as will be seen, obtaining an optimal relation between ϵ, δ and n will be the content of the *Large Deviation Principle.* It happens that the precise relation between ϵ and n will involve the thermodynamic potentials of the system under consideration (free energy, pressure). A simplified model (the ideal gas) of the above situation will described in details, using large deviations techniques, in Chapter 4.

Before going further, let us summarize the previous discussion into a few starting principles regarding the statistical mechanical description of large systems composed of simple elements.

(1) **Randomness:** If a system is composed of a large number of elements, it is hopeless and useless to obtain a theory aimed at describing the exact state of each individual elements. Adopting a probabilistic viewpoint, an *observation* of the system is a random realization of a random experiment (of which the probability space, in particular the probability measure P, must be specified). The global properties of the system can then be studied using the tools and methods from probability theory.

(2) **Micro and Macroscopic quantities:** There are different types of observables related to the different *scales* of the system. Some, like the individual variables, or like local quantities which depend only on a finite number of variables, are called *microscopic.* Others, like the average total momentum or the average total magnetization, depend on the whole system and are essentially insensitive to the change of a finite number of variables [2], and are called *macroscopic.* These are relevant in thermodynamics, for example, which is a theory giving detailed relations among macroscopic variables.

[2] In probability theory, these observables are called *tail* measurable.

(3) **Equilibrium:** To simplify, we aim at describing only systems which have had the time to adapt (mechanically) to all exterior constraints, and therefore to introduce a probability distribution independent of time. Although the notion of equilibrium is rather subtle, it coincides with our intuition that same measurments of macroscopic quantities made in the same conditions lead to same results (within a certain range of precision). For example, the density of a gas remains constant regardless of the microscopic changes, which occur constantly along the time evolution. Therefore, a characterisation of equilibrium is that although the microscopic variables are random, the macroscopic ones *are deterministic*, i.e. constant with probability one.

With these basic principles at hand, the aim of equilibrium statistical mechanics (and, partially, of these notes) is then

(1) *Decide which probability measures P are best suited for the description of large systems.* Part of this problem is to determine how certain parameters (temperature [3], external magnetic field, etc.) enter in the definition of P.
(2) *Through the study of the chosen measure P, relate the microscopic details of the model to the large scale macroscopic behaviour.* In this study, a natural way of testing this correspondence is to consider the *thermodynamic limit*, in which the size of the system goes to infinity. The fluctuations of the macroscopic quantities must be studied in details, related to the size of the system, and should be shown to become negligible in the thermodynamic limit, leading to a deterministic macroscopic description.
(3) *If there are different possible choices for the measure P describing a system, then these must be shown to lead to equivalent results in the thermodynamic limit.* Namely, the various microscopic descriptions should all lead to the same macroscopic behaviour.

These three long-term objectives of equilibrium statistical mechanics will be seen to fit together naturally in the framework of Large deviation Theory (LDT). Point (1) is a recurrent theme of these notes; we will first provide a simple and natural way of defining equilibrium probability measures, using the Maximum Entropy Principle in Chapter 2. This will require in particular the definition of the fundamental quantity of information theory, the Shannon Entropy. Point (2) will be studied from the point of view of LDT; we will use the concentration results provided by Large Deviation Principles to simple models of statistical mechanics, and establish the relations between the rate functions of these principles and the thermodynamic functions of statistical mechanics. (3) is the subject of the *Equivalence of Ensembles*, in which various tools from LDT will be used, following a series of papers by Lewis, Pfister and Sullivan.

[3]Probability theory does not say how the notion of temperature must be introduced into the definition of a probability measure. Rather, the temperature will be introduced by *analogy*, by comparison with thermodynamics.

Remark 1.1. Part of point (2) is to give a description of *phase transitions*. Although this is a central problem in statistical mechanics, it will only be slightly studied in these notes.

We move on to a short description of LDT.

2. Large deviations

Large Deviation Theory is a classical chapter of probability theory. It can be resumed as a theory giving a deeper analysis of the concentration provided by the Law of Large Numbers (LLN), in the sense that it studies exponential convergence of sequences of certain random objects around their expected value. These can be averages of collections of variables or their empirical measure, but also more general objects like probability measures on metric spaces, in which case the theory takes its more general form nowadays.

As an illustration, consider a sequence of i.i.d. random variables $X_1, X_2, \dots$ with common distribution μ. Let $S_n = X_1 + \cdots + X_n$. The Strong Law of Large Numbers (SLLN) states that if $m := E[X_1] = \int x\mu(dx)$ exists, then the empirical mean $\frac{S_n}{n}$ converges to m in the limit $n \to \infty$:

$$\frac{S_n}{n} \to m\,, \quad a.s. \tag{3}$$

As a consequence, the Weak Law of Large Numbers (WLLN) states that for all $\epsilon > 0$,

$$P\Big(\Big|\frac{S_n}{n} - m\Big| \geq \epsilon\Big) \to 0\,. \tag{4}$$

The event $\{\frac{S_n}{n} - m| \geq \epsilon\}$ is called a *large deviation*, in the sense that it describes a deviation of order n of S_n far from its mean value mn. The LLN thus states that large deviations of the mean have small probability. When looked at on a finer scale around mn, standard deviations of order $\sqrt{n}$ are probable and random. Namely, if X_1 has finite variance $\sigma^2 < \infty$, then the Central Limit Theorem (CLT) states that

$$\frac{S_n - mn}{\sigma\sqrt{n}} \Rightarrow \mathcal{N}(0,1)\,. \tag{5}$$

(5) clearly implies (4), but a natural question is to know if the convergence (4) can be described in more details.

Large Deviation Theory tipically gives sharp bounds for the concentration of $\frac{S_n}{n}$ around m, in that it characterises the speed at which the convergence in (4) occurs. Under fairly general hypothesis, this convergence happens to be *exponential* in n:

$$e^{-c_1 n} \leq P\Big(\Big|\frac{S_n}{n} - m\Big| \geq \epsilon n\Big) \leq e^{-c_2 n}\,, \tag{6}$$

where $c_1 > c_2$. The detailed study of the constants c_1, c_2 in (6) is the main concern of LDT. In many cases, c_1 and c_2 can be shown to be equal in the limit $n \to \infty$. As a simple example where this can be done explicitely, consider an i.i.d. sequence $X_1, X_2, \dots$ of Bernoulli random variables: X_i takes values in the finite set $\mathbb{A} = \{0, 1\}$. We also simplify by considering the symmetric case, where $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$, so that $m = \frac{1}{2}$. We fix some $x > m$ and study the probability of having a large deviation $\{\frac{S_n}{n} \geq x\}$. Since

$$P(S_n \geq xn) = \sum_{xn \leq k \leq n} P(S_n = k) = 2^{-n} \sum_{xn \leq k \leq n} \binom{n}{k}, \tag{7}$$

we have

$$H_n(x) 2^{-n} \leq P(S_n \geq xn) \leq (n+1) H_n(x) 2^{-n}, \tag{8}$$

where

$$H_n(x) := \max_{xn \leq k \leq n} \binom{n}{k}.$$

Since $k \to \binom{n}{k}$ is increasing when $k \leq \frac{n}{2}$ and decreasing when $k \geq \frac{n}{2}$, the maximum in $H_n(x)$ is attained for $k = \lceil xn \rceil$, and as can be easily computed using the Stirling Formula,

$$\lim_{n \to \infty} \frac{1}{n} \log H_n(x) = -x \log x - (1-x) \log(1-x).$$

Therefore,

$$\lim_{n \to \infty} \frac{1}{n} \log P(S_n \geq an) = -I(x), \tag{9}$$

where

$$I(x) = \begin{cases} \log 2 - x \log x - (1-x) \log(1-x) & \text{if } x \in [0,1], \\ \infty & \text{if } x \notin [0,1]. \end{cases}$$

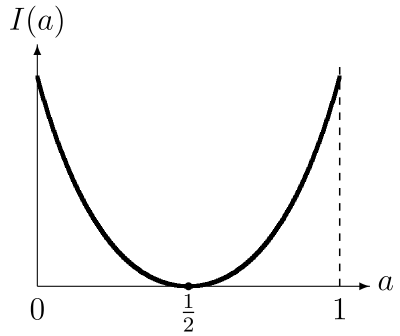

FIGURE 1. The rate function $I(x)$ for Bernoulli variables.

Clearly, $I \geq 0$, it is strictly convex, symmetric around the point $a = \frac{1}{2}$, at which it has its unique minimum: $I(\frac{1}{2}) = 0$. We have actually obtained

$$\lim_{n \to \infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in [a, \infty)\Big) = - \inf_{x \in [a, \infty)} I(x) \tag{10}$$

By the symmetry of I, we can write

$$\lim_{n \to \infty} \frac{1}{n} \log P\Big(\Big|\frac{S_n}{n} - m\Big| \geq \epsilon\Big) = - \inf_{x : |x - m| \geq \epsilon} I(x) = -I(\epsilon) < 0. \tag{11}$$

This is therefore a case where the two constants c_1, c_2 above can be computed exactly, and shown to be equal. They are expressed in terms of a variational problem involving the rate function I.

This result will be generalized to any sequence of i.i.d. random variables in the Theorem of Cramér. Nevertheless, it will not always be possible to obtain I in an explicit form as above, neither will it be possible to obtain the exact limit (10). The general setup will be the following.

Definition 1.1. *A sequence of random variables $Z_1, Z_2, \dots$ satisfies a Large Deviation Principle (LDP) if there exists a lower semicontinuous function $I : \mathbb{R} \to [0, \infty]$ with compact level sets such that*

(1) *for all closed set $F \subset \mathbb{R}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\Big(Z_n \in F\Big) \leq - \inf_{x\in F} I(x) \tag{12}$$

(2) *for all open set $G \subset \mathbb{R}$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(Z_n \in G\Big) \leq - \inf_{x\in G} I(x) \tag{13}$$

It will be seen in the Theorem of Cramér (Section 7 below) that under a finiteness condition on, Λ, the logarithmic moment generating function of X_1, the sequence $\frac{S_n}{n}$ satisfies a LDP with a rate function given by the Legendre transform of Λ.

A LDP also holds (the Theorem of Gärtner-Ellis) in the case where some dependence is introduced among the variables Z_k, which is a typical situation encountered in statistical mechanics. In that case, the rate function cannot be obtained by the distribution of a single variable X_i, but through a limiting process equivalent to the definition of the free energy/pressure in the thermodynamic limit.

The LDP isn't restricted to sequences of $\mathbb{R}$-valued random variables, as the sequence $\frac{S_n}{n}$ above, but can be defined for more general objects living on more abstract metric spaces. These appear naturally, even in the study of real i.i.d. sequences. For example, rather than using $\frac{S_n}{n}$ as a macroscopic observable, a finer description of a large sample is obtained by considering the empirical measure $L_n \in \mathcal{M}_1(\mathbb{R})$, defined by

$$L_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j} \,,$$

where δ_{X_j} is a Dirac mass at X_j. In these terms, the WLLN can be reformulated by saying that L_n converges weakly to the distribution of X_1, ν, in the sense that $\int f(x) L_n(dx) \to \int f(x)\nu(dx)$ for all bounded continuous function $f : \mathbb{R} \to \mathbb{R}$. One can then wonder if some concentration speed for L_n around ν can be obtained. Namely, the Theorem of Sanov (see Section) says that L_n also satisfies a large deviation principle: there exists a convex lower semicontinuous function

$\mathcal{I} : \mathcal{M}_1(\mathbb{R}) \to [0, +\infty]$ such that for all $E \subset \mathcal{M}_1(\mathbb{R})$,

$$- \inf_{\mu \in \mathring{E}} \mathcal{I}(\mu) \leq \liminf_{n\to\infty} \frac{1}{n} \log P\big(L_n \in E\big)$$
$$\leq \limsup_{n\to\infty} \frac{1}{n} \log P\big(L_n \in E\big) \leq - \inf_{\mu \in \overline{E}} \mathcal{I}(\mu)\,.$$

Here, the rate function is in fact given by $\mathcal{I}(\mu) = D(\mu|\nu)$, the relative entropy of μ with respect to ν. The above is called a *LDP of Level* 2. It involves the convergence of measures on $\mathcal{M}_1(\mathbb{R})$ and gives a more precise information than the level 1. Somehow, the Level 1-LDP should follow from the Level 2-LDP. Namely, since

$$\frac{S_n}{n} = \int x\, L_n(dx) \equiv \Phi(L_n)\,,$$

we expect that the concentration of L_n around μ should imply the concentration of $\frac{S_n}{n} = \Phi(L_n)$ around $m = \Phi(\mu)$. This indeed holds, due to the continuity of the map $\Phi : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R}$, as will be seen in the *Contraction Principle*.

A large part of these notes is to make a close link between the rate functions of LDT with the thermodynamic potentials of ESM.

3. Spin systems

CHAPTER 2

The Maximum Entropy Principle

As we saw in the introduction, it is more natural to describe a large system of particles at equilibrium using a probability distribution, rather than to seek for the solution of a system of 10^{25} differential equations. In this section we present a simple procedure which allows to select this probability measure, in the most natural way. The method, called the Maximum Entropy Principle, will select probability measures *a priori*, under certain constraints. Such constraints appear in statistical mechanics, where one studies, for example, systems of particles whose total energy is fixed. This method is very clearly explained in the papers of Jaynes, [**?**] and [**?**]. Later, we use it to introduce the Gibbs distribution, the most widely used in equilibrium statistical mechanics.

The technique presented hereafter does not only apply to systems of particles and has a wide range of applicability. Notice that although it should be considered as fundamental, the Maximum Entropy Principle will find a justification in the large deviation theorems of subsequent chapters.

We start with a simple example. Suppose we are given a dice with 6 faces: throwing the dice results in a random number $X \in \{1, 2, \dots, 6\}$. We put ourselves in the situation where the probabilities $p_i = P(X = i)$ are unknown, and our aim is to associate to this dice a suitable probability distribution $P = (p_1, \dots, p_6)$, $p_i \geq 0$, $\sum_{i=1}^{6} p_i = 1$. We are thus assuming that P *exists*, and our aim is to find it.

Of course, one way of doing, call it *empirical*, is to throw the dice a large number n of times, $X_1, \dots, X_n$, and to count the number of times each face appeared: for each $i \in \{1, 2, \dots, 6\}$,

$$p_i^{(n)} := \frac{\sharp\{1 \leq k \leq n : X_k = i\}}{n}\,.$$

Then the Law of Large Numbers guarantees that when n becomes large, the empirical ratios $p_i^{(n)}$ converge to the true values p_i:

$$(p_1^{(n)}, \dots, p_6^{(n)}) \to (p_1, \dots, p_6) \qquad \text{almost surely.}$$

Unfortunately, we have no time to throw the dice an infinite number of times, and our aim is to find a way of choosing a distribution P *a priori* to any empirical manipulation: our choice must be made taking only into account the information available at hand concerning the dice. It is also important that this association be done by taking *all* available information into account. Assuming that our method leads to a candidate distribution P, the LLN can then be used as a way of *testing* if our choice for P matches with the frequencies $p_i^{(n)}$ observed experimentally.

In general, the problem of choosing a distribution a priori [1] is under determined: the a priori available information usually doesn't allow to determine P uniquely. For example, since the dice has six faces and since the constraint $\sum_{i=1}^{6} p_i = 1$ determines for example p_6 in function of the other p_is, one would need another five conditions in order to determine P uniquely (if these conditions are not contradictory). We are thus faced with the problem of making a *choice* between many possibilities, and the point is to decide which choice is most natural.

The simplest situation is when *no* information is available. In this case, any choice of 6-tuple $(p_1, \ldots, p_6)$ seems possible, as long as $\sum_{i=1}^{6} p_i = 1$. For example,

p_1	p_2	p_3	p_4	p_5	p_6
0	0.1	0.4	0.1	0.1	0.3

Nevertheless, something seems to go wrong with this choice. Namely, the fact that *no information is available* is in itself a piece information (!), in the sense that it obliges us to treat all the possible outcomes in an equivalent way: if nothing indicates a priori that the outcome $X = 3$ is more likely than $X = 2$, then we have no reason to choose $p_3 > p_2$. Therefore, the most reasonable choice, in absence of information, seems to be choosing all the p_is equal. That is, to consider the uniform distribution:

p_1	p_2	p_3	p_4	p_5	p_6
1/6	1/6	1/6	1/6	1/6	1/6

Of course, someone could argue "but what if the true distribution *is* actually given by $(0, 0.1, 0.4, 0.1, 0.1, 0.3)$?" In this case our response is that if this is true, then necessarily some important information is missing and we can therefore not make any reasonable choice for the distribution.

This above reasoning leading us to choose the uniform distribution is called the *Principle of Indifference*: when nothing is known, the choice must be made in order that the probability be spread in the most uniform way among the outcomes. As explained by Penrose [**?**]:

> *[...] the Principle of Indifference, according to which a person who sees no essential difference between two possible alternatives assigns them equal subjective probabilities.*

Assume now that one additional piece of information is known. For example,

$$E[X] = 4\,. \tag{14}$$

(Observe that in the uniform case above, $E[X] = 3.5$.) Here, the presence of a constraint necessarily induces an assymetry among the p_is. Again, an allowed choice satisfying the constraint would be

p_1	p_2	p_3	p_4	p_5	p_6
0	0	0	1	0	0

[1] Jaynes calls these *prior probabilities*, see [**?**].

but again, this choice seems to favorize overwhelmingly the outcome $X = 4$. Rather, one needs to apply something analogous to the Principle of Indifference *among* the set of distributions that satisfy (14). Since the distribution we are looking for is obviously non-uniform, it is not clear how exactly this principle should be applied: the distribution we are looking for must satisfy (14) and be at the same time "closest" to the uniform distribution. Ideally, one needs to seek for a *function* quantifying the spread of a distribution, which should be maximal when the distribution is uniform.

It might be surprising to learn that there is essentially a unique way of defining this function, introduced by Shannon in 1948. This function associates to $(p_1, \dots, p_k)$ a number $H = H(p_1, \dots, p_k)$ called *entropy*. H has many interpretations, but its main feature is that it allows to measure, in some sense (which will be made clearer when introducing relative entropy), the distance to the uniform distribution. The spread of a distribution is larger when the distribution is close to uniform. This can be expressed by saying that the spreading of the distribution turns the outcome of a realization more *uncertain*. Entropy therefore provides a way of measuring our uncertainty with respect to the outcome of a random experiment. Citing again Jaynes [**?**],

> *Our problem is to find a probability assignment which avoids bias, while agreeing with whatever information is given. [...] The great advance provided by Information Theory lies in the discovery that there is a unique, unambiguous criterion for the amount of uncertainty represented by a discrete probability distribution, which agrees with our intuition that a broad distribution represents more uncertainty than does a sharply peaked one.*

The following definition of entropy was given by C.E. Shannon in [**?**].

Definition 2.1. *The* entropy *of a probability distribution* $(p_1, \dots, p_k)$ *is defined by*

$$H(p_1, p_2, \dots, p_k) := -\sum_{j=1}^{k} p_j \log p_j \,, \tag{15}$$

where it is assumed that the logarithm is with respect to the base e, and where we make the convention that $0 \log 0 := 0$.

Let us verify that this definition suits our requirements for a function measuring our ignorance with respect to the outcome of the random experiment. First, observe that H is a positive quantity which attains its minimal value $H = 0$ exactly when all but one p_is are zero, and the last one equals 1. In any of these cases, the outcome of the experiment is certain, and so the uncertainty must be zero.

Then, we show that entropy is maximal for the uniform distribution:

$$H(p_1, \dots, p_k) \leq H\left(\tfrac{1}{k}, \dots, \tfrac{1}{k}\right) = \log k \,. \tag{16}$$

Namely, consider the strictly concave function $\psi(x) := -x \log x$ for $x \in (0,1]$, $\psi(0) := 0$. We have

$$\begin{aligned} H(p_1, \dots, p_k) = \sum_{j=1}^{k} \psi(p_j) &= k \sum_{j=1}^{k} \tfrac{1}{k} \psi(p_j) \\ &\leq k\psi(\tfrac{1}{k}) = \log k = H\big(\tfrac{1}{k}, \dots, \tfrac{1}{k}\big) . \end{aligned}$$

Therefore, (16) fullfills our previous requirement: the uncertainty with respect to the outcome of the experience is maximal when the distribution is uniform.. Moreover, in Shannon's own words, *any change towards equalization of the probabilities* $(p_1, \dots, p_k)$ *increases* $H(p_1, \dots, p_k)$. This can be seen by explicit calculation, by considering the variation of H when, say $p_1(s) = p_0 + s$, $p_2(s) = p_0 - s$, with $s \searrow 0$, and where the other $k-2$ variables are kept fixed:

$$\begin{aligned} \frac{d}{ds} H &= \frac{d}{ds} \big[-(p_0 + s) \log(p_0 + s) - (p_0 - s) \log(p_0 - s) \big] \\ &= \log \frac{p_0 - s}{p_0 + s} , \end{aligned}$$

which is < 0 when $s > 0$. This means that if s decreases, i.e. when p_1 and p_2 tend to equalize, then the entropy increases.

A further property of H is that it is a concave function of $\mathbf{p} = (p_1, \dots, p_k)$. To gain geometric intuition, we can identify the set of probability distributions $(p_1, \dots, p_k)$ with the simplex

$$\mathcal{M}_1 := \Big\{ \mathbf{p} = \sum_{j=1}^{k} p_j \mathbf{e}_j : p_j \geq 0, \sum_{j=1}^{k} p_j = 1 \Big\} \subset \mathbb{R}^k .$$

Any $\mathbf{p} \in \mathcal{M}_1$ can thus be considered as a convex combination of the extreme elements of $\mathcal{M}_1$, which are the unit vectors of the canonical basis of $\mathbb{R}^k$: $\mathbf{e}_1, \dots, \mathbf{e}_k$. At each extreme element $\mathbf{e}_j$, $H(\mathbf{e}_j) = 0$. By the concavity of ψ, for any $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{M}_1$, $0 \leq \lambda \leq 1$,

$$H(\lambda \mathbf{p}_1 + (1-\lambda)\mathbf{p}_2) \geq \lambda H(\mathbf{p}_1) + (1-\lambda) H(\mathbf{p}_2) .$$

One thus has a picture of H as a concave function on $\mathcal{M}_1$ which attains its maximum $(\log k)$ at the barycenter of $\mathcal{M}_1$, namely $(\frac{1}{k}, \dots, \frac{1}{k})$.

Going back to our problem of selecting a probability distribution under a constraint, (Jaynes [**?**])

> *In making inference on the basis of partial information we must use that probability distribution which has the maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information, which by hypothesis we don't have.*

Consider the dice problem. Following Jaynes, the proper unbiased probability distribution is the one that maximises $H(p_1, \dots, p_6)$, with possible additional constraints. When no constraint is fixed (other than $(p_1,\dots,p_6)$ being a probability), the problem then reduces to

Maximise $H(p_1,\dots,p_6)$ over $\mathcal{M}_1$.

As we have seen, the unique solution to this problem is the uniform probability $(\frac{1}{6},\dots,\frac{1}{6})$. Now if one imposes that $E[X]=4$, then with $\mathcal{M}_1' := \{(p_1,\dots,p_6)\in\mathcal{M}_1 : \sum_{j=1}^6 jp_j = 4\}$, the problems becomes:

Maximise $H(p_1,\dots,p_6)$ over $\mathcal{M}_1'$.

This optimization problem can be solved using the method of Lagrange multipliers. Since there are two constraints, we introduce two Lagrange multipliers λ,β, and define

$$L(p_1,\dots,p_6) := H(p_1,\dots,p_6) - \lambda\sum_{i=1}^6 p_i - \beta\sum_{i=1}^6 ip_i\,.$$

The optimization problem then turns into the analytic resolution of the system

$$\begin{cases} \nabla L = 0\,, \\ \sum_{i=1}^6 p_i = 1\,, \\ \sum_{i=1}^6 ip_i = 4\,. \end{cases}$$

As can be easily verified using the Lagrange multipliers , the solution of this problem is given by

$$p_i = \frac{e^{-\beta_* i}}{Z(\beta_*)}\,,$$

where

$$Z(\beta) = \sum_{i=1}^6 e^{-\beta i}\,,$$

and where β_* denotes the unique solution of

$$-\frac{d}{d\beta}\log Z(\beta) = 4\,.$$

The solution $(p_1,\dots,p_6)$ is therefore the solution of the variational problem

$$\sup\{H(p) : p\in\mathcal{M}_1,\ E_p[X]=4\}\,.$$

As can be verified numerically, the result is, within ± 0.01,

p_1	p_2	p_3	p_4	p_5	p_6
$\simeq 0.10$	$\simeq 0.12$	$\simeq 0.15$	$\simeq 0.17$	$\simeq 0.21$	$\simeq 0.25$

1. The Gibbs Distribution

We reformulate the above setting in a slightly different way, by applying the Maximum Entropy Principle to the study of a finite system encountered constantly in statistical mechanics. Assume a physical system can be in a finite number of states, denoted $\omega\in\Omega$, $|\Omega|<\infty$. It is typical, depending on the situation under consideration, to assume that some physical observables of the system have fixed average values. For example, these mean values can be determined by some experimental restrictions. Assume therefore that a function (observable) $U:\Omega\to\mathbb{R}$

is given. Although it is arbitrary, in most cases $U(\omega)$ represents the energy of the microscopic state ω. If no other information is given , what is the most natural (or "least biased") probability distribution $\mu = (\mu(\omega), \omega \in \Omega)$, satisfying the constraint that $E_\mu[U] = E_0$? (Here, $E_0 \in (\min U, \max U)$.) As before, the Maximum Entropy Principle leads to the following optimisation problem: maximise the Shannon Entropy

$$H(\mu) = -\sum_{\omega\in\Omega} \mu(\omega)\log\mu(\omega)\,,$$

under the constraints

$$\mu(\omega) \geq 0, \quad \sum_{\omega\in\Omega}\mu(\omega) = 1\,, \quad \sum_{\omega\in\Omega}\mu(\omega)U(\omega) = E_0\,.$$

As we saw, by a direct application of the method of the Lagrange multipliers, the solution is obtained by first finding the Lagrange multiplier $\beta_* = \beta_*(E_0)$, solution of

$$-\frac{d}{d\beta}\log Z(\beta) = E_0\,,$$

where $Z(\beta)$ is the partition function, defined by

$$Z(\beta) := \sum_{\omega\in\Omega} e^{-\beta U(\omega)}\,.$$

(As can be verified, β_* exists as soon as $E_0 \in (\min U, \max U)$.) Then, the maximiser $\mu_* = (\mu_*(\omega), \omega \in \Omega)$ of the Shannon Entropy is given by

$$\mu_*(\omega) = \frac{e^{-\beta_* U(\omega)}}{Z(\beta_*)}\,, \forall \omega \in \Omega\,. \tag{17}$$

The measure μ_* has thus been constructed so that $E_{\mu_*}[U] = E_0$, and its Shannon Entropy is maximal among all measures μ satisfying this condition. By a direct computation,

$$H(\mu_*) = \beta_* E_{\mu_*}[U] + \log Z(\beta_*) = \beta_* E_0 + \log Z(\beta_*)\,.$$

Therefore,

$$\begin{aligned}
\frac{\partial H}{\partial E_0} &= E_0\frac{\partial \beta_*}{\partial E_0} + \beta_* + \frac{\partial}{\partial E_0}\log Z(\beta_*) \\
&= E_0\frac{\partial \beta_*}{\partial E_0} + \beta_* + \underbrace{\frac{\partial}{\partial \beta}\log Z(\beta)\Big|_{\beta=\beta_*}}_{=-E_0}\frac{\partial \beta_*}{\partial E_0} \\
&= \beta_*\,.
\end{aligned} \tag{18}$$

Assume for a while that $U(\omega)$ is the energy of the state ω. Then we can compare this last display with the fundamental thermodynamic relation between entropy S, internal energy E and temperature T:

$$\frac{\partial S}{\partial E} = \frac{1}{T}\,.$$

Comparing this with (18), and denoting the dependence of $H(\mu_*)$ on E_0 by: $S(E_0) = H(\mu_*)$, we can therefore interpret the Lagrange multiplier β_* as an inverse temperature:

$$\beta_* \equiv \frac{1}{T}\,. \tag{19}$$

We will not pursue this delicate comparison [2], but rather consider it natural, from now on, to study probability distributions of the form (17), but where the parameter β_* is free, not necessarily associated to the solution of a entropy maximisation problem. For physical reasons, we will only consider this free parameter as non-negative.

DEFINITION 2.2. *Let Ω be a finite set. If $U : \Omega \to \mathbb{R}$ and $\beta > 0$, then the Gibbs distribution with potential U and inverse temperature β is the distribution μ_β on Ω defined by*

$$\mu_\beta(\omega) := \frac{e^{-\beta U(\omega)}}{Z(\beta)}\,, \forall \omega \in \Omega\,, \tag{20}$$

where the normalizing factor $Z(\beta) := \sum_{\omega\in\Omega} e^{-\beta U(\omega)}$ is called partition function.

Since we sometimes need to consider μ_β as a real probability measure, i.e. an element of $\mathcal{M}_1(\Omega)$, we will write it as

$$\mu_\beta = \sum_{\omega\in\Omega} \mu_\beta(\omega)\delta_\omega\,. \tag{21}$$

The Gibbs distribution is one of the main themes of these notes. We will see that it also appears naturally as a limiting distribution for a sequence of conditionned measures.

2. Uniqueness of Entropy

In this section, following Khinchin [?], we show that the entropy H defined in (15) is the unique function satisfying a set of natural conditions that are suggested by the intuitive notion of *uncertainty* about the outcome of a random experience, or its *unpredictability*.

It is useful to formulate the problem in a slightly more general manner. Consider a random experiment modelized by some probability space $(\Omega, \mathcal{F}, P)$. A finite scheme is a partition $\mathcal{A}$ of Ω into a finite number of sets $A_k \in \mathcal{F}$ (called atoms) together with their associated probabilities $p_k := P(A_k)$. Altough we will usually denote a scheme by $\mathcal{A} = (A_1, \dots, A_n)$, it should always be remembered that the probabilities $(P(A_1), \dots, P(A_n))$ are part of the information contained in $\mathcal{A}$. A scheme can represent a simple experiment, like the throw of a dice (where $A_k = \{\text{the } k\text{th face shows up}\}$), but it can also modelize the coarse-graining of a more complicated experiment where instead of the true result of the random experiment, i.e. ω, one is interested in the atom A_k of the partition $\mathcal{A}$ to which ω belongs. The coarse-grained result of the experience is therefore an index $k = k(\omega) \in \{1, 2, \dots, n\}$, giving the unique atom $A_k \ni \omega$, and can be considered

[2] See [?] where this identification is made in details.

as partial knowledge about the result of the experiment.

How can one define the unpredictability of a random experience modelized by a finite scheme $\mathcal{A} = (A_1, \ldots, A_n)$? The unpredictability of $\mathcal{A}$ can of course depend only on the probabilities $(P(A_1), \ldots, P(A_n))$. That is, we are looking for a class of functions

$$H(\mathcal{A}) = H(P(A_1), \ldots, P(A_n)) .$$

We will define four conditions (see (I)-(IV) below) that H should satisfy, most of which will be natural in terms of unpredictability [3], and then show that these lead unambiguously to the only possibility $H = H_{\mathrm{Sh}}$ (up to a multiplicative constant), where according to Definition 2.1, the Shannon Entropy H_{Sh} is defined by

$$H_{\mathrm{Sh}}(\mathcal{A}) = -\sum_{A \in \mathcal{A}} P(A) \log P(A) . \tag{22}$$

A first condition we should impose is that the unpredictability of a random experience be continuous in its arguments:

$$H \text{ is continuous.} \tag{I}$$

Next, as we have seen for H_{Sh}, we need that unpredictability be maximal when the distribution is uniform. Let us call a scheme uniform when its atoms have equal probabilities. The second requirement we impose on H is therefore that

$$H \text{ is maximal on uniform schemes.} \tag{II}$$

To state the third requirement, we introduce a few notations. Consider two schemes, say $\mathcal{A} = (A_1, \ldots, A_k)$ and $\mathcal{B} = (B_1, \ldots, B_k)$, giving a coarse-grained description of the same random experiment. Consider the composite scheme $\mathcal{A} \vee \mathcal{B}$ defined by

$$\mathcal{A} \vee \mathcal{B} := \{A \cap B : A \in \mathcal{A}, B \in \mathcal{B}\} .$$

This scheme is a refinement of both $\mathcal{A}$ and $\mathcal{B}$, since each of its atoms contains a double information ($\omega \in A$ *and* $\omega \in B$ for some couple (A, B)). It is easy to see that when $\mathcal{A}$ and $\mathcal{B}$ are independent, that is if $P(A \cap B) = P(A)P(B)$ for all pair $A \in \mathcal{A}$, $B \in \mathcal{B}$, then

$$H_{\mathrm{Sh}}(\mathcal{A} \vee \mathcal{B}) = H_{\mathrm{Sh}}(\mathcal{A}) + H_{\mathrm{Sh}}(\mathcal{B}) . \tag{23}$$

This property is called extensivity. In the general case (i.e. without assuming independence) we always have

$$H_{\mathrm{Sh}}(\mathcal{A} \vee \mathcal{B}) \geq -\sum_{(A,B)} P(A \cap B) \log P(A) = H_{\mathrm{Sh}}(\mathcal{A}) .$$

The exact excess in the previous inequality,

$$H_{\mathrm{Sh}}(\mathcal{B}|\mathcal{A}) := H_{\mathrm{Sh}}(\mathcal{A} \vee \mathcal{B}) - H_{\mathrm{Sh}}(\mathcal{A}) , \tag{24}$$

[3] The only exception being maybe (IV), which has to do with extensivity; this condition seems natural to me although I understand some might find it completely arbitrary.

measures the average excess of information produced by the outcome of the experiment $\mathcal{A} \vee \mathcal{B}$ over the information produced by the experiment $\mathcal{A}$. For any event A, $P(A) > 0$, define

$$H_{\mathrm{Sh}}(\mathcal{B}|A) := -\sum_{B\in\mathcal{B}} P(B|A)\log P(B|A)\,.$$

As can be verified explicitely,

$$H_{\mathrm{Sh}}(\mathcal{B}|\mathcal{A}) = \sum_{A\in\mathcal{A}} P(A) H_{\mathrm{Sh}}(\mathcal{B}|A)\,. \tag{25}$$

The above quantities can be defined in general for finite schemes:

$$H(\mathcal{B}|A) := H(P(B_1|A),\dots,P(B_k|A))\,,$$

and

$$H(\mathcal{B}|\mathcal{A}) := \sum_{A\in\mathcal{A}} P(A) H(\mathcal{B}|A)\,.$$

A fundamental property that should be satisfied by entropy is therefore that

$$H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}|\mathcal{A})\,. \tag{III}$$

The interpretation of this identity in terms of unpredictability is clear. The final condition says that unpredictability is not affected by the presence of atoms whose probability is zero:

> If $\mathcal{A} = (A_1,\dots,A_k)$ with $P(A_k) = 0$, and $\mathcal{A}' = (A_1,\dots,A_{k-2},A'_{k-1})$, where $A'_{k-1} = A_{k-1}\cup A_k$, then $H(\mathcal{A}) = H(\mathcal{A}')$. (IV)

Theorem 2.1. *Let H be a function defined on probability distributions of finite schemes, such that* (I)-(IV) *hold. Then $H = H_{\mathrm{Sh}}$, up to a positive constant.*

We first prove Theorem 2.1 for uniform schemes, which leads to (26), the Entropy of Boltzmann in its simplest form. If $\mathcal{U}$ is a scheme, $|\mathcal{U}|$ denotes the number of elements of its partition.

Lemma 2.1. *Let H be a function defined on uniform finite schemes which is monotone increasing in $|\mathcal{U}|$, and which is extensive, in the sense that if $\mathcal{U}$, $\mathcal{U}'$ are two independent uniform schemes, then*

$$H(\mathcal{U}\vee\mathcal{U}') = H(\mathcal{U}) + H(\mathcal{U}')\,.$$

Then there exists a constant $\lambda > 0$ such that for all uniform scheme $\mathcal{U}$,

$$H(\mathcal{U}) = \lambda \log|\mathcal{U}|\,. \tag{26}$$

Proof. Let $L(k)$ denote the value of H on schemes of size k. By the first assumption, $L(k)$ is non-decreasing. Let $\mathcal{U}_1,\dots,\mathcal{U}_m$ be independent, containing each r elements. On one hand, by extensivity,

$$H(\mathcal{U}_1\vee\cdots\vee\mathcal{U}_m) = \sum_{j=1}^{m} H(\mathcal{U}_j) = mL(r).$$

On the other hand, $\mathcal{U}_1\vee\cdots\vee\mathcal{U}_m$ is also uniform and contains r^m elements. Therefore, $H(\mathcal{U}_1\vee\cdots\vee\mathcal{U}_m) = L(r^m)$. This shows that $L(r^m) = mL(r)$. We then check that the logarithm is the unique function satisfying this condition for all $m \geq 1$, $r \geq 1$. Namely, fix any $s \geq 1$, $n \geq 1$ and find some $m \geq 1$ such that $r^m \leq s^n < r^{m+1}$. Therefore, $m \log r \leq n \log s < (m+1) \log r$, and since L is monotone and $L(s^n) = nL(s)$, we have

$$\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m+1}{n} \, .$$

Altogether, these imply that

$$\left| \frac{L(s)}{L(r)} - \frac{\log s}{\log r} \right| \leq \frac{1}{n} \, .$$

Since n was arbitrary, this shows that there exists λ such that $L(r) = \lambda \log r$. λ must be positive in order for L to be monotone increasing. □

Proof of Theorem 2.1: Assume H satisfies (I)-(IV). For a while, let H_k denote the function associated to schemes with k elements. Then clearly

$$H_k(\tfrac{1}{k}, \dots, \tfrac{1}{k}) \overset{\text{(IV)}}{=} H_{k+1}(\tfrac{1}{k}, \dots, \tfrac{1}{k}, 0) \overset{\text{(II)}}{\leq} H_{k+1}(\tfrac{1}{k+1}, \dots, \tfrac{1}{k+1}, \tfrac{1}{k+1}) \, ,$$

which implies that for uniform schemes, H is increasing of the number of elements. By (III), H is of course extensive for independent partitions. H therefore satisfies the conditions of the lemma, and so $H(\mathcal{U}) = \lambda \log |\mathcal{U}|$ for all uniform scheme. Let then $\mathcal{A} = (A_1, \dots, A_n)$ be a scheme with rational probabilities p_k (the extension to arbitrary probabilities p_k then follows by the continuity hypothesis (I)). That is, $p_k = \frac{w_k}{W}$, where $w_k \in \mathbb{N}$ and $W = \sum_{k=1}^n w_k$. It is useful to reinterpret the probabilities as follows. Consider a set of W balls numbered from 1 to W, w_1 of which are painted with a color c_1, w_2 of which are painted with a color c_2, etc. (we assume all the colors are different). A ball is drawn at random, uniformly. Then if we define A'_j as the even "the ball drawn has color c_j", then $P(A'_j) = \frac{\omega_j}{W} \equiv P(A_j)$. Therefore, $H(\mathcal{A}) = H(\mathcal{A}')$, where $\mathcal{A}' = (A'_1, \dots, A'_n)$. Let $\mathcal{B} = (B_1, \dots, B_W)$ be made of the smallest possible atoms of the experience: B_i is the event "the ball i was drawn". Then clearly,

$$P(B_i | A'_j) = \begin{cases} \frac{1}{w_j} & \text{if } i \in A'_j \\ 0 & \text{otherwise.} \end{cases}$$

By the lemma, $H(\mathcal{B}|A'_j) = \lambda \log w_j$. Therefore,

$$H(\mathcal{B}|\mathcal{A}') = \sum_{j=1}^{n} p_j H(\mathcal{B}|A'_j) = \lambda \sum_{j=1}^{n} p_j \log p_j + \lambda \log W$$

But since $\mathcal{A}' \vee \mathcal{B}$ contains events of equal probabilities (W of which are non-zero), we have again by (IV) and the lemma that $H(\mathcal{A}' \vee \mathcal{B}) = \lambda \log W$. Therefore, by assumption (III),

$$H(\mathcal{A}') = H(\mathcal{A}' \vee \mathcal{B}) - H(\mathcal{B}|\mathcal{A}') = -\lambda \sum_{j=1}^{n} p_j \log p_j \, .$$

Since $H(\mathcal{A}) = H(\mathcal{A}')$, this finishes the proof. □

CHAPTER 3

Large Deviations for finite alphabets

In this chapter, we expose large deviation result for i.i.d. sequences $X_1, X_2, \ldots$ taking values in a finite set. The advantage of working with finite alphabets is that it allows to rely on combinatorial rather than analytical methods, and already gives a framework in which many interesting statistical mechanical systems can be studied, for example the ideal gas of the next section. On the other hand, the results obtained are already far-reaching, and illustrate the main concepts that will be encountered later in more general settings.

This chapter will contain two kinds of results: first, those describing the concentration properties of the empirical mean $\frac{S_n}{n}$, called Level-1 LDPs. Second, those describing the concentration of the empirical measure L_n, called Level-2 LDPs.

1. The Theorem of Sanov

Note that in the case where the variables X_j are Bernoulli, i.e. taking values in the alphabet $\mathbb{A} = \{0, 1\}$, we have obtained in (10), Chapter 1, a precise asymptotic behaviour of the frequency of 1s, i.e. $\frac{S_n}{n}$. Since the alphabet contains only two letters, this of course automatically gives the frequency of 0s, equal to $1 - \frac{S_n}{n}$. The LLN thus implies that the *empirical measure* $(\frac{S_n}{n}, 1 - \frac{S_n}{n})$, which give the frequency of each symbol in a sample of size n, converges to its theoretical value $(\frac{1}{2}, \frac{1}{2})$, exponentially fast with n. In the general case, where the alphabet can contain more than two letters, the combinatorics is a little more complicated, but a similar result holds, known as the Theorem of Sanov.

In this section, we study a more general situation, in which the random variables X_j take values in a finite alphabet $\mathbb{A}$. Let $\mathcal{M}_1 = \mathcal{M}_1(\mathbb{A})$ denote the set of all probabilty distributions on $\mathbb{A}$, which we identify with the vectors $\mu = (\mu(a), a \in \mathbb{A})$ with $\mu(a) \geq 0$, $\sum_{a \in \mathbb{A}} \mu(a) = 1$. We equip $\mathcal{M}_1$ with the L^1-metric, (also called total variation distance), defined by

$$\|\mu - \nu\|_1 := \sum_{a \in \mathbb{A}} |\mu(a) - \nu(a)| \,.$$

Teh interior of a set $\mathcal{E} \subset \mathcal{M}_1$ is denoted $\text{int}\mathcal{E}$, and its closure $\overline{\mathcal{E}}$. We will denote the common distribution of the X_js by ν: $\nu(a) := P(X_j = a)$. Without loss of generality, we can assume that $\nu(a) > 0$ for all $a \in \mathbb{A}$. We also assume that the sequence $X_1, X_2, \ldots$ is constructed canonically on the sequence space $\mathbb{A}^{\mathbb{N}}$, endowed with the σ-algebra generated by cylinders, and denote the product measure on this space by P_ν.

Rather than just $\frac{S_n}{n}$, we study the frequency of appearance of each of the symbols of the alphabet up to time n. Therefore, the empirical measure associated to a finite sample $X_1, \dots, X_n$, is the probability distribution $L_n \in \mathcal{M}_1$ defined by

$$L_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j} \,.$$

L_n is also called the type of the sample $X_1, \dots, X_n$. In terms of L_n, the Law of Large Numbers takes the following form: if $\nu \in \mathcal{M}_1$ denotes the common distribution of the X_is, then

$$L_n \Rightarrow \nu\,, \quad P_\nu\text{-almost surely} \tag{27}$$

where $\Rightarrow$ denotes weak convergence, which in the case of a finite alphabet reduces to $\|L_n - \nu\|_1 \to 0$.

The Theorem of Sanov describes the exponential concentration of L_n in the vincinity of ν. The cost of the empirical measure being far from ν will be measured with the following function.

Definition 3.1. *Let μ and ν be two probability distributions on $\mathbb{A}$. The relative entropy of μ with respect to ν, is defined by*

$$D(\mu\|\nu) := \sum_{a\in\mathbb{A}} \mu(a)\log\frac{\mu(a)}{\nu(a)}\,.$$

Our convention is that $0\log 0 := 0$, *and* $0 \log 0 := \frac{0}{0}$.

D is also called Kullback-Leibler distance, or information divergence between μ and ν. Observe that if μ is not absolutely continuous with respect to ν (i.e. if there exists $a \in \mathbb{A}$ such that $\mu(a) > 0$, $\nu(a) = 0$), then $D(\mu\|\nu) = +\infty$. We list a few properties of $D(\cdot\|\cdot)$.

Proposition 3.1. *Assume $\nu > 0$. D satisfies the following properties:*

1. *$D(\mu\|\nu) \geq 0$, with equality if and only if $\mu = \nu$,*
2. *$\mu \mapsto D(\mu\|\nu)$ is strictly convex and continuous on $\mathcal{M}_1$,*
3. *$(\mu, \nu) \mapsto D(\mu\|\nu)$ is convex.*

Proof. Write, temporarily,

$$D(\mu\|\nu) = \sum_{a\in\mathbb{A}} \nu(a)\psi\Big(\frac{\mu(a)}{\nu(a)}\Big)\,,$$

where ψ is the strictly convex function $\psi(x) = x\log x$. Jensen's inequality gives $D(\mu\|\nu) \geq \psi(1) = 0$, with equality if and only if $\mu(a)/\nu(a) = 1$ for all $a \in \mathbb{A}$, thus proving (1). For (2), the strict convexity and continuity of $D(\cdot|\nu)$ follow by that of ψ. For (3), we use the following (called log-concave inequality): if $a_1, \dots, a_n \geq 0$, $b_1, \dots, b_n \geq 0$,

$$\sum_{j=1}^{n} a_j \log\frac{a_j}{b_j} \geq \Big(\sum_{j=1}^{n} a_j\Big)\log\frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}\,,$$

with equality if and only if there exists a constant c such that $a_j = cb_j$ for all j. Therefore if $(\mu', \nu') = \sum_{i=1}^n \lambda_i(\mu_i, \nu_i)$ with $\sum_{i=1}^n \lambda_i = 1$, we have

$$\begin{aligned} D(\mu'\|\nu') &= \sum_{a\in\mathbb{A}} \Big(\sum_{i=1}^n \lambda_i\mu_i(a)\Big) \log \frac{\sum_{i=1}^n \lambda_i\mu_i(a)}{\sum_{i=1}^n \lambda_i\nu_i(a)} \\ &\leq \sum_{a\in\mathbb{A}} \sum_{i=1}^n \lambda_i\mu_i(a) \log \frac{\lambda_i\mu_i(a)}{\lambda_i\nu_i(a)} \\ &= \sum_{i=1}^n \lambda_i D(\mu_i\|\nu_i)\,, \end{aligned}$$

with equality if and only if $\mu_i = \nu_i$ for all i. □

Although these properties remind those of a distance, D is actually *not* a distance: it is not symetric in general, as can be seen by constructing a simple example on an alphabet with two letters. Nevertheless, we have a usefull comparison with the total variation distance:

PROPOSITION 3.2 (Pinsker's Inequality). *For all* $\mu, \nu \in \mathcal{M}_1$,

$$D(\mu\|\nu) \geq \frac{1}{2}\|\mu - \nu\|_1^2 \tag{28}$$

Therefore, if μ and ν are close in the sense of $D(\cdot\|\cdot)$, they are also close in the sense of $\|\cdot\|_{\mathrm{TV}}$.

Observe that if ν is the uniform distribution, $\nu(a) = 1/|\mathbb{A}|$, then

$$D(\mu\|\nu) = \log|\mathbb{A}| - H(\mu)\,. \tag{29}$$

Then (2) implies that $\mu \to H(\mu)$ is concave, as we already knew, but also that in this case, D measures the same discrepancy as the Shannon Entropy does with respect to the uniform distribution. Relative entropy can be thus be considered as a generalization of the Shannon Entropy to the case where the reference measure is ν rather than the uniform distribution.

We now state the theorem.

THEOREM 3.1 (Theorem of Sanov). *Let* $X_1, X_2, \ldots$ *be i.i.d. with common distribution* ν. *For all* $\mathcal{E} \subset \mathcal{M}_1$,

$$P_\nu(L_n \in \mathcal{E}) \leq (n+1)^{|\mathbb{A}|} \exp\big(-n \inf_{\mu\in\mathcal{E}} D(\mu\|\nu)\big)\,. \tag{30}$$

Moreover,

$$-\inf_{\mu\in\mathrm{int}\mathcal{E}} D(\mu\|\nu) \leq \liminf_{n\to\infty} \frac{1}{n}\log P_\nu(L_n \in \mathcal{E}) \tag{31}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n}\log P_\nu(L_n \in \mathcal{E}) \leq -\inf_{\mu\in\mathcal{E}} D(\mu\|\nu)\,. \tag{32}$$

REMARK 3.1. A particular feature of the upper bound (32) is that it does not impose any restriction on the set $\mathcal{E}$. Nevertheless, if $\mathcal{E}$ is open, or more generally, if $\mathcal{E} \subset \overline{\text{int}\mathcal{E}}$ (this forbids $\mathcal{E}$ to have isolated points, in particular), then

$$\inf_{\mu\in\text{int}\mathcal{E}} D(\mu\|\nu) \geq \inf_{\mu\in\mathcal{E}} D(\mu\|\nu) \geq \inf_{\mu\in\overline{\text{int}\mathcal{E}}} D(\mu\|\nu) = \inf_{\mu\in\text{int}\mathcal{E}} D(\mu\|\nu)$$

The last equality follows from the continuity of $D(\cdot\|\nu)$. Therefore the two bounds in the theorem coincide, and:

$$\lim_{n\to\infty} \frac{1}{n} \log P_\nu(L_n \in \mathcal{E}) = -\inf_{\mu\in\mathcal{E}} D(\mu\|\nu)\,.$$

For later reference, we say that the open sets have the $D(\cdot\|\nu)$-*continuity* property.

The proof of Theorem 3.5 relies on simple combinatorial arguments. Denote by $\mathcal{L}_n$ the set of all possible types of sequences of size n. For example, if $|\mathbb{A}| = 2$, $\mathcal{L}_n = \{(0,1), (\frac{1}{n}, \frac{n-1}{n}), \dots, (\frac{n-1}{n}, \frac{1}{n}), (1,0)\}$.

LEMMA 3.1. $|\mathcal{L}_n| \leq (n+1)^{|\mathbb{A}|}$.

PROOF. Each $L \in \mathcal{L}_n$ can be identified with an n-tuple $(l_1, \dots, l_{|\mathbb{A}|})$, with $l_j \in \{0,1,2,\dots,n\}$, $\sum_j l_j = n$. But without this last constraint, the number of such n-tuples is exactly $(n+1)^{|\mathbb{A}|}$. □

If $x = (x_1, \dots, x_n) \in \mathbb{A}^n$, we denote its type by L_x, and write $P_\nu(x) := P_\nu(X_1 = x_1, \dots, X_n = x_n)$.

LEMMA 3.2. *For all $x \in \mathbb{A}^n$, $P_\nu(x)$ depends only on the type of x, and*

$$P_\nu(x) = \exp\left[-n(H(L_x) + D(L_x\|\nu))\right]. \tag{33}$$

(33) nicely expresses the fact that $D(L_x\|\nu)$ represents the cost of L_x being different from ν.

PROOF. First, $H(L_x) + D(L_x\|\nu) = -\sum_{a\in\mathbb{A}} L_x(a) \log \nu(a)$. Then, by independence,

$$\begin{aligned} P_\nu(x) = \prod_{i=1}^n \nu(x_i) &= \prod_{a\in\mathbb{A}} \nu(a)^{|\{i\leq n : x_i = a\}|} \\ &= \prod_{a\in\mathbb{A}} \nu(a)^{nL_x(a)} = e^{n\sum_{a\in\mathbb{A}} L_x(a)\log\nu(a)}\,, \end{aligned}$$

which gives (33). □

If $\mu \in \mathcal{L}_n$, we denote the type class of μ as $\mathbb{A}^n(\mu) := \{x \in \mathbb{A}^n : L_x = \mu\}$. We have $|\mathbb{A}^n| = |\mathbb{A}|^n = e^{n\log|\mathbb{A}|}$, but

LEMMA 3.3. *For all type $\mu \in \mathcal{L}_n$,*

$$(n+1)^{-|\mathbb{A}|} e^{nH(\mu)} \leq |\mathbb{A}^n(\mu)| \leq e^{nH(\mu)}\,. \tag{34}$$

Since the Shannon Entropy is maximal for the uniform distribution, (34) says that the uniform distribution also has the largest type class.

PROOF. For the upper bound, we use Lemma 3.2:

$$1 \geq P_\mu(\mathbb{A}^n(\mu)) = \sum_{x\in\mathbb{A}^n(\mu)} e^{-nH(\mu)} = |\mathbb{A}^n(\mu)|e^{-nH(\mu)}\,.$$

For the lower bound, we use Lemma 3.1:

$$1 = \sum_{\mu'\in\mathcal{L}_n} P_\mu(\mathbb{A}^n(\mu')) \leq |\mathcal{L}_n| \max_{\mu'\in\mathcal{L}_n} P_\mu(\mathbb{A}^n(\mu'))\,.$$

We claim that $\max_{\mu'\in\mathcal{L}_n} P_\mu(\mathbb{A}^n(\mu')) \leq P_\mu(\mathbb{A}^n(\mu))$, which will give the lower bound since $P_\mu(\mathbb{A}^n(\mu)) = |\mathbb{A}^n(\mu)|e^{-nH(\mu)}$, as already seen.

$$\frac{P_\mu(\mathbb{A}^n(\mu))}{P_\mu(\mathbb{A}^n(\mu'))} = \frac{|\mathbb{A}^n(\mu)|\prod_{a\in\mathbb{A}}\mu(a)^{N(a)}}{|\mathbb{A}^n(\mu')|\prod_{a\in\mathbb{A}}\mu(a)^{N'(a)}}\,,$$

where $N(a) = n\mu(a)$, $N'(a) = n\mu'(a)$. Since, by a simple combinatorial argument,

$$|\mathbb{A}^n(\mu)| = \frac{n!}{\prod_{a\in\mathbb{A}} N(a)!}\,,$$

we have

$$\frac{P_\mu(\mathbb{A}^n(\mu))}{P_\mu(\mathbb{A}^n(\mu'))} = \prod_{a\in\mathbb{A}}\Big\{\frac{N'(a)!}{N(a)!}\mu(a)^{N(a)-N'(a)}\Big\}$$

Since $\frac{m!}{n!} \geq n^{m-n}$ for all m, n, this product is bounded below by

$$\prod_{a\in\mathbb{A}} N(a)^{N'(a)-N(a)}\mu(a)^{N(a)-N'(a)} = \prod_{a\in\mathbb{A}} n^{N'(a)-N(a)} = 1\,,$$

since $\sum_{a\in\mathbb{A}}\big(N'(a)-N(a)\big) = n\sum_{a\in\mathbb{A}}\big(\mu'(a)-\mu(a)\big) = n-n = 0$. ☐

PROPOSITION 3.3. *Let $\nu\in\mathcal{M}_1$. For any $\mu\in\mathcal{L}_n$,*

$$(n+1)^{-|\mathbb{A}|}e^{-nD(\mu\|\nu)} \leq P_\nu(\mathbb{A}^n(\mu)) \leq e^{-nD(\mu\|\nu)}\,. \tag{35}$$

PROOF. The proof follows by using Lemma 3.2,

$$P_\nu(\mathbb{A}^n(\mu)) = \sum_{x\in\mathbb{A}^n(\mu)} P_\nu(x) = |\mathbb{A}^n(\mu)|\exp\big[-n(H(\mu)+D(\mu\|\nu))\big]\,,$$

and by using Lemma 3.3 to estimate $|\mathbb{A}^n(\mu)|$. ☐

PROOF OF THEOREM 3.5: The upper bound (30) (from which (32) follows immediately) is obtained as follows:

$$\begin{aligned} P_\nu(L_n\in\mathcal{E}) &= \sum_{\mu\in\mathcal{E}\cap\mathcal{L}_n} P_\nu(\mathbb{A}^n(\mu)) \qquad (36)\\ \text{by Proposition 3.3} &\leq \sum_{\mu\in\mathcal{E}\cap\mathcal{L}_n} e^{-nD(\mu\|\nu)}\\ &\leq |\mathcal{L}_n|e^{-n\inf_{\mu\in\mathcal{E}} D(\mu\|\nu)}\,, \end{aligned}$$

and $|\mathcal{L}_n|$ is estimated using Lemma 3.1. For the lower bound (31),

$$P_\nu(L_n \in \mathcal{E}) = \sum_{\mu \in \mathcal{E}\cap\mathcal{L}_n} P_\nu(\mathbb{A}^n(\mu))$$
$$\geq (n+1)^{-|\mathbb{A}|} \sup_{\mu \in \mathcal{E}\cap\mathcal{L}_n} e^{-nD(\mu\|\nu)},$$

where we used Proposition 3.3. Therefore,

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\nu(L_n \in \mathcal{E}) \geq -\limsup_{n\to\infty} \inf_{\mu \in \mathcal{E}\cap\mathcal{L}_n} D(\mu\|\nu).$$

Let $\mu_0 \in \text{int}\mathcal{E}$. Then clearly, there exists a sequence $\mu_n \in \mathcal{E} \cap \mathcal{L}_n$ such that $\|\mu_n - \mu_0\|_1 \to 0$. Therefore,

$$\limsup_{n\to\infty} \inf_{\mu \in \mathcal{E}\cap\mathcal{L}_n} D(\mu\|\nu) \leq \limsup_{n\to\infty} D(\mu_n\|\nu) = D(\mu_0\|\nu),$$

by the continuity of D (Proposition 3.1). This gives (31). □

As we have seen, when ν is the uniform distribution, $D(\mu\|\nu) = \log|\mathbb{A}| - H(\mu)$. Therefore, if $\mathcal{E} \subset \mathcal{M}_1$, we have

$$\inf_{\mu\in\mathcal{E}} D(\mu\|\nu) = \log|\mathbb{A}| - \sup_{\mu\in\mathcal{E}} H(\mu).$$

The variational problem that appears in the Sanov estimates is thus related, in this case, to a maximization of the Shannon Entropy. We will come back to this later.

1.1. The Theorem of Sanov, General Case. The Theorem of Sanov is general and holds when the random variables X_j take values in a complete separable metric space S [1]. Without proof, we will state this theorem in the simple case where $S = \mathbb{R}$. The first step towards this generalization is to redefine the relative entropy.

DEFINITION 3.2. *Let $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ be two probability measures on $\mathbb{R}$. The **relative entropy of μ with respect to ν** is defined by*

$$D(\mu\|\nu) := \begin{cases} \int f \log f \, d\nu & \text{if } \mu \ll \nu, \text{ and } f = \frac{d\mu}{d\nu}, \\ +\infty & \text{otherwise.} \end{cases} \tag{37}$$

As before, it can be shown that $\mu \to D(\mu\|\nu)$ is convex. We equip $\mathcal{M}_1(\mathbb{R})$ with the topology of weak convergence. That is, say that a sequence μ_n converges to μ if

$$\int f \, d\mu_n \to \int f \, d\mu$$

for all continuous bounded function $f : \mathbb{R} \to \mathbb{R}$. It can be shown that with respect to the weak topology, $\mu \to D(\mu\|\nu)$ is lower semicontinuous and has compact level sets.

[1]This happens to be a corollary of an even more general result, the Theorem of Cramér in infinite dimensional spaces. See [?].

Theorem 3.2 (Theorem of Sanov). *Let $X_1, X_2, \dots$ be i.i.d. with common distribution $\nu \in \mathcal{M}_1(\mathbb{R})$.*

(1) *For all closed set $\mathcal{C} \subset \mathcal{M}_1(\mathbb{R})$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_\nu(L_n \in \mathcal{C}) \leq - \inf_{\mu\in\mathcal{C}} D(\mu\|\nu)\,. \tag{38}$$

(2) *For all open set $\mathcal{G} \subset \mathcal{M}_1(\mathbb{R})$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\nu(L_n \in \mathcal{G}) \geq - \inf_{\mu\in\mathcal{G}} D(\mu\|\nu) \tag{39}$$

Proof. See [**?**], p.58. □

2. Applications

We go back to the case of a finite alphabet. Since the Theorem of Sanov 3.5 gives precise information on the convergence of the empirical measure L_n, it seems natural to ask whether it allows to derive similar informations about the empirical mean $\frac{S_n}{n}$. Let us be a little more general, and consider partial sums of the form $S_n^f := f(X_1) + \dots + f(X_n)$, where $f : \mathbb{A} \to \mathbb{R}$, and where as before the X_is are i.i.d., $\mathbb{A}$-valued with distribution ν. We introduce the notation

$$\langle f, \mu\rangle := \int_{\mathbb{A}} f\, d\mu = \sum_{a\in\mathbb{A}} \mu(a) f(a)\,,$$

with which $\frac{S_n^f}{n} = \langle f, L_n\rangle$. Therefore, for $A \subset \mathbb{R}$

$$\frac{S_n^f}{n} \in A \Leftrightarrow L_n \in \mathcal{E}_A^f \tag{40}$$

where $\mathcal{E}_A^f := \{\mu \in \mathcal{M}_1 : \langle f, \mu\rangle \in A\}$. Using this observation, we give two applications of the Theorem of Sanov.

2.1. The Law of Large Numbers.

Here we assume that $\mathbb{A} \subset \mathbb{R}$. Let $m := E[X_1] = \sum_a a\nu(a)$. Let $\epsilon > 0$, and $\mathcal{E}_\epsilon := \{\mu : D(\mu\|\nu) \geq \frac{1}{2}\epsilon^2\}$. By the Pinsker Inequality (28) and the upper bound (30),

$$P_\nu(\|L_n - \nu\|_1 \geq \epsilon) \leq P_\nu(L_n \in \mathcal{E}_\epsilon) \leq (n+1)^{|\mathbb{A}|} e^{-\frac{1}{2}\epsilon^2 n}\,. \tag{41}$$

To study the empirical mean $\frac{S_n}{n}$, we use (40) with $f = \mathrm{id}$,

$$\begin{aligned}
P_\nu\Big(\Big|\frac{S_n}{n} - m\Big| \geq \delta\Big) &= P_\nu(|\langle \mathrm{id}, L_n\rangle - m| \geq \delta) \\
&= P_\nu\Big(\Big|\sum_{a\in\mathbb{A}} a(L_n(a) - \nu(a))\Big| \geq \delta\Big) \\
&\leq P_\nu\Big(\sum_{a\in\mathbb{A}} |L_n(a) - \nu(a)| \geq \delta/a_{\max}\Big) \\
&= P_\nu\big(\|L_n - \nu\|_1 \geq \delta/2a_{\max}\big)\,,
\end{aligned}$$

where $a_{\max} := \max_{a\in\mathbb{A}} |a|$. Using (41), we conclude that there exists some constant $c_\delta > 0$ such that for large enough n,

$$P_\nu\Big(\Big|\frac{S_n}{n} - m\Big| \geq \delta\Big) \leq e^{-c_\delta n}\,.$$

Since this holds for all $\delta > 0$, the Borel-Cantelli Lemma implies that $\frac{S_n}{n} \to m$ almost surely.

2.2. The Theorem of Cramér. In the previous section we derived a Strong Law of Large Numbers for $\frac{S_n^f}{n}$ from the Theorem of Sanov. We now show that a Large Deviation Principle also holds for $\frac{S_n^f}{n}$, called the Theorem of Cramér. Later (see Chapter 7) this theorem will be generalized to arbitrary real-valued random variables.

Let $X_1, X_2, \dots$ be i.i.d., $\mathbb{A}$-valued, with $X_1 \sim \nu$. Let Λ denote the logarithmic moment generating function of X_1, i.e. for all $t \in \mathbb{R}$,

$$\Lambda(t) := \log \sum_{a\in\mathbb{A}} \nu(a) e^{tf(a)}\,. \tag{42}$$

The rate function for the LDP of the sequence $\frac{S_n^f}{n}$ will be given by the Legendre transform of Λ, i.e.

$$\Lambda^*(x) := \sup_{t\in\mathbb{R}}\{tx - \Lambda(t)\}\,. \tag{43}$$

Theorem 3.3 (Theorem of Cramér for finite alphabets). *Let $X_1, X_2, \dots$ be an i.i.d. sequence taking values in a finite alphabet $\mathbb{A}$ with common distribution ν. Let $f : \mathbb{A} \to \mathbb{R}$, $S_n^f := f(X_1) + \cdots + f(X_n)$. Then for all $A \subset \mathbb{R}$,*

$$-\inf_{x\in\mathring{A}} \Lambda^*(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log P_\nu\Big(\frac{S_n^f}{n} \in A\Big) \tag{44}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log P_\nu\Big(\frac{S_n^f}{n} \in A\Big) \leq -\inf_{x\in A} \Lambda^*(x)\,. \tag{45}$$

Remark 3.2. Remark that when I is continuous and when $A \subset \overline{\mathrm{int}A}$, then the upper and lower bounds coincide and we have

$$\lim_{n\to\infty} \frac{1}{n} \log P_\nu\Big(\frac{S_n^f}{n} \in A\Big) = -\inf_{x\in A} I(x)\,.$$

Before proving the theorem, we show how the Legendre transform Λ^* is related to the relative entropy D. Let $f_{\min} := \inf f$, $f_{\max} := \sup f$, $K := [f_{\min}, f_{\max}]$.

Lemma 3.4. *Let $\nu \in \mathcal{M}_1$, $f : \mathbb{A} \to \mathbb{R}$, and for all $x \in \mathbb{R}$, define*

$$I(x) := \inf_{\mu:\langle f,\mu\rangle = x} D(\mu\|\nu)\,. \tag{46}$$

Then $I = +\infty$ outside K, and for all $x \in K$, $I(x) = \Lambda^*(x)$. In particular, I is continuous on $\mathrm{int}K$. Moreover, I is strictly convex, and $I(x) \geq 0$ with equality if and only if $x = \langle f, \nu\rangle = E_\nu[f(X_1)]$.

Proof. If $x \notin K$, then there are no μ such that $\langle f, \mu \rangle = x$, and so $I(x) = +\infty$. For all $\mu \in \mathcal{M}_1$ (with $\mu(a) > 0$ for all $a \in \mathbb{A}$), Jensen Inequality gives

$$\begin{aligned}\Lambda(t) &= \log \sum_{a \in \mathbb{A}} \mu(a) \Big(e^{tf(a)} \frac{\nu(a)}{\mu(a)} \Big) \\ &\geq t\langle f, \mu \rangle - D(\mu \| \nu) \,, \end{aligned} \tag{47}$$

with equality if and only if $\mu = \mu^t$, where μ^t is the Gibbs distribution

$$\mu^t(a) = \frac{e^{tf(a)}}{Z(t)} \nu(a) \quad \forall a \in \mathbb{A} \,, \tag{48}$$

with $Z(t) = e^{\Lambda(t)}$. Therefore, $D(\mu \| \nu) \geq t\langle f, \mu \rangle - \Lambda(t)$, and by infimizing this lower bound over those μ for which $\langle f, \mu \rangle = x$, this shows that

$$I(x) \geq \Lambda^*(x) \,, \tag{49}$$

We then shown that equality holds when $x \in \text{int}K$. Clearly, Λ is C^∞ on $\mathbb{R}$. Moreover, a simple calculation shows that

$$\Lambda''(t) = \text{Var}_{\mu^t}[f] > 0 \,,$$

so that Λ is strictly convex, and $t \mapsto tx - \Lambda(t)$ is strictly concave. We can thus try to compute the supremum in Λ^* by simple derivation:

$$\Lambda^*(x) = t_* x - \Lambda(t_*) \,,$$

where $t_* = t_*(x)$ is the solution of $\Lambda'(t) = x$. The existence of t_* is guaranteed when $f_{\min} < x < f_{\max}$ (i.e. $x \in \text{int}K$), since

$$\lim_{t \to -\infty} \Lambda'(t) = f_{\min} \qquad \lim_{t \to +\infty} \Lambda'(t) = f_{\max} \,.$$

By the Implicit Function Theorem, $x \mapsto t_*(x)$ is C^∞. Now observe that $\langle f, \mu^{t_*} \rangle = \Lambda'(t_*) = x$, and so

$$\begin{aligned} I(x) \leq D(\mu^{t_*} \| \nu) &= \sum_{a \in \mathbb{A}} \mu^{t_*}(a) \log \frac{e^{t_* f(a)}}{Z(t_*)} \\ &= t_* \langle f, \mu^{t_*} \rangle - \log Z(t_*) \\ &= t_* x - \Lambda(t_*) \\ &= \Lambda^*(x) \,. \end{aligned}$$

We show that $I(x) \leq \Lambda^*(x)$ also on the boundary of K. Let for instance $x = f_{\max}$. Let a be the point of $\mathbb{A}$ at which f takes the value $f_{\max}$. We can assume that this point is unique (the general case follows by continuity). Let $\mu_{\max} := \delta_a$. Of course, $\langle f, \mu_{\max} \rangle = x$. It easy to verify that

$$\Lambda^*(x) \geq \limsup_{t \to \infty} \{tx - \Lambda(t)\} \geq -\log \nu(a) \equiv D(\mu_{\max} \| \nu) \geq I(x) \,.$$

The continuity of I on $\text{int}K$ follows from the fact that there $I(x) = \Lambda^*(x) = t_*(x)x - \Lambda(t_*(x))$, which is made of continuous functions. The strict convexity of I follows from the fact that on $\text{int}K$, $I''(x) = 1/\text{Var}_{\mu^{t_*(x)}}[f] > 0$. Observe that by Jensen, $\Lambda(t) \geq \langle f, \nu \rangle$, and so $I(\langle t, \nu \rangle) \leq 0$. The last claim follows from the fact that $I(x) \geq 0 \cdot x - \Lambda(0) = 0$, which implies that $I(\langle f, \nu \rangle) = 0$, and from the fact that I is strictly convex. □

More properties of $I(x) = \sup_t \{tx - \Lambda(t)\}$ will be described in Chapter 7.

The proof above shows that for reasonable values of the parameter x,

$$\inf_{\mu : \langle f, \mu \rangle = x} D(\mu \| \nu) = D(\mu^{t_*} \| \nu)\,, \tag{50}$$

where μ^{t_*} is the Gibbs distribution (48), with t_* chosen such that $\langle f, \mu^{t_*} \rangle = x$. In the following section, we will study more general variational problems related to the microcanonical distribution.

Proof of Theorem 3.3: The bounds (44)-(45) follow essentially from the correspondence (40) and the Theorem of Sanov. By Lemma 3.4,

$$\inf_{\mu \in \mathcal{E}^f_A} D(\mu \| \nu) = \inf_{x \in A} \inf_{\mu : \langle f, \mu \rangle = x} D(\mu \| \nu) \equiv \inf_{x \in A} \Lambda^*(x)\,,$$

from which the upper bound follows immediately. For the lower bound, observe that

$$\begin{aligned} \liminf_{n \to \infty} \frac{1}{n} \log P_\nu \Big(\frac{S^f_n}{n} \in A \Big) &\geq \liminf_{n \to \infty} \frac{1}{n} \log P_\nu \Big(\frac{S^f_n}{n} \in \mathrm{int} A \Big) \\ &= \liminf_{n \to \infty} \frac{1}{n} \log P_\nu \big(L_n \in \mathcal{E}^f_{\mathrm{int} A} \big)\,. \end{aligned}$$

Since $\mu \to \Phi(\mu) := \langle f, \mu \rangle$ is continuous, $\mathcal{E}^f_{\mathrm{int} A} = \Phi^{-1}(\mathrm{int} A)$ is open. Therefore, by the lower bound in the Theorem of Sanov and again Lemma 3.4,

$$\liminf_{n \to \infty} \frac{1}{n} \log P_\nu \big(L_n \in \mathcal{E}^f_{\mathrm{int} A} \big) \geq - \inf_{\mu \in \mathcal{E}^f_{\mathrm{int} A}} D(\mu \| \nu) = - \inf_{x \in \mathrm{int} A} \Lambda^*(x)\,.$$

This proves the theorem. □

Remark 3.3. What was done here is called a *contraction*: a large deviation result for L_n was used to derive a large deviation result for the empirical mean $\frac{S_n}{n} = \Phi(L_n)$, where $\Phi(\cdot) = \langle \mathrm{id}, \cdot \rangle$ is a continuous mapping from $\mathcal{M}_1$ to the real line. The new rate function I is obtained from the old D by the formula (46). This principle will be used at many places in these notes. It will be presented in its most general form in Chapter 5.

2.3. The Gibbs Conditioning Principle. Let X_i be i.i.d. with common distribution ν, and $\alpha > E_\nu[X_1]$. What is the distribution of X_1, conditionned on the event that $\frac{1}{n} \sum_{i=1}^n X_i \geq \alpha$? This type of problem is typical in statistical mechanics in the study of the microcanonical ensemble (see next chapter on the ideal gas), and can be solved using the Theorem of Sanov. We will see below that the limiting distribution of X_1 is a Gibbs distribution μ_β with a well chosen inverse temperature $\beta = \beta(\alpha)$. This is a weak form of the *Equivalence of Ensembles* for independent variables [2]. Moreover, if ν is uniform, μ_β also has the property of having maximal Shannon entropy among those distributions μ

[2]The general problem of the equivalence of ensembles will be treated in Chapter **??**.

that satisfy the constraint $\langle \mathrm{id}, \mu\rangle \geq \alpha$. This is essentially the Maximal Entropy Principle exposed in Chapter 2.

We thus consider an i.i.d. sequence $X_1, \dots, X_n$ with common distribution ν, and study the distribution of X_1, *conditioned* on an event of the form $\{L_n \in \mathcal{E}\}$:

$$\mu_n^{\mathcal{E}}(a) := P_\nu(X_1 = a | L_n \in \mathcal{E})\,. \tag{51}$$

FIGURE 1. The conditionning $L_n \in \mathcal{E}$, with two minimizers of $D(\cdot\|\nu)$.

Here, we assume that $\mathcal{E} \subset \mathcal{M}_1$ (non-empty) is atypical, in the sense that $P_\nu(L_n \in \mathcal{E}) \to 0$. For this we can for example assume that the closure of $\mathcal{E}$ does not contain ν. Namely, since D is continuous,

$$\inf_{\mu\in\mathcal{E}} D(\mu\|\nu) = \inf_{\mu\in\overline{\mathcal{E}}} D(\mu\|\nu) \equiv D_{\mathcal{E}} > 0\,,$$

and therefore by Sanov, $P_\nu(L_n \in \mathcal{E}) \to 0$ exponentially fast in n, with an exponent $D_{\mathcal{E}} > 0$.

The event $L_n \in \mathcal{E}$ is thus very rare, and the following question is: when a rare event happens, *how* does it happen? More generally, what can one say on the distribution of the variables $X_1, \dots, X_n$ when *conditioned* on a rare event to occur? In our framework, we will see that under a conditionning of the form $L_n \in \mathcal{E}$, the most likely empirical distribution is the one that *infimizes the relative entropy over* $\overline{\mathcal{E}}$. Let therefore

$$\mathcal{N} := \left\{\mu_* \in \overline{\mathcal{E}} : D(\mu_*\|\nu) = \inf_{\mu\in\mathcal{E}} D(\mu\|\nu) = D_{\mathcal{E}}\right\}.$$

It is then natural to ask if the probability distributions in $\mathcal{N}$ are candidates for being limit points of the sequence $\mu_n^{\mathcal{E}}$. We will show below that indeed the accumulation points of $\mu_n^{\mathcal{E}}$ are given by the closure of the convex hull of $\mathcal{N}$. In this sense, the probability distributions of $\mathrm{co}\mathcal{N}$ are candidates to describe the "most probable among the worst of all scenarios".

Throughout, we will assume that $\mathcal{E}$ has a dense relative interior, i.e. $\mathcal{E} \subset \overline{\mathrm{int}\mathcal{E}}$, so that the bounds in the Theorem of Sanov coincide (see Remark 3.1):

$$\inf_{\mu\in\mathrm{int}\mathcal{E}} D(\mu\|\nu) = \inf_{\mu\in\overline{\mathcal{E}}} D(\mu\|\nu) = D_{\mathcal{E}}\,. \tag{52}$$

THEOREM 3.4. *Assume $\mathcal{E} \subset \overline{\mathrm{int}\mathcal{E}}$. Then in the limit $n \to \infty$, the sequence $(\mu_n^{\mathcal{E}})_{n\geq 1}$ concentrates on the closure of the convex hull of $\mathcal{N}$, denoted $\overline{\mathrm{co}\mathcal{N}}$.*

PROOF. We first note that $\mu_n^{\mathcal{E}}$ can be written entirely in terms of L_n. Namely, let $f : \mathbb{A} \to \mathbb{R}$. Since the variables X_k are identically distributed,

$$E[f(X_1)|L_n \in \mathcal{E}] = E[f(X_2)|L_n \in \mathcal{E}] = \cdots = E[f(X_n)|L_n \in \mathcal{E}]\,.$$

Therefore

$$\begin{aligned}\langle f, \mu_n^{\mathcal{E}} \rangle &= E[f(X_1)|L_n \in \mathcal{E}] \\ &= E\Big[\frac{f(X_1)+\cdots+f(X_n)}{n}\Big|L_n \in \mathcal{E}\Big] \\ &= E[\langle f, L_n\rangle | L_n \in \mathcal{E}] \\ &= \langle f, E[L_n|L_n \in \mathcal{E}]\rangle\,,\end{aligned}$$

which gives $\mu_n^{\mathcal{E}} = E[L_n|L_n \in \mathcal{E}]$. Using this representation, we will show that for all $\delta > 0$,

$$d_1(\mu_n^{\mathcal{E}}, (\mathrm{co}\mathcal{N})_\delta) \to 0 \tag{53}$$

exponentially fast when $n \to \infty$, where we used the notation $\mathcal{U}_\delta := \{\mu : d_1(\mu, \mathcal{U}) < \delta\}$ for the open δ-thickening of the set $\mathcal{U}$, and $d_1(\mu, \mathcal{U}) := \inf_{\eta\in\mathcal{U}} \|\mu - \eta\|_1$. As can be seen, for any set $\mathcal{U} \subset \mathcal{M}_1$ and any $a \in \mathbb{A}$, the difference $\mu_n^{\mathcal{E}}(a) - \mu_n^{\mathcal{E}\cap\mathcal{U}}(a)$ can be expressed as

$$P(L_n \in \mathcal{U}^c|L_n \in \mathcal{E})\Big\{E[L_n(a)|L_n \in \mathcal{E}\cap\mathcal{U}^c] - E[L_n(a)|L_n \in \mathcal{E}\cap\mathcal{U}]\Big\}$$

Taking absolute values, summing over $a \in \mathbb{A}$, and using $\|\cdot\|_1 \le 2$,

$$\|\mu_n^{\mathcal{E}} - \mu_n^{\mathcal{E}\cap\mathcal{U}}\|_1 \le 2P(L_n \in \mathcal{U}^c|L_n \in \mathcal{E})\,.$$

Since $\mu_n^{\mathcal{E}\cap\mathcal{U}} \in \mathrm{co}\mathcal{U}$ [3],
this gives

$$d_1(\mu_n^{\mathcal{E}}, \mathrm{co}\mathcal{U}) \le 2P(L_n \in \mathcal{U}^c|L_n \in \mathcal{E})\,.$$

With $\mathcal{U} = \mathcal{N}_\delta$,

$$d_1(\mu_n^{\mathcal{E}}, \mathrm{co}(\mathcal{N}_\delta)) \le 2P(L_n \in (\mathcal{N}_\delta)^c|L_n \in \mathcal{E})\,.$$

By the Theorem of Sanov and (52),

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\log d_1(\mu_n^{\mathcal{E}}, \mathrm{co}(\mathcal{N}_\delta)) &\le -\Big\{\inf_{\mu\in(\mathcal{N}_\delta)^c\cap\mathcal{E}} D(\mu\|\nu) - \inf_{\mu\in\mathcal{E}} D(\mu\|\nu)\Big\} \\ &\le -\Big\{\inf_{\mu\in(\mathcal{N}_\delta)^c\cap\overline{\mathcal{E}}} D(\mu\|\nu) - D_{\mathcal{E}}\Big\} < 0\,.\end{aligned}$$

Namely, assume $\inf_{\mu\in(\mathcal{N}_\delta)^c\cap\overline{\mathcal{E}}} D(\mu\|\nu) = D_{\mathcal{E}}$. Since $(\mathcal{N}_\delta)^c \cap \overline{\mathcal{E}}$ is closed, it must contain some μ_0 for which $D(\mu_0\|\nu) = D_{\mathcal{E}}$. But this would mean $\mu_0 \in \mathcal{N}$, a contradiction. We have thus shown that $d_1(\mu_n^{\mathcal{E}}, \mathrm{co}(\mathcal{N}_\delta)) \to 0$ exponentially fast in n. To obtain (53), it suffices to observe that $(\mathrm{co}\mathcal{N})_\delta$ is convex and that it contains $\mathcal{N}_\delta$, so that $(\mathrm{co}\mathcal{N})_\delta \supset \mathrm{co}(\mathcal{N}_\delta)$. Therefore, $d_1(\mu_n^{\mathcal{E}}, (\mathrm{co}\mathcal{N})_\delta) \le d_1(\mu_n^{\mathcal{E}}, \mathrm{co}(\mathcal{N}_\delta))$. This finishes the proof. □

In most interesting situations, $\mathcal{E}$ will be convex. In this case, the strict convexity of the relative entropy implies that $\mathcal{N}$ is a singleton: $\mathcal{N} = \{\mu_*\}$. μ_* is then called the I-projection of ν on $\mathcal{E}$ (see Figure 2).

[3] Namely, $\mu_n^{\mathcal{E}\cap\mathcal{U}}$ is a convex combination of elements of $\mathcal{E}\cap\mathcal{U} \subset \mathcal{U}$:

$$\mu_n^{\mathcal{E}\cap\mathcal{U}} = E[L_n|L_n \in \mathcal{E}\cap\mathcal{U}] = \int_{L_n\in\mathcal{E}\cap\mathcal{U}} L_n dP_n^{\mathcal{E}\cap\mathcal{U}}\,,$$

where $P_n^{\mathcal{E}\cap\mathcal{U}}(\cdot) = P(\cdot|L_n \in \mathcal{E}\cap\mathcal{U})$.

FIGURE 2. The I-projection of ν on a convex set $\mathcal{E}$.

COROLLARY 3.1. *Assume that* $\mathcal{E} \subset \overline{\text{int}\mathcal{E}}$ *and that* $\mathcal{E}$ *is convex. Then* $\mathcal{N} = \{\mu_*\}$, *and* $\|\mu_n^{\mathcal{E}} - \mu_*\|_1 \to 0$.

Let us then answer the question asked at the beginning of the section. Let X_i be i.i.d. with $X_i \sim \nu$. We consider the distribution of X_1 when conditioned on $(X_1 + \cdots + X_n)/n \geq \alpha$. In terms of the empirical measure, this represents the event $\mathcal{E} = \{\mu : \langle \text{id}, \mu \rangle \geq \alpha\}$. Assume that $\langle \text{id}, \nu \rangle < \alpha$, so that $\mathcal{E}$ does not contain ν. Clearly, $\mathcal{E}$ is convex, and $\mathcal{E} = \overline{\text{int}\mathcal{E}}$. By Corollary 3.1

$$P_\nu\Big(X_1 = a \Big| \frac{X_1 + \cdots + X_n}{n} \geq \alpha\Big) \to \mu_*(a)\,,$$

where μ_* is the I-projection of ν on $\mathcal{E}$, i.e. the unique measure μ_* which satisfies

$$D(\mu_* \| \nu) = \inf_{\mu \in \overline{\mathcal{E}}} D(\mu \| \nu)\,. \tag{54}$$

As we have already seen, μ_* can be found by writing

$$\inf_{\mu \in \overline{\mathcal{E}}} D(\mu\|\nu) = \inf_{x \geq \alpha} \inf_{\mu : \langle \text{id}, \mu \rangle = x} D(\mu\|\nu) \equiv \inf_{x \geq \alpha} \Lambda^*(x)\,,$$

where Λ^* is the Legendre transform of $\Lambda(t) = \log \sum_a e^{ta}\nu(a)$. Since Λ^* is strictly convex and has a unique minimum at $m = E_\nu[X_1]$, and since we are assuming $\alpha > m$, we have

$$\inf_{x \geq \alpha} \Lambda^*(x) = \Lambda^*(\alpha) = \inf_{\mu : \langle \text{id}, \mu \rangle = \alpha} D(\mu\|\nu) = D(\mu^{t_*}\|\nu)\,,$$

where $t_* = t_*(\alpha)$ is the unique value of t for which $\Lambda'(t) = \alpha$.

REMARK 3.4. Assume for a while that ν is the uniform measure. Then $D(\cdot\|\nu) = \log|\mathbb{A}| - H(\cdot)$, and therefore μ_* satisfies

$$H(\mu_*) = \sup_{\mu \in \mathcal{E}} H(\mu)\,. \tag{55}$$

In other words, μ_* is a measure with maximal entropy, under the constraint $\mu \in \mathcal{E}$. We thus see that the limiting distribution of an independent sequence under conditionning is of maximal entropy, in the sense of Chapter 2. Typically, if $\mathcal{E}$ is a convex set of the form $\{\mu \in \mathcal{M}_1 : E_\mu[f] \in [a, b]\}$, then the method of Lagrange multipliers applies, and μ_* can be shown to be a Gibbs distribution with potential f and a well defined inverse temperature β (depending on $[a, b]$). We will come back to this application in Chapter 4.

3. The Theorem of Sanov for Pairs

The empirical measure L_n studies the frequency with which the individual letters appear in a sample of size n. More detailed information can be extracted from the sample by studying the occurence of *words*. The case of words of size two already contains the main difficulty, and we shall stick to this case.

The **empirical pair measure** is the probability distribution on words of size two, defined by

$$L_n^2 := \frac{1}{n}\sum_{j=1}^{n} \delta_{(X_j,X_{j+1})} . \tag{56}$$

As will be seen, it is convenient to define L_n^2 by identifying that $X_{n+1} \equiv X_1$ (which, when n is large, has no influence on the statistics of pairs). $L_n^2 \in \mathcal{M}_1(\mathbb{A}^2)$, the set of probability distributions on pairs $(a,b) \in \mathbb{A}^2$. A generic element of $\mathcal{M}_1(\mathbb{A}^2)$ will usually be denoted by π. The topology on $\mathcal{M}_1(\mathbb{A}^2)$ is that of the L^1-norm, defined as before:

$$\|\pi - \pi'\|_1 := \sum_{a,b} |\pi(a,b) - \pi'(a,b)| .$$

The marginal of the first (resp. second) coordinate of π, denoted $\overline{\pi} \in \mathcal{M}_1(\mathbb{A})$ (resp. $\underline{\pi} \in \mathcal{M}_1(\mathbb{A})$), is given by $\overline{\pi}(a) := \sum_b \pi(a,b)$ (resp. $\overline{\pi}(b) := \sum_a \pi(a,b)$). By the cyclicity condition $X_{n+1} \equiv X_1$, we actually have $L_n^2 \in \mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2)$, where

$$\mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2) := \{\pi \in \mathcal{M}_1(\mathbb{A}^2) : \overline{\pi} = \underline{\pi}\} .$$

We of course expect L_n^2 to be close to $\nu \otimes \nu$ when n is large. Actually, as can be shown using the Ergodic Theorem of Birkhoff (or simply the Strong Law of Large Numbers, if suitably applied), when $n \to \infty$,

$$L_n^2 \Rightarrow \nu \otimes \nu \quad P_\nu - a.s. \tag{57}$$

We will describe the concentration of L_n^2 in the vicinity of $\nu \otimes \nu$ in terms of the following rate function: if $\pi \in \mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2)$,

$$I_\nu^2(\pi) := \sum_{a,b} \pi(a,b) \log \frac{\pi(a,b)}{\overline{\pi}(a)\nu(b)} . \tag{58}$$

Observe that if D denotes the relative entropy between distributions on $\mathcal{M}_1(\mathbb{A}^2)$ (see Definition 3.1), then

$$I_\nu^2(\pi) = D(\pi \| \overline{\pi} \otimes \nu) , \tag{59}$$

and not $D(\pi\|\nu \otimes \nu)$ as one could have naively expected. The presence of $\overline{\pi} \otimes \nu$ comes from the fact that unlike single letters, words can overlap. Nevertheless, I_ν^2 has the expected properties (compare with Proposition 3.1):

Proposition 3.4. *$I_\nu^2 : \mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2) \to \mathbb{R}$ has the following properties:*

1. *$I_\nu^2(\pi) \geq 0$, with equality if and only if $\pi = \nu \otimes \nu$.*
2. *$\pi \to I_\nu^2(\pi)$ is strictly convex, except on segments $[\pi, \pi']$ which satisfy $\pi(a,b)/\overline{\pi}(a) = \pi'(a,b)/\overline{\pi}'(a)$, on which it is affine.*

Proof. (Similar to proof of Proposition 3.1.) Write

$$I_\nu^2(\pi) = \sum_{a,b} \overline{\pi}(a)\nu(b)\psi(X_{a,b}) ,$$

where $\psi(x) := x \log x$ and $X_{a,b} := \frac{\pi(a,b)}{\overline{\pi}(a)\nu(b)}$. By Jensen's Inequality, $I_\nu^2(\pi) \geq \psi(1) = 0$, with equality if and only if $\pi(a,b) = \overline{\pi}(a)\nu(b)$. But since $\pi \in \mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2)$, this implies $\overline{\pi} = \underline{\pi} = \nu$, and so $\pi = \nu \otimes \nu$. Convexity follows from the representation

(59) and from (3) in Proposition 3.1, which implies that $\pi \mapsto D(\pi \| \overline{\pi} \otimes \nu)$ is convex. □

THEOREM 3.5 (Theorem of Sanov for pairs). *Let $X_1, X_2, \dots$ be i.i.d. with common distribution ν. For all $\mathcal{E} \subset \mathcal{M}_1(\mathbb{A}^2)$,*

$$-\inf_{\pi \in \mathrm{int}\mathcal{E}} I_\nu^2(\pi) \le \liminf_{n\to\infty} \frac{1}{n} \log P_\nu(L_n^2 \in \mathcal{E}) \tag{60}$$

$$\le \limsup_{n\to\infty} \frac{1}{n} \log P_\nu(L_n^2 \in \mathcal{E}) \le -\inf_{\pi \in \mathcal{E}} I_\nu^2(\pi)\,, \tag{61}$$

PROOF. Let as before $\mathcal{L}_n^2 \subset \mathcal{M}_1^{\mathrm{cycl}}(\mathbb{A}^2)$ denote the set of (cyclic) types of size n, i.e. the set of all empirical measures L_n^2 associated to the sequences $(X_1, \dots, X_n) \in \mathbb{A}^n$. We start as in the proof of Theorem 3.5, and decompose as we did in (36):

$$P_\nu(L_n^2 \in \mathcal{E}) = \sum_{\pi \in \mathcal{E} \cap \mathcal{L}_n^2} P_\nu(L_n^2 = \pi)\,. \tag{62}$$

The key of the proof is to observe that a sequence $X = (X_1, \dots, X_n)$ can be encoded into a path in a graph with vertices $\mathbb{A}$, and that L_n^2 can be read off the path by counting the number of times each edge is visited.

A multigraph (with vertex set $\mathbb{A}$), $G = (\mathbb{A}, E)$, is specified by a multiset E of oriented edges, that is, a set of pairs (a, b) with $a, b \in \mathbb{A}$, in which the same pair can appear more than once. The multiset E is therefore completely characterized by a family of numbers $k_G(a, b) \in \{0, 1, \dots, n\}$, where $k_G(a, b)$ gives the number of times the edge (a, b) appears in E.

We associate to each sample $X = (X_1, \dots, X_n)$ a multigraph $G(X) = (\mathbb{A}, E(X))$ by walking along the vertices $\mathbb{A}$ and adding a new edge (a, b) to the edge-set each time it is traversed. In the end, the edge (a, b) appears the same number of times in $E(X)$ as the number of is for which $(X_i, X_{i+1}) = (a, b)$. By the identification $X_{n+1} \equiv X_1$, the path is back at its starting point X_1 at time $n+1$. The multigraph $G(X)$ contains all the information about L_n^2: for all pair (a, b),

$$L_n^2(a, b) = \frac{k_{G(X)}(a, b)}{n}\,.$$

For a given $\pi \in \mathcal{L}_n^2$, of the form $\pi = \frac{k}{n}$, let G_π denote the multigraph associated to π, in which the edge (a, b) appears exactly $k(a, b) \equiv n\pi(a, b)$ times. Then

$$P_\nu(L_n^2 = \pi) = \sum_{X : G(X) = G_\pi} P_\nu(X)\,.$$

Clearly, all the sequences X appearing in the sum have same probability. Namely, in each of them, the symbol $a \in \mathbb{A}$ appears exactly $\sum_b k_{G(X)}(a, b) = \sum_b k(a, b) \equiv \overline{k}(a)$ times, and so

$$P_\nu(L_n^2 = \pi) = \Big(\prod_a \nu(a)^{\overline{k}(a)} \Big) \sharp\{X : G(X) = G_\pi\}\,. \tag{63}$$

Let G'_π denote the graph with vertex set $\mathbb{A}$ such that each pair of points (a,b) is joined by exactly $k(a,b)$ distinct oriented edges. Now, the number of elements in $\{X : G(X) = G_\pi\}$ can be obtained by first counting the number of distinct Eulerian paths on G'_π (that is paths that visit each edge of G'_π exactly once), and then take care of the number of overcountings that have been made. Since exactly $\prod_{a,b} k(a,b)!$ distinct Eulerian paths yield the same sequence X we get

$$\sharp\{X : G(X) = G_\pi\} = \mathrm{cycl}(\pi) \cdot \frac{\sharp\{\text{Eulerian paths on } G'_\pi\}}{\prod_{a,b} k(a,b)!} ,$$

where $1 \leq \mathrm{cycl}(\pi) \leq n$ is a number that accounts for the fact that each Eulerian path can yields to at most n distinct sequences X, obtained by cyclic permutation.

LEMMA 3.5. *For all* $\pi = \frac{k}{n} \in \mathcal{L}_n^2$,

$$\prod_a (\overline{k}(a) - 1)! \leq \sharp\{\text{Eulerian paths on } G'_\pi\} \leq |\mathbb{A}| \prod_a \overline{k}(a)! . \tag{64}$$

Since $1 \leq \prod_a \overline{k}(a) \leq n^{|\mathbb{A}|}$, we can thus write (63) as:

$$P_\nu(L_n^2 = \pi) = \Big(\prod_a \nu(a)^{\overline{k}(a)}\Big) e^{O(\log n)} \frac{\prod_a \overline{k}(a)!}{\prod_{a,b} k(a,b)!} , \tag{65}$$

uniformly in $\pi \in \mathcal{L}_n^2$. Using the Stirling formula for each of the factorials $k(a,b)!$, $\overline{k}(a)!$, gathering carefully the terms and remembering (58),

$$P_\nu(L_n^2 = \pi) = e^{-nI_\nu^2(\pi) + O(\log n)} . \tag{66}$$

The rest of the proof proceeds as in Theorem 3.5. For example, for the upper bound,

$$\begin{aligned} P_\nu(L_n^2 \in \mathcal{E}) &\leq |\mathcal{E} \cap \mathcal{L}_n^2| \sup_{\pi \in \mathcal{E} \cap \mathcal{L}_n^2} P_\nu(L_n^2 = \pi) \\ &\leq |\mathcal{E} \cap \mathcal{L}_n^2| e^{O(\log n)} \exp\big(-n \inf_{\pi \in \mathcal{E} \cap \mathcal{L}_n^2} I_\nu^2(\pi)\big) \\ &\leq |\mathcal{E} \cap \mathcal{L}_n^2| e^{O(\log n)} \exp\big(-n \inf_{\pi \in \mathcal{E}} I_\nu^2(\pi)\big) , \end{aligned}$$

from which (61) follows, since $|\mathcal{E} \cap \mathcal{L}_n^2| \leq |\mathcal{L}_n^2| \leq (n+1)^{|\mathbb{A}^2|}$. □

The Theorem of Sanov for pairs will be used later when studying large deviations for Markov chains. COMPLETER

CHAPTER 4

The Ideal Gas

A system is called *ideal* when it is made of particles that do not interact. Mathematically, their study is greatly simplified by the fact that they can be reduced to sequences of independent random variables. In the case of the ideal gas, this will allow us to apply the large deviations results of the previous chapter. Although the model presented hereafter is oversimplified, the results obtained are far-reaching, and provide a rigorous answer to some of the questions raised in the Introduction about large systems of particles. We follow the treatment of Ellis in [?].

Consider a system of numbered n particles in a box, all with equal mass $m \equiv 1$. For simplicity, let this box be a one-dimensional bounded interval $\Delta \subset \mathbb{R}$. The state of each particle is specified by a pair (q_j, p_j), where $q_i \in \Delta$ is the position, and $p_j \in \mathbb{R}$ is the momentum. To simplify, we discretize the problem and assume that the momenta take on only a finite number of possible values, say in a symmetric set $\Gamma := \{v_i, i = -r, -r+1, \dots, r-1, r\}$, where $v_{-i} = -v_i$, and where we assume that $v_{-r} < v_{-r+1} < \cdots < v_{-1} < v_1 < \cdots < v_{r-1} < v_r$. The microscopic states of the system are therefore the elements $\omega \in \Omega_n$, where

$$\Omega_n := \{\omega = (q_1, p_1, \dots, q_n, p_n) : q_j \in \Delta, \, p_j \in \Gamma\} \equiv (\Delta \times \Gamma)^n \, .$$

Since we are assuming that the gas is ideal, i.e. that the particles don't interact, neither on short or long range. The dynamics of this model is thus quite simple: since it doesn't feel the others, the particle j with momentum $p_j \in \Gamma$ travels at constant speed until it hits the boundary of Δ. It then reverses its momentum, $p_j \to -p_j$ and continues at constant speed until hitting again the boundary of Δ.

Although it is possible to write down the detailed trajectory of each particle, the complete evolution of the system is complicated to described. Rather, as we said in the introduction, we are more interested in finding a probability measure on Ω_n, describing the equilibrium properties of the system. The main measure will be the *microcanonical measure*, which is simply obtained by assuming that the total energy of the system if fixed. We know from thermodynamics that this is essentially equivalent to fixing the temperature of the system (although this will be proved in the present section). The main result (Theorems 4.1 and 4.3) will be that under the microcanonical measure, i.e. when the total energy of the system is fixed in a small interval C, then all other macroscopic observables are essentially deterministic, in the sense that they have exponentially small fluctuations around their mean value. We will also see that these mean values can be computed using a Gibbs distribution whose inverse temperature is chosen as a function of the interval C. This should be considered as a satisfactory description of equilibrium.

1. The Microcanonical Distribution

Let $\mathcal{B}(\Delta)$ denote the Borel σ-field on Δ, and $\mathcal{B}(\Gamma)$ the discrete σ-field on Γ. We will consider Ω_n endowed with the product σ-field $\mathcal{F}_n := (\mathcal{B}(\Delta) \otimes \mathcal{B}(\Gamma))^{\otimes n}$. If no particular information is provided about the state of the gas, the Maximum Entropy Principle leads us to choose an a priori probability measure which does not favor any special microscopic state ω: the uniform measure on $(\Omega_n, \mathcal{F}_n)$, which will be denoted ν_n. Let λ denote the normalized [1] Lebesgue measure on $(\Delta, \mathcal{B}(\Delta))$, i.e. such that $\lambda(\Delta) = 1$, and ρ the uniform measure on $(\Gamma, \mathcal{B}(\Gamma))$, i.e.

$$\rho := \frac{1}{|\Gamma|} \sum_{i=-r}^{r} \delta_{v_i} \, .$$

Our reference measure on $(\Omega_n, \mathcal{F}_n)$ is thus the product measure

$$\nu_n := (\lambda \otimes \rho)^{\otimes n} \, .$$

Expectation under ν_n is denoted E_n. The probability space $(\Omega_n, \mathcal{F}_n, \nu_n)$ modelizes a system of particles with no interactions.

We shall be interested in the behaviour of certain observables associated to this system, namely random variables $F : \Omega_n \to \mathbb{R}$ (or $\mathbb{R}^2$). The observables we will be interested in will typically be macroscopic, i.e. involving all the particles of the system. These are usually of the form

$$F(\omega) = \sum_{j=1}^{n} f(q_j, p_j) \, .$$

For example, with $f(q,p) = 1_A(q)$, the observable

$$I_A(\omega) := \sum_{j=1}^{n} 1_A(q_j)$$

counts the number of particles contained in the Borel set $A \subset \Delta$. Our main example will be that of the total kinetic energy,

$$U_n(\omega) := \sum_{j=1}^{n} \frac{p_j^2}{2} \, .$$

which corresponds to the choice $f(q,p) = \frac{1}{2}p^2 \equiv g(p)$. By introducing $Y_j(\omega) := p_j$, $j = 1, \dots, n$, U_n writes

$$U_n = \sum_{j=1}^{n} g(Y_j) \, .$$

[1] This normalization condition on λ is not necessary.

Since $\frac{1}{2}v_1^2 \leq g(p) \leq \frac{1}{2}v_r^2$, U_n takes values in $[\frac{1}{2}nv_1^2, \frac{1}{2}nv_r^2] = n[g(v_1), g(v_r)]$. Under the reference measure ν_n, the variables Y_j are i.i.d. This implies that

$$E_n\Big[\frac{U_n}{n}\Big] = E_n[g(Y_1)] = \int_\Gamma g(p)\rho(dp) = \frac{1}{2r}\sum_{i=1}^r v_i^2 \equiv \alpha\,. \tag{67}$$

In the second equality, we have integrated out all variables p_j, $j = 1, \dots, n$, and all the q_j, $j = 2, \dots, n$. Under ν_n, the total kinetic energy satisfies a LLN: when $n \to \infty$,

$$\nu_n\Big(\Big|\frac{U_n}{n} - \alpha\Big| \geq \epsilon\Big) \to 0\,. \tag{68}$$

1.1. The microcanonical mesure. In general, physical systems are *not* described by product measures, since some external influence usually fixes the value of certain observables of the system. A typical situation is where the system is in contact with a heat reservoir that keeps it at constant temperature. It is well known from thermodynamics that the temperature is proportional to the average kinetic energy per particle:

$$\big\langle \tfrac{1}{2}mv^2 \big\rangle = \tfrac{3}{2}kT\,, \tag{69}$$

where k is the Boltzmann constant. We thus see that imposing a restriction on the temperature is equivalent to imposing a restriction on the kinetic energy. Although it will be one of our tasks in this chapter to settle this equivalence on rigorous bases, we will take as a starting point a pointwise restriction on the kinetic energy, by assuming that U_n is constrained to take values in a small interval. This leads to a first interesting probability measure (considered often as the starting point of statistical mechanics):

DEFINITION 4.1. *Let $C \subset [g(v_1), g(v_r)]$ be a closed interval. The microcanonical distribution is the probability measure ν_n^C on $(\Omega_n, \mathcal{F}_n)$ defined by*

$$\nu_n^C(\cdot) := \nu_n\Big(\cdot \,\Big|\, \frac{U_n}{n} \in C\Big)\,. \tag{70}$$

(We are assuming that n is large enough, so that $\{\frac{U_n}{n} \in C\} \neq \emptyset$.) Expectation under ν_n^C is denoted E_n^C.

The microcanonical distribution is thus the uniform measure over the energy shell $\{\omega \in \Omega_n : U_n(\omega)/n \in C\}$ [2]. Our aim is thus to study the properties of the system under distribution ν_n^C, for fixed C and large n. As we will see, all observables are exponentially concentrated around a deterministic mean value, which can be computed in function of C. This concentration will become sharper when n gets larger, and it will thus be natural to consider the limit in which the number of particles goes to infinity. The limit $n \to \infty$ is called the thermodynamic

[2] In Statistical Mechanics textbooks, the microcanonical measure is introduced by invoking the *ergodic hypothesis*: "over long periods of time, the time spent by a particle in some region of the phase space of microstates with the same energy is proportional to the volume of this region, i.e., that all accessible microstates are equally probable over a long period of time." (Wikipedia)

limit [3], in the sense that for large n the properties of the system become close to those of a large system (of thermodynamic size, i.e. for which typically $n \simeq 10^{25}$).

2. The equilibrium value of the kinetic energy

We saw in (68) that under ν_n, $U_n/n \to \alpha$. We will now see that under ν_n^C, $U_n/n \to u_*$, where u_* is the unique minimizer of Λ_1^* over C, where Λ_1^* is the Legendre Transform of the logarithmic moment generating function of the kinetic energy of the first particle:

$$\inf_{x \in C} \Lambda_1^*(x) = \Lambda_1^*(u_*)\,. \tag{71}$$

In other words, u_* equals to α if $\alpha \in C$, and to the point of C closest to C if $\alpha \notin C$.

THEOREM 4.1. *For all $\epsilon > 0$, there exists $\gamma_\epsilon = \gamma_\epsilon(C) > 0$ such that when $n \to \infty$,*

$$\nu_n^C\Big(\Big|\frac{U_n}{n} - u_*\Big| \geq \epsilon\Big) \leq e^{-\gamma n}\,. \tag{72}$$

As (72) shows, U_n/n has a deterministic limit under the microcanonical measure: we call u_* the macroscopic equilibrium kinetic energy (per particle). If one interprets $-\Lambda_1^*$ as the entropy of the system, which is concave, then (71) says that the equilibrium value observed by the system is the one that maximizes the entropy under the constraint C.

We will denote by $K_j := \frac{1}{2}Y_j^2 = g(Y_j)$ the kinetic energy of the particle j. K_j takes values in the finite set $\mathbb{K} := \{k_1, \dots, k_r\}$, where $k_j = g(v_j)$.

PROOF OF THEOREM 4.1: Let $A := \{z : |z - u_*| \geq \epsilon\}$, assuming that ϵ is small. By integrating out the variables q_j,

$$\nu_n^C\Big(\frac{U_n}{n} \in A\Big) = \frac{\nu_n\big(\frac{U_n}{n} \in A \cap C\big)}{\nu_n\big(\frac{U_n}{n} \in C\big)} = \frac{\rho^{\otimes n}\big(\frac{U_n}{n} \in A \cap C\big)}{\rho^{\otimes n}\big(\frac{U_n}{n} \in C\big)} \equiv \frac{P_\rho\big(\frac{U_n}{n} \in A \cap C\big)}{P_\rho\big(\frac{U_n}{n} \in C\big)}\,,$$

where $P_\rho = \rho^{\otimes \infty}$. Since the variables K_j are independent and take values in a finite alphabet, the Theorem of Cramér 3.3 applies: if Λ_1 denotes the logarithmic moment generating function of K_1,

$$\Lambda_1(t) = \log E_\rho[e^{tK_1}] = \log \int_\Gamma e^{tg(y)} \rho(dy)\,,$$

and Λ_1^* denotes its Legendre Transform, then $\frac{U_n}{n} = \frac{1}{n}\sum_{j=1}^n K_j$ satisfies a Large Deviation principle with rate function Λ_1^*. In particular (see Lemma 3.4), Λ_1^* is strictly convex and continuous on $[k_1, k_r]$, with a unique minimum at $E_\rho[K_1] = \alpha$.

[3] Actually, the true thermodynamic limit should involve also a limit in which the size of the system Δ goes to infinity with n:$\Delta_n \nearrow \infty$. Since we are here studying a system of particles that don't interact, this will not be necessary.

Therefore, by the Theorem of Sanov and since both $A \cap C$ and C have dense relative interiors (see Remark 3.2),

$$-\lim_{n\to\infty} \frac{1}{n} \log \nu_n^C\Big(\frac{U_n}{n} \in A\Big) = \inf_{x\in A\cap C} \Lambda_1^*(x) - \inf_{x\in C} \Lambda_1^*(x) \equiv 2\gamma_\epsilon\,.$$

We consider separately the cases $C \ni \alpha$ and $C \not\ni \alpha$. If $C \ni \alpha$, then $u_* = \alpha$, and so $\inf_{x\in C} \Lambda_1^*(x) = 0$. Moreover, since A is closed and does not contain α, we have $\inf_{x\in A\cap C} \Lambda_1^*(x) > 0$. Therefore, $\gamma_\epsilon > 0$. On the other hand, if $C \not\ni \alpha$, the convexity of Λ_1^* gives $\inf_{x\in C} \Lambda_1^*(x) = \Lambda_1^*(u_*)$, and again since A is closed and does not contain u_*, we get $\inf_{x\in A\cap C} \Lambda_1^*(x) > \Lambda_1^*(u_*)$. This shows that $\gamma_\epsilon > 0$. □

We then study the distribution of the kinetic energy of particle 1 under ν_n^C.

THEOREM 4.2. *Let $\nu_n^C \circ K_1^{-1} \in \mathcal{M}_1(\mathbb{K})$ denote the distribution of K_1. Let β_* denote the unique value of β for which $\Lambda_1'(-\beta) = u_*$. Then $\nu_n^C \circ K_1^{-1} \to \mu_{\beta_*}$, where $\mu_{\beta_*}(k) = e^{-\beta_* k}/Z(\beta_*)$ is the Gibbs distribution on $\mathbb{K}$. More precisely,*

$$d_1\big(\nu_n^C \circ K_1^{-1}, \mu_{\beta_*}\big) \to 0\,, \tag{73}$$

exponentially fast when $n \to \infty$.

PROOF. Again, since the microcanonical constraint $\{U_n/n \in C\}$ concerns only the variables K_j and since we are studying K_1, it is natural to formulate everything in terms of the K_js. Let $\hat\rho \in \mathcal{M}_1(\mathbb{K})$, $\hat\rho(k) := P_\rho(K_j = k)$, denote the distribution of K_j under P_ρ, and $P_{\hat\rho} := \hat\rho^{\otimes\infty}$. Then

$$\nu_n^C\big(K_1 = k\big) = P_{\hat\rho}\big(K_1 = k \big| L_n \in \mathcal{E}\big)\,,$$

where L_n is the empirical measure associated to the sequence K_j, and $\mathcal{E} = \{\mu \in \mathcal{M}_1(\mathbb{K}) : \langle \mathrm{id}, \mu\rangle \in C\}$. Then

$$\inf_{\mu\in\mathcal{E}} D(\mu\|\hat\rho) = \inf_{x\in C} \Lambda_1^*(x) = \Lambda_1^*(u_*) = D(\mu_{\beta_*}\|\hat\rho)\,.$$

Clearly, $\mathcal{E}$ is convex and therefore μ_{β_*} is the unique infimizer of $D(\cdot\|\hat\rho)$ over $\mathcal{E}$, and (73) is a consequence of Corollary 3.1. □

It can be shown (c.f. [?]) that in the limit $n \to \infty$, the variables Y_j become independent, with distribution $\mu_{\beta_*} \circ g^{-1}$.

The convergence $\nu_n^C(K_1 = k) \to \mu_{\beta_*}(k)$ of the previous theorem can be formulated in When expressed in terms of the moments, the convergence of Theorem 4.2 takes the form:

$$\nu_n^C(Y_1 = v) \to \frac{e^{-\frac{1}{2}\beta_* v^2}}{Z(\beta_*)}\,,$$

which is known as the Maxwell-Boltzmann distribution of velocities in a gas. In the continuous setting, it takes the form

$$\Big(\frac{m}{2\pi kT}\Big)^{\frac{3}{2}} e^{-\frac{1}{2}\frac{m}{kT}\vec{v}^2}$$

We have therefore started with identically, uniformly distributed velocities, and seen that under a global constraint on the kinetic energy, the distribution becomes

Gibbsian for a given temperature. This equivalence, which holds only after the thermodynamic limit, is a simplified form of the Equivalence of Ensembles.

3. The equilibrium value of other observables

Having showed that when the system is conditionned on $\{\frac{U_n}{n} \in C\}$, the average kinetic energy $\frac{U_n}{n}$ converges to u_*, we go on and study the behaviour of other observables, always of the form $F_n := \sum_{j=1}^n f(q_j, p_j)$.

THEOREM 4.3. *Let $f : \Delta \times \Gamma \to \mathbb{R}$ be bounded and continuous and let F_n be defined as above. Then for all $\epsilon > 0$ there exists $\gamma_\epsilon = \gamma_\epsilon(f, C) > 0$ such that when $n \to \infty$,*

$$\nu_n^C\Big(\Big|\frac{F_n}{n} - \langle f \rangle_*\Big| \geq \epsilon\Big) \leq e^{-\gamma_\epsilon n}\,, \tag{74}$$

where

$$\langle f \rangle_* := \int_{\Delta\times\Gamma} f(q,p)(\lambda \otimes \mu_{\beta_*})(dq, dp)\,. \tag{75}$$

PROOF. We will give a proof in the case where f depends on p only: $f(q,p) = f(p)$. Since $f(p)$ may not be necessarily related to the kinetic energy, we work on Γ (rather than on $\mathbb{K}$). Let $B = \{u : |u - \langle f \rangle_*| \geq \epsilon\}$. We use the correspondence (40) to write the probability in terms of the empirical measure $L_n = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$,

$$\nu_n^C\Big(\Big|\frac{F_n}{n} - \langle f \rangle_*\Big| \geq \epsilon\Big) = \nu_n^C\Big(\frac{1}{n}\sum_{j=1}^n g(Y_j) \in B\Big) \tag{76}$$

$$= \frac{\nu_n\big(\langle f, L_n\rangle \in B,\, \langle g, L_n\rangle \in C\big)}{\nu_n(\langle g, L_n\rangle \in C)} \tag{77}$$

$$= \frac{P_\rho(L_n \in \mathcal{E}_B^f \cap \mathcal{E}_C^g)}{P_\rho(L_n \in \mathcal{E}_C^g)}\,. \tag{78}$$

where $\mathcal{E}_B^f = \{\mu \in \mathcal{M}_1(\Gamma) : \langle f, \mu\rangle \in B\}$, $\mathcal{E}_C^g = \{\mu \in \mathcal{M}_1(\Gamma) : \langle g, \mu\rangle \in C\}$. These two sets have obviously dense relative interiors. We can thus apply the Theorem of Sanov on Γ:

$$-\lim_{n\to\infty}\frac{1}{n}\log \nu_n^C\Big(\Big|\frac{F_n}{n} - \langle f \rangle_*\Big| \geq \epsilon\Big) = \inf_{\mu\in\mathcal{E}_B^f\cap\mathcal{E}_C^g} D(\mu\|\rho) - \inf_{\mu\in\mathcal{E}_C^g} D(\mu\|\rho) \equiv 2\gamma_\epsilon\,.$$

For the same reasons as before, $\gamma_\epsilon > 0$.
In the general case, when $f = f(q,p)$, the Theorem of Sanov must be applied to the empirical measure $L_n' = \frac{1}{n}\sum_{j=1}^n \delta_\Delta \otimes \delta_{Y_j}$. □

DISCUTER LE PROBLEME $C \searrow \{c\}$ apres la limite thermodynamique.

CHAPTER 5

The Large Deviation Principle

The previous chapters have studied various theorems in which a random object, taking values in a given set, concentrates around its mean value. It is now time to take a step back and look for the common features of these theorems, and to gather them into a single general theory. We will then state general results that will shed some light on what has been seen until now.

This chapter follows the concise presentation of den Hollander [**?**].

1. Definition of the LDP

We have studied the empirical mean S_n/n taking values in $\mathbb{R}$ in the Theorem of Cramér, and the empirical measure L_n taking values in $\mathcal{M}_1(\mathbb{R})$ in the Theorem of Sanov. What happened to play an important was not the random object itself but its distribution: $\mu_n(\cdot) = P(S_n/n \in \cdot) \in \mathcal{M}_1(\mathbb{R})$ in the first case, $\mu_n(\cdot) = P(L_n \in \cdot) \in \mathcal{M}_1(\mathbb{A})$ in the second.

It thus seems natural to forget about the underlying random variables and to consider only a sequence of probability measures μ_n. To be sufficiently general, we assume that these measures live on a metric space $\mathcal{X}$ (with metric d), whose Borel σ-algebra we denote by $\mathcal{B}(\mathcal{X})$ The basic definition of the abstract theory will thus be that of the Large Deviation Principle for a sequence of probability measures (μ_n) on the metric space $\mathcal{X}$. This level of abstraction has apparently been settled by Varadhan [**?**].

We first state a few generalities about rate functions. We denote the open ball of radius $r > 0$ centered at x by $B_r(x) := \{y : d(y, x) < r\}$.

DEFINITION 5.1. *A function $f : \mathcal{X} \to [-\infty, \infty]$ is called* lower semicontinuous *if whenever $x_n, x \in \mathcal{X}$ are such that $x_n \to x$, then*

$$\liminf_{n\to\infty} f(x_n) \geq f(x)\,. \tag{79}$$

A level set of f is a set of the form $\{x : f(x) \leq c\}$, for some $c \in \mathbb{R}$.

LEMMA 5.1. *The following are equivalent.*

(1) *f is lower semicontinuous.*
(2) *for all $x \in \mathcal{X}$, $\lim_{r\searrow 0} \inf_{y\in B_r(x)} f(y) = f(x)$.*
(3) *f has closed level sets.*

PROOF. (1) implies (2): $\inf_{y\in B_\epsilon(x)} \leq f(x)$ is trivial. Let $\delta > 0$. There exists $y_\epsilon \in B_\epsilon(x)$ such that $\inf_{y\in B_\epsilon(x)} \geq f(y_\epsilon) - \epsilon$. Since $y_\epsilon \to x$ when $\epsilon \searrow 0$, lower semitontinuity gives

$$\lim_{\epsilon\searrow 0} \inf_{y\in B_\epsilon(x)} f(y) \geq \liminf_{\epsilon\searrow 0} f(y_\epsilon) - \delta \geq f(x) - \delta\,.$$

(2) implies (3): let $x_n \in \{f \leq c\}$, $x_n \to x$. Let $\epsilon_n := 2d(x_n, x)$. Then $x_n \in B_{\epsilon_n(x)}$, and

$$f(x) = \lim_{n\to\infty} \inf_{y\in B_{\epsilon_n}(x)} f(y) = \liminf_{n\to\infty} \inf_{y\in B_{\epsilon_n}(x)} f(y) \leq \liminf_{n\to\infty} f(x_n) \leq c\,,$$

and so $x \in \{f \leq c\}$.
(3) implies (1): assume the implication is wrong, i.e. there exists x, a sequence $x_n \to x$, and $\epsilon > 0$ such that $\liminf_n f(x_n) \leq f(x) - \epsilon < f(x)$. Consider the level set $L := \{y : f(y) \leq f(x) - \epsilon\}$. Then there exists a subsequence x_{n_k} such that $y_k := x_{n_k} \in L$ for large enough k. But $y_k \to x$, and $x \notin L$, and so L is not closed. □

LEMMA 5.2. *A lower semicontinuous function attains its infimum on any non-empty compact set $K \subset \mathcal{X}$.*

PROOF. Let K be non-empty, compact, and set $\lambda := \inf_{x\in K} f(x)$ (possibly $= -\infty$). There exists a sequence $x_n \in K$ such that $f(x_n) \searrow \lambda$. By compacity, there exists a subsequence n_k and $x_* \in K$ such that $x_{n_k} \to x_*$ when $k \to \infty$. We have $f(x_*) \geq \lambda$. But, by lower semicontinuity of f, $f(x_*) \leq \liminf_k f(x_{n_k}) = \lambda$. Therefore, $f(x_*) = \lambda$. □

DEFINITION 5.2. *$f : \mathcal{X} \to [0, \infty]$ is called a good rate function if*

(1) *$f \not\equiv \infty$,*
(2) *f is lower semicontinuous,*
(3) *f has compact level sets.*

The qualifyer *good* is associated to condition (3). Since in metric spaces, all compacts are closed, and by the equivalence of Lemma 5.1, (3) implies (2). One could therefore skip the requirement (2) in the above definition. Nevertheless, we leave it as it since a weaker form of rate function will be introduced below, droping (3).

We define a set-function on the subsets $A \subset \mathcal{X}$ by,

$$f(A) := \inf_{x\in A} f(x)\,.$$

DEFINITION 5.3. *Let (μ_n) be a sequence of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and $I : \mathcal{X} \to [0, \infty]$ be a rate function. (μ_n) satisfies a large deviation principle (LDP) with rate function I if*

(1) *for all closed set $F \subset \mathcal{X}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq -I(F)\,. \tag{80}$$

(2) *for all open set $G \subset \mathcal{X}$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq -I(G)\,. \tag{81}$$

Remark 5.1. The above defines a LDP **with speed** n. Nevertheless, everything that will be said in what follows holds for an arbitrary speed a_n, which is any sequence $a_n > 0$, $a_n \nearrow \infty$. This will be necessary for example in Chapter 11 when studying large deviations for the empirical measure of lattice systems. There, $a_n = |\Lambda_n|$, where Λ_n is a sequence of growing boxes.

Remark 5.2. We have already seen cases, like the Theorem of Sanov on finite alphabets, where no restriction was imposed on the sets of the upper bound.

Remark 5.3. One can wonder why the LDP is not defined by requiring that

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(B) = -I(B) \tag{82}$$

holds for all measurable set $B \subset \mathcal{X}$. Unfortunetaly, this is too restrictive. For instance, if each μ_n is non-atomic, i.e. $\mu_n(\{x\}) = 0$ for all $x \in \mathcal{X}$, then $I = \infty$ everywhere, and therefore (82) doesn't provide interesting information. Nevertheless, it makes sense to ask whether there exist measurable sets B for which (82) holds. It does clearly hold for the I-**continuous**, i.e. those that satisfy $I(\text{int}B) = I(\overline{B})$. As can be easily shown, if B is I-continuous, then (82) holds. Observe also that if I is continuous, then all sets that satisfy $B \subset \overline{\text{int}B}$ are I-continuous.

The following definition is natural:

Definition 5.4. *A sequence of $\mathbb{R}^d$-valued random variables $(Z_n)_{n\geq 1}$ (possibly defined on different probability spaces)* **large deviation principle (LDP) with rate function** $I : \mathbb{R}^d \to [0,\infty]$ *if the sequence of their distributions $\mu_n := \mu_{Z_n}$ satisfies a LDP with rate function I (in the sense of Definition 5.3, with $\mathcal{X} = \mathbb{R}^d$, and d the Euclidian metric).*

We then list some consequences of the LDP.

Lemma 5.3. *The rate function of a LDP is unique.*

Proof. Let I, J be two rate functions for the sequence (μ_n). Fix $x \in \mathcal{X}$. Then for all $\epsilon > 0$,

$$\begin{aligned} -I(x) \leq -I(B_\epsilon(x)) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\epsilon(x)) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(B_\epsilon(x)) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(\overline{B_\epsilon(x)}) \leq -J(\overline{B_\epsilon(x)}) \leq -J(B_{2\epsilon}(x))\,, \end{aligned}$$

since $\overline{B_\epsilon(x)} \subset B_{2\epsilon}(x)$. By the lower semicontinuity of J, $\lim_{\epsilon\searrow 0} J(B_{2\epsilon}(x)) = J(x)$, which shows that $I(x) \geq J(x)$. Similarly, $J(x) \geq I(x)$. □

Lemma 5.4. *There exists $x_* \in \mathcal{X}$ such that $I(x_*) = 0$.*

Observe that in the Theorem of Cramér, the infimum of the rate function was always attained at the unique point $m = E[X_1]$, giving the Law of Large Numbers. In general, however, x_* need not be unique. See for example the rate function of the two-dimensional Ising model in Chapter 10.

PROOF OF LEMMA 5.4: Using the large deviation bounds for the set $\mathcal{X}$, we get $I(\mathcal{X}) = 0$, which is clearly I-continuous. Therefore, there exists a sequence x_n such that $I(x_n) \to 0$. In particular, $x_n \in F := \{I \leq 1\}$ for large enough n. Since F is compact, there exists a subsequence X_{n_k} and $x_* \in F$ such that $x_{n_k} \to x_*$. By lower semicontinuity $0 \leq I(x_*) \leq \liminf_k f(x_{n_k}) = 0$. □

The following result will be used repeatedly in the sequel, in the particular in the study of upper bounds:

LEMMA 5.5. *Let $(a_n^{(i)})_{n\geq 1}$, $i = 1, \ldots, k$ be sequences of non-negative numbers. Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{k} a_n^{(i)} = \max_{i=1,\ldots,k} \limsup_{n\to\infty} \frac{1}{n} \log a_n^{(i)} . \tag{83}$$

PROOF. For simplicity, we prove the lemma considering the case $k = 2$. Let therefore $a_n, b_n \geq 0$ be two sequences. We have

$$\frac{1}{n}\log(a_n \vee b_n) \leq \frac{1}{n}\log(a_n + b_n) \leq \frac{1}{n}\log(a_n \vee b_n) + \frac{\log 2}{n} ,$$

which implies

$$\limsup_{n\to\infty} \frac{1}{n}\log(a_n + b_n) = \limsup_{n\to\infty} \frac{1}{n}\log(a_n \vee b_n)$$

But since log is increasing,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\log(a_n \vee b_n) &= \limsup_{n\to\infty} \Big(\frac{1}{n}\log a_n \vee \frac{1}{n}\log b_n\Big) \\ &= \Big(\limsup_{n\to\infty} \frac{1}{n}\log a_n\Big) \vee \Big(\limsup_{n\to\infty} \frac{1}{n}\log b_n\Big)\end{aligned}$$

□

1.1. The weak LDP.

DEFINITION 5.5. *$f : \mathcal{X} \to [0, \infty]$ is called a* weak rate function *if*

(1) *$f \not\equiv \infty$,*
(2) *f is lower semicontinuous.*

DEFINITION 5.6. *Let (μ_n) be a sequence of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and $I : \mathcal{X} \to [0, \infty]$ be a weak rate function. (μ_n) satisfies a* weak large deviation principle (WLDP) with speed n and rate function I *if*

(1) *for all compact set $K \subset \mathcal{X}$,*

$$\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(K) \leq -I(K) . \tag{84}$$

(2) *for all open set $G \subset \mathcal{X}$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq -I(G)\,. \tag{85}$$

A LDP can be recuperated from a WLDP under a concentration condition of (μ_n) on compact sets.

DEFINITION 5.7. *A sequence (μ_n) is* **exponentially tight** *if for all $M > 0$ there exists a compact set $K \subset \mathcal{X}$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(K^c) < -M\,. \tag{86}$$

LEMMA 5.6. *If (μ_n) satisfies a WLDP with weak rate function I, and if (μ_n) is exponentially tight, then I is a rate function, and (μ_n) satisfies a LDP with rate function I.*

PROOF. We first show the validity of the upper bound for closed sets F. Since the bound is trivial when $I(F) = 0$, we assume that $I(F) > 0$. Let $0 < b < I(F)$ (observe that $I(F)$ might be infinite). Let K_b be the compact set associated to b, for which (86) holds:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(K_b^c) < -b\,. \tag{87}$$

Then $\mu_n(F) \leq \mu_n(F \cap K_b) + \mu_n(K_b^c)$. Now $F \cap K_b$ is compact, and by the WLDP,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F \cap K_b) \leq -I(F \cap K_b) \leq -I(F) < -b\,. \tag{88}$$

Therefore, by Lemma 5.5,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq -b\,.$$

We can then take the limit $b \nearrow I(F)$. To show that the level sets of I are compact. it suffices to show that $\{I \leq b\} \subset K_b$ for all $b \geq 0$ (since a closed subset of a compact is also compact). Fix $b \geq 0$. Applying the lower large deviation bound to the open set K_b^c, we get $I(K_b^c) > b$, which implies $K_b^c \subset \{I > b\}$. □

LEMMA 5.7. *If $(\mathcal{X}, d)$ is complete and separable, and if (μ_n) satisfies a LDP, then (μ_n) is exponentially tight.*

PROOF. See [**?**]. □

2. The Varadhan Lemma

We now state and prove one of the main result of the chapter, which will allow to compute various objects when considering the thermodynamic limit, like the pressure.

THEOREM 5.1 (Varadhan Lemma). *Let (μ_n) satisfy a LDP with rate function I. Let $f : \mathcal{X} \to \mathbb{R}$ be continuous and bounded from above. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \int e^{nf(x)} \mu_n(dx) = \sup_{x\in\mathcal{X}} \bigl\{ f(x) - I(x) \bigr\}\,. \tag{89}$$

It will be useful to define the following sequence of finite measures on $\mathcal{X}$:

$$J_n(B) := \int_B e^{nf(x)} \mu_n(dx) .$$

PROOF. *Upper bound:* Consider the interval $[a, b]$, where

$$a := \sup_{x \in \mathcal{X}} \{f(x) - I(x)\} , \quad b := \sup_{x \in \mathcal{X}} f(x) .$$

Clearly, $-\infty \le a \le b < \infty$. Consider the closed set $F := f^{-1}([a, b])$. Fix some integer $N \ge 1$ and divide $[a, b]$ into N intervals of the form

$$\Delta_k^{(N)} := \Big[a + \frac{b-a}{N} k, a + \frac{b-a}{N}(k+1)\Big] , \quad k = 0, \dots, N-1 ,$$

to which correspond the closed sets $F_k^{(N)} := f^{-1}(\Delta_k^{(N)})$. We have, for all k,

$$\limsup_{n\to\infty} \frac{1}{n} \log J_n(F_k^{(N)}) \le \sup_{x \in F_k^{(N)}} f(x) - I(F_k^{(N)}) .$$

By Lemma 5.5,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log J_n(F) &\le \max_{k=1,\dots,N} \Big\{ \sup_{x \in F_k^{(N)}} f(x) - I(F_k^{(N)}) \Big\} \\
&\le \max_{k=1,\dots,N} \Big\{ \inf_{x \in F_k^{(N)}} f(x) - I(F_k^{(N)}) \Big\} + \frac{b-a}{N} \\
&\le \max_{k=1,\dots,N} \sup_{x \in F_k^{(N)}} \big\{ f(x) - I(x) \big\} + \frac{b-a}{N} \\
&= \sup_{x \in F} \big\{ f(x) - I(x) \big\} + \frac{b-a}{N} .
\end{aligned}$$

Take $N \to \infty$. Write then $J_n(\mathcal{X}) = J_n(F) + J_n(\mathcal{X} \backslash F)$. Of course $J_n(\mathcal{X} \backslash F) \le e^{an}$, and so using again Lemma 5.5,

$$\limsup_{n\to\infty} \frac{1}{n} \log J_n(\mathcal{X}) \le \sup_{x \in F} \big\{ f(x) - I(x) \big\} .$$

Lower bound: Fix $x \in \mathcal{X}$. For all $\epsilon > 0$, consider the open set $A := \{y : f(y) > f(x) - \epsilon\}$. We have

$$\liminf_{n\to\infty} \frac{1}{n} \log J_n(\mathcal{X}) \ge \liminf_{n\to\infty} \frac{1}{n} \log J_n(A) \ge f(x) - \epsilon - I(A) \ge f(x) - \epsilon - I(x) ,$$

which gives the lower bound after taking $\epsilon \to 0$ and supremizing over x. ☐

3. LDP for Tilted Measures

In this section and in the following, we "obtain new LDPs from old ones".

THEOREM 5.2 (LDP for Tilted measures). *Let (μ_n) be a sequence of probability measures on $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}$ be continuous and bounded from above, and define the sequence (ν_n) by*

$$\nu_n(dx) := \frac{e^{nf(x)}}{Z_n} \mu_n(dx) , \tag{90}$$

where $Z_n := \int e^{nf(x)}\mu_n(dx)$*. If* (μ_n) *satisfies a LDP with rate function* I*, then* (ν_n) *satisfies a LDP with rate function* J*, where*

$$J(x) = \sup_{y\in\mathcal{X}} \{f(y) - I(y)\} - \{f(x) - I(x)\} \tag{91}$$

$$= I(x) - f(x) - \inf_{y\in\mathcal{X}} \{I(x) - f(x)\}\,. \tag{92}$$

PROOF. Using the definition of J_n above, for all measurable set B,

$$\nu_n(B) = \frac{J_n(B)}{J_n(\mathcal{X})}\,.$$

The asymptotic behaviour of $J_n(\mathcal{X})$ was computed in the Lemma of Varadhan. For $J_n(\cdot)$, a similar analysis shows that

$$\limsup_{n\to\infty} \frac{1}{n}\log J_n(F) \le \sup_{x\in F}\{f(x) - I(x)\}$$

for all closed set F, whereas

$$\liminf_{n\to\infty} \frac{1}{n}\log J_n(G) \le \sup_{x\in G}\{f(x) - I(x)\}\,.$$

We verify that J is a good rate function: clearly $J \ge 0$, and $\not\equiv \infty$ since $a := \sup\{f - I\} \le \sup f < \infty$, and since $I \not\equiv \infty$. J is lower semicontinuous since I also is, and since f is continuous. Finally J has compact level sets because the closed set $\{J \le c\} \subset \{I - f \le c - a\} \subset \{I \le c - a + \sup f\}$, which is compact. □

4. The Contraction Principle

The Contraction Principle has been applied various times, for example when deriving the Cramér Theorem from the Theorem of Sanov for finite alphabets.

THEOREM 5.3 (Contraction Principle). *Let* (μ_n) *be a sequence of probability measures on* $\mathcal{X}$*. Let* $\mathcal{Y}$ *be another metric space and* $T : \mathcal{X} \to \mathcal{Y}$ *be continuous. Consider the* image measures $\nu_n := \mu_n \circ T^{-1}$*. If* (μ_n) *satisfies a LDP on* $\mathcal{X}$ *with function* I*, then* (ν_n) *satisfies a LDP on* $\mathcal{Y}$ *with rate function* J*, where*

$$J(y) = \inf_{x:Tx=y} I(x)\,. \tag{93}$$

PROOF. Let $F' \subset \mathcal{Y}$ be closed. Then $T^{-1}F' \subset \mathcal{X}$ is closed, and so

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\log \nu_n(F') = \limsup_{n\to\infty} \frac{1}{n}\log \mu_n(T^{-1}F') &\le -I(T^{-1}F') \\ &= -\inf_{y\in F'}\inf_{x\in T^{-1}\{y\}} I(x) \\ &\equiv -J(F')\,.\end{aligned}$$

The same can be done for an open set $G' \subset \mathcal{Y}$. To show that J has compact level sets, it suffices to observe that

$$\{J \le c\} = T(\{I \le c\})\,. \tag{94}$$

Namely, since $\{I \le c\}$ is compact and since the image of a compact by a continuous mapping is compact, we are done. To show (94), observe first that

$\{J \le c\} \supset T(\{I \le c\})$ is trivial. For the reverse inclusion, let $y \in \{J \le c\}$, i.e. $\lambda := \inf_{x \in T^{-1}\{y\}} I(x) \le c$. Let $x_n \in T^{-1}\{y\}$ be such that $I(x_n) \searrow \lambda$. Then clearly $x_n \in \{I \le 2\lambda\}$ for large n and since this set is compact, one can extract some subsequence x_{n_k} and some x_* such that $x_{n_k} \to x_*$. Since $Tx_* = \lim_k Tx_{n_k} = y$, and since $I(x_*) \le \liminf_k I(x_{n_k}) \le c$, we have that $y \in T\{I \le c\}$. $\square$

CHAPTER 6

The Curie-Weiss Model

In this chapter, we consider the first statistical mechanical model with interactions, the Curie-Weiss model, which is the simplest that can be described analytically, and whose large deviation properties fit nicely in the framework of the abstract results of Chapter 5. Although it is an oversimplified model of a ferromagnet (see the more realistic Ising model in further chapters), it gives a clear description of an important mechanism in the theory of phase transition, the energy-entropy argument.

Consider a system of n spins ± 1, whose set of configurations is

$$\Omega_n := \{\pm 1\}^n \equiv \{\sigma_\cdot : \{1, 2, \dots, n\} \to \{\pm 1\}\} .$$

Let $\mathcal{F}_n$ denote the discrete σ-algebra on Ω_n. As we know, if there are no interactions between the spins then the distribution describing them should be uniform. Let therefore $\rho = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ denote the uniform measure on $\{\pm 1\}$, and let $\rho_n := \rho^{\otimes n}$ be the product measure on $(\Omega_n, \mathcal{F}_n)$.

The interaction between the spins is a function $H_n : \Omega_n \to \mathbb{R}$, called hamiltonian, defined by

$$H_n^h(\sigma) := -\frac{1}{2n}\sum_{i=1}^n \sum_{j=1}^n \sigma_i\sigma_j - h\sum_{i=1}^n \sigma_i \quad \sigma \in \Omega_n . \tag{95}$$

This interaction is called ferromagnetic since the function $(\sigma_i, \sigma_j) \to -\sigma_i\sigma_j$ is lowest when the two spins are equal (aligned). The hamiltonian thus has two ground states, which are the minimizing configurations σ^+ (all the spins $\sigma_i^+ = +1$) and σ^- (all the spins $\sigma_i^+ = -1$).

We will consider the Gibbs distribution with inverse temperature $\beta > 0$, given by

$$\mu_n^{\beta,h}(d\sigma) := \frac{1}{Z_n^{\beta,h}} e^{-\beta H_n^{\beta,h}(\sigma)} \rho_n(d\sigma) , \quad \sigma \in \Omega_n , \tag{96}$$

where the partition function is given by

$$Z_n^{\beta,h} = \int_{\Omega_n} e^{-\beta H_n^{\beta,h}(\sigma)} \rho_n(d\sigma) \equiv 2^{-n} \sum_{\sigma \in \Omega_n} e^{-\beta H_n^{\beta,h}(\sigma)} .$$

We denote expectation with respect to $\mu_n^{\beta,h}$ by $E_n^{\beta,h}$. Observe that by symmetry, $E_n^{\beta,h}[\sigma_j] = E_n^{\beta,h}[\sigma_1]$ for all $1 \le j \le n$. Moreover, $E_n^{\beta,0}[\sigma_1] = 0$.

The main interest of the model lies in the fact that it exhibits a phase transition when $h = 0$, which from a probabilistic point of view is related to a breakdown of the Law of Large Numers. Let

$$m_n(\sigma) := \frac{1}{n}\sum_{j=1}^{n}\sigma_j$$

denote the empirical magnetization. Let $\beta_c := 1$ denote the critical inverse temperature of the model. We will show that

(1) if $\beta \leq \beta_c$, then the Law of Large Numbers holds: for all $\epsilon > 0$,

$$\mu_n^{\beta,0}\Big\{\Big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j - E_n^{\beta,0}[\sigma_1]\Big| \geq \epsilon\Big\} \to 0\,,$$

exponentially in n.

(2) if $\beta > \beta_c$, then the Law of Large Numbers is violated: for all small enough $\epsilon > 0$ (depending on β),

$$\mu_n^{\beta,0}\Big\{\Big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j - E_n^{\beta,0}[\sigma_1]\Big| \geq \epsilon\Big\} \to 1\,,$$

These two very different behaviours will be obtained by a large deviation analysis of the magnetization.

1. The LDP for the magnetization

The main simplyfing feature of the Curie-Weiss model is that the Hamiltonian can be written as a *function* of the magnetization:

$$H_n^h(\sigma) = -\frac{1}{2}nm_n(\sigma)^2 - hnm_n(\sigma) \equiv -nf(m_n(\sigma))\,, \tag{97}$$

where $f(z) := \frac{z^2}{2} + hz$. This has the advantage that the large deviation properties of the distribution of m_n can be studied without going through the (more cumbersome) large deviation properties of the Gibbs distribution. Namely, let $Q_n^{\beta,h}(\cdot) := \mu_n^{\beta,h}(m_n \in \cdot)$ be the distribution of m_n on $[-1,1]$. Then for all Borel set $B \subset [-1,1]$, a simple change of variable formula gives

$$\begin{aligned} Q_n^{\beta,h}(B) = \int_{m_n^{-1}(B)} \mu_n^{\beta,h}(d\sigma) &= \int_{m_n^{-1}(B)} \frac{e^{-\beta H_n^{\beta,h}(\sigma)}}{Z_n^{\beta,h}}\rho_n(d\sigma) \\ &= \int_{m_n^{-1}(B)} \frac{e^{n\beta f(m_n(\sigma))}}{Z_n^{\beta,h}}\rho_n(d\sigma) \\ &= \int_B \frac{e^{n\beta f(z)}}{Z_n^{\beta,h}}Q_n^0(dz)\,, \end{aligned}$$

where Q_n^0 is the distribution of m_n under ρ_n, and

$$Z_n^{\beta,h} = \int e^{n\beta f(z)}Q_n^0(dz)\,.$$

The distribution $Q_n^{\beta,h}$ is therefore a tilted measure of the form (90) in Theorem 5.2. Observe that the original sequence Q_n^0 satisfies a LDP by the Theorem of Cramér 3.3, with rate function

$$I(x) = \sup_{t \in \mathbb{R}} \{tx - \Lambda(t)\} \,,$$

where Λ is the logarithmic moment generating function of ρ, $\Lambda(t) = \log(\frac{1}{2}(e^t + e^{-t}))$. As seen in an exercise,

$$I(x) = \log 2 + \frac{1-x}{2} \log \frac{1-x}{2} + \frac{1+x}{2} \log \frac{1+x}{2} \,,$$

which is clearly a good rate function, strictly convex, with a unique minimum at $\frac{1}{2}$. By Theorem 5.2, $Q_n^{\beta,h}$ satisfies a LDP with rate function

$$\begin{aligned} J(z) &= \sup_{x \in [-1,1]} \{\beta f(x) - I(x)\} - \{\beta f(z) - I(z)\} \\ &= \{I(z) - \beta f(z)\} - \inf_{x \in [-1,1]} \{I(x) - \beta f(x)\} \end{aligned}$$

The asymptotic values of the magnetization are thus studied by the analysis of the function $z \mapsto \chi(z) := I(z) - \beta f(z)$. First,

$$\chi'(z) = \frac{1}{2} \log \frac{1+z}{1-z} - \beta(z+h) = \operatorname{arctanh}(z) - \beta(z+h) \,.$$

The derivative of χ thus vanishes at the points z solution of the mean field equation

$$\tanh(\beta(z+h)) = z \,. \tag{98}$$

This equation can be solved qualitatively (the case $h = 0$ is solved on Figure 1).

FIGURE 1. The solutions of the mean field equation (98) in the absence of magnetic field. a) When $\beta \le 1$, the straight line intersects the curve $\tanh(\beta z)$ at the only point $z = 0$, which is also the unique minimum of χ. b) When $\beta > 1$, the straight line intersects the curve $\tanh(\beta z)$ at the three point $z = \pm z_(\beta, h), 0$, of which only $\pm z(\beta, 0)$ are global minima of χ.

As can be seen geometrically, the particular value $\beta = \beta_c := 1$ of the inverse temperature plays an important role in the structure of the global minima of χ. This leads to the following information about the points where J attains its minima, i.e. where it equals zero.

- If $h \neq 0$, then J has a unique global minimum $z(\beta, h)$ for all $\beta > 0$, whose sign is the same as that of h.
- If $h = 0$ and $\beta \le \beta_c$, then $z = 0$ is the unique global minimum of J, whereas when $h = 0$ and $\beta > \beta_c$, then there exists two distinct global minima, $-z(\beta, 0) < 0 < z(\beta, 0)$.

We can thus answer the problem of the concentration of the magnetization and its dependence on (β, h). When $h \neq 0$, or when $h = 0$ and $\beta \leq \beta_c$, m_n satisfies a Law of Large Numbers, and converges to 0 exponentially rapidly: for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n^{\beta,h}(-\epsilon,\epsilon)^c = -\inf_{z\in(-\epsilon,\epsilon)^c} J(z) < 0\,.$$

On the other hand, if $h = 0$ and $\beta > \beta_c$, then

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n^{\beta,h}(K_{\beta,h}^c) = -\inf_{z\in K_{\beta,h}^c} J(z) < 0\,,$$

where $K_{\beta,h}$ is the set of points in $[-1, 1]$ that are at distance less than ϵ from either of the points $\pm z(\beta, h)$.

Stated in other words, we have the following complete description of the weak limit of $Q_n^{\beta,h}$ when $n \to \infty$. Let $\Rightarrow$ denote weak convergence on the set of probability distributions on $[-1, 1]$, and δ_x the Dirac mass at x.

Theorem 6.1.

$$Q_n^{\beta,h} \Rightarrow \begin{cases} \delta_0 & \text{if } h \neq 0 \text{ or } h = 0 \text{ and } \beta \leq \beta_c\,, \\ \frac{1}{2}\delta_{-z(\beta,h)} + \frac{1}{2}\delta_{+z(\beta,h)} & \text{if } h = 0 \text{ and } \beta > \beta_c\,. \end{cases} \tag{99}$$

Remark 6.1. It is interesting to notice that unlike all the other rate functions encountered up to now, J is *not convex* when β is large. Namely,

$$\chi''(z) = \frac{1}{1-z^2} - \beta\,.$$

Therefore, when $\beta \leq 1$, χ is strictly convex, but when $\beta > 1$, it is non-convex on the interval $-\sqrt{1-1/\beta} \leq z \leq \sqrt{1-1/\beta}$.

2. The free energy

FAIRE, en utilisant le Lemme de Varadhan.

CHAPTER 7

The Theorem of Cramér

As a consequence of the Theorem of Sanov, we have obtained a Large Deviation Principle for sequences of independent variables taking values in a finite alphabet. In this chapter (based on [?]) we show a more general result, for $\mathbb{R}$-valued random variables.

1. The logarithmic moment generating function

Remember that if X is any real valued random variable, its logarithmic moment generating function is defined by $\Lambda(t) := \log E[e^{tX}]$. Observe that $\Lambda(0) = 0$ and that $\Lambda : \mathbb{R} \to (-\infty, \infty]$. The Legendre Transform of Λ is defined by

$$\Lambda^*(x) := \sup_{t \in \mathbb{R}} \{tx - \Lambda(t)\} . \tag{100}$$

We have $\Lambda^*(x) \geq 0 \cdot x - \Lambda(0) = 0$ for all x.

LEMMA 7.1. *Λ and Λ^* are convex and lower semicontinuous.*

PROOF. The convexity of Λ follows from the Hölder inequality:

$$E\big[e^{(\lambda t_1 + (1-\lambda)t_2)X}\big] = E\big[(e^{t_1 X})^{\lambda} (e^{t_2 X})^{1-\lambda}\big] \leq E[e^{t_1 X}]^{\lambda} E[e^{t_2 X}]^{1-\lambda} .$$

The lower semicontinuity of Λ follows from the Lemma of Fatou: if $t_n \to t$,

$$\Lambda(t) = \log E[\lim_{n\to\infty} e^{t_n X}] \leq \log \liminf_{n\to\infty} E[e^{t_n X}] = \liminf_{n\to\infty} \log E[e^{t_n X}] = \liminf_{n\to\infty} \Lambda(t_n) .$$

The convexity of Λ^* is straightforward:

$$\begin{aligned} \Lambda^*(\lambda x_1 + (1-\lambda)x_2) &= \sup_t \big\{\lambda(tx_1 - \Lambda(t)) + (1-\lambda)(tx_2 - \Lambda(t))\big\} \\ &\leq \lambda \Lambda^*(x_1) + (1-\lambda)\Lambda^*(x_2) . \end{aligned}$$

For the lower semicontinuity, it suffices to remark that if $x_n \to x$, then for all n, $\Lambda^*(x_n) \geq tx_n - \Lambda(t)$, and so $\liminf_n \Lambda^*(x_n) \geq tx - \Lambda(t)$, which gives $\liminf_n \Lambda^*(x_n) \geq \Lambda^*(x)$. □

The effective domain of $f : \mathbb{R} \to (-\infty, \infty]$ is defined by $\mathcal{D}_f := \{t : f(t) < \infty\}$.

LEMMA 7.2. *Assume $m := E[X]$ exists.*

1. *$\Lambda^*(m) = 0$.*
2. *If $x \geq m$, then $\Lambda^*(x) = \sup_{t\geq 0}\{tx - \Lambda(t)\}$. Moreover, $x \mapsto \Lambda^*(x)$ is non-decreasing on $[m, \infty)$.*
3. *If $x \leq m$, then $\Lambda^*(x) = \sup_{t\leq 0}\{tx - \Lambda(t)\}$. Moreover, $x \mapsto \Lambda^*(x)$ is non-increasing on $(-\infty, m]$.*

(4) Λ *is* C^∞ *on* $\mathrm{int}\mathcal{D}_\Lambda$*, and there* $\Lambda'(t) = \frac{E[Xe^{tX}]}{E[e^{tX}]}$.

Proof. Using Jensen, $\Lambda(t) \geq E[\log e^{tX}] = tm$. Therefore, $\Lambda^*(m) \leq 0$, i.e. $\Lambda^*(m) = 0$. To show (2), observe that if $x \geq m$, $t \leq 0$, then $tx - \Lambda(t) \leq tm - \Lambda(t) \leq 0$, and so $\sup_{t\leq 0}\{tx - \Lambda(t)\} \leq 0$, which gives $\Lambda^*(x) = \sup_{t\geq 0}\{tx - \Lambda(t)\}$. Since $x \mapsto tx - \Lambda(t)$ is non-decreasing for all $t \geq 0$, then so is $x \mapsto \Lambda^*(x)$ for all $x \geq m$. Let $M(t) := E[e^{tX}]$. Let $t \in \mathrm{int}\mathcal{D}_\Lambda$. If μ_X denotes the distribution of X, then for small enough $\epsilon > 0$,

$$\frac{M(t+\epsilon) - M(t)}{\epsilon} = \int \frac{e^{(t+\epsilon)x} - e^{tx}}{\epsilon} \mu_X(dx)\,.$$

By dominated convergence,

$$M'(t) = \lim_{\epsilon\to 0} \frac{M(t+\epsilon) - M(t)}{\epsilon} = \int x e^{tx} \mu_X(dx) \equiv E[Xe^{tX}]\,.$$

Then, $\Lambda'(t) = M'(t)/M(t)$. C^∞-differentiability follows by induction. □

We know that $\Lambda(0) = 0 < \infty$. We now show that if finiteness of Λ holds in a neighbourhood of the origin, then Λ^* diverges at infinity. When $\mathcal{D}_\Lambda = \mathbb{R}$, this divergence is faster than linear.

Lemma 7.3. *If* $0 \in \mathrm{int}\mathcal{D}_\Lambda$*, then* $\lim_{x\to\pm\infty} \Lambda^*(x) = \infty$*, and therefore* Λ^* *has compact level sets. Moreover, if* $\mathcal{D}_\Lambda = \mathbb{R}$*, then* $\lim_{x\to\pm\infty} \Lambda^*(x)/|x| = \infty$.

Proof. We have, for all t, $\Lambda^*(x) \geq tx - \Lambda(t)$. Therefore,

$$\frac{\Lambda^*(x)}{|x|} \geq t\,\mathrm{sgn}(x) - \frac{\Lambda(t)}{|x|}\,.$$

If $0 \in \mathrm{int}\mathcal{D}_\Lambda$, then there exist $t_- < 0 < t_+$ such that $\Lambda(t_\pm) < \infty$. Therefore,

$$\liminf_{x\to+\infty} \frac{\Lambda^*(x)}{|x|} \geq t_+ > 0 \qquad \liminf_{x\to-\infty} \frac{\Lambda^*(x)}{|x|} \geq -t_- > 0\,.$$

Let $c \geq 0$ and $K := \{x : \Lambda^*(x) \leq c\}$. Since Λ^* is lower semicontinuous, K is closed. Since Λ^* diverges when $x \to \pm\infty$, K is bounded, hence compact. □

Therefore, if $0 \in \mathrm{int}\mathcal{D}_\Lambda$, Λ^* is a good rate function.

2. Main Theorem

We move on to the main result of this chapter.

Theorem 7.1 (Theorem of Cramér on $\mathbb{R}$). *Let* $X_1, X_2, \ldots$ *be an i.i.d. sequence such that* $m = E[X_1] < \infty$*. Let* Λ *denote the logarithmic moment generating function of* X_1 *and* Λ^* *its Legendre transform. Then,*

(1) *for all closed set* $F \subset \mathbb{R}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in F\Big) \leq -\inf_{x\in F} \Lambda^*(x)\,. \tag{101}$$

(2) *for all open set* $G \subset \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in G\Big) \geq - \inf_{x\in G} \Lambda^*(x) \,. \tag{102}$$

In particular, if $0 \in \mathrm{int}\mathcal{D}_\Lambda$, *then the sequence of distributions* $\mu_n(\cdot) := P(\frac{S_n}{n} \in \cdot)$ *satisfies a Large Deviation Principle with good rate function* Λ^*.

Proof. Let μ_n denote the distribution of $\frac{S_n}{n}$.
Upper bound: If $\inf_{x\in F} \Lambda^*(x) = 0$, the bound is trivial, so assume $\inf_{x\in F} \Lambda^*(x) > 0$. In particular, F does not contain m. Let $x \geq m$. For all $t \geq 0$, the Chebychev Inequality gives

$$\mu_n[x, \infty) = P(S_n \geq xn) = P(e^{tS_n} \geq e^{txn}) \leq \frac{E[e^{tS_n}]}{e^{txn}} = e^{-n(tx-\Lambda(t))} \,,$$

since $E[e^{tS_n}] = \prod_{j=1}^n E[e^{tX_j}] = e^{n\Lambda(t)}$. Therefore, by (2) of Lemma 3.4,

$$\mu_n[x, \infty) \leq e^{-n\Lambda^*(x)} \,.$$

In a similar was, if $x \leq m$, $\mu_n(-\infty, x] \leq e^{-n\Lambda^*(x)}$. Let then (x_-, x_+) denote the union of all open finite intervals containing m, not intersecting F. Then $x_- < m < x_+$, and possibly $x_- = -\infty$ or $x_+ = \infty$. We have $F \subset (-\infty, x_-] \cup [x_+, \infty)$, and so $\mu_n(F) \leq \mu_n(-\infty, x_-] + \mu_n[x_+, \infty)$. If $x_+ = \infty$ then $\mu_n[x_+, \infty) = 0$. Otherwise when $x_+ < \infty$ then $x_+ \in F$ and we have

$$\mu_n[x_+, \infty) \leq e^{-n\Lambda^*(x_+)} \leq e^{-n \inf_{x\in F} \Lambda^*(x)} \,.$$

In the same way, $\mu_n(-\infty, x_-] \leq e^{-n \inf_{x\in F} \Lambda^*(x)}$. This gives

$$\mu_n(F) \leq 2e^{-n \inf_{x\in F} \Lambda^*(x)} \,, \tag{103}$$

which implies the upper bound (101).

Remark 7.1. Observe that the proof of the upper bound has actually lead to a stronger result than (101), since (103) holds for *all* n.

Lower bound: Since G is open, then for each $x_0 \in G$, one can find some $\delta > 0$ such that $G \supset B_\delta(x_0)$ ($B_\delta(x)$ denotes the open interval os size 2δ centered at x). If we can show that

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_\delta(x_0)\Big) \geq -\Lambda^*(x_0) \,, \tag{104}$$

then the lower bound follows by infimizing over $x_0 \in G$. It will actually be sufficient to prove the inequality in the case where $x_0 = 0$.

Proposition 7.1. *Let* $X_1, X_2, \dots$ *be i.i.d. with logarithmic moment generating function* Λ. *Then for all* $\delta > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_\delta(0)\Big) \geq -\Lambda^*(0) \,. \tag{105}$$

Assume the proposition is true (the proof is given below). Set $\widetilde{X}_j := Y_j - x_0$. Then $\widetilde{X}_1$ has logarithmic generating function $\widetilde{\Lambda}(t) = \Lambda(t) - tx_0$, and so $\widetilde{\Lambda}^*(x) = \Lambda^*(x + x_0)$. Applying the proposition to the sequence $\widetilde{X}_k$,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_\delta(x_0)\Big) &= \liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{\widetilde{S}_n}{n} \in B_\delta(0)\Big) \\ &\geq -\widetilde{\Lambda}^*(0) \\ &= -\Lambda^*(x_0)\,.\end{aligned}$$

□

Proof of Proposition 7.1: To show (105), we need to consider separately the cases concerning the structure of the support of the distribution of X_1, denoted μ.

a) $\underline{\mu(-\infty, 0) > 0 \textit{ and } \mu(0, \infty) > 0, \textit{ and } \mu \textit{ has bounded support}}$: in this case, it is easy to verify that $\lim_{t\to\pm\infty} \Lambda(t) = +\infty$. Since we also have $\mathcal{D}_\Lambda = \mathbb{R}$, Λ is differentiable everywhere. As a consequence, there exists $t_* \in \mathbb{R}$ such that $\Lambda'(t_*) = 0$. Define then

$$\mu_*(dx) := \frac{e^{t_* x}}{Z_*} \mu(dx) \equiv e^{t_* x - \Lambda(t_*)} \mu(dx)\,,$$

where $Z_* = \int e^{t_* x} \mu(dx) = e^{\Lambda(t_*)}$. Let $X_1^*, X_2^*, \ldots$ be i.i.d. with distribution μ_*. Observe that by Lemma 7.2,

$$E[X_1^*] = \int x \mu_*(dx) = \frac{1}{Z_*} \int x e^{t_* x} \mu(dx) = \Lambda'(t_*) = 0.$$

Then, by independence of the X_js,

$$\begin{aligned}P\Big(\frac{S_n}{n} \in B_\delta(0)\Big) &= \int_{\frac{x_1+\cdots+x_n}{n} \in B_\delta(0)} \mu(dx_1)\ldots\mu(dx_n) \\ &= e^{n\Lambda(t_*)} \int_{\frac{x_1+\cdots+x_n}{n} \in B_\delta(0)} e^{-t_*(x_1+\cdots+x_n)} \mu_*(dx_1)\ldots\mu_*(dx_n) \\ &\geq e^{n\Lambda(t_*)} e^{-\delta|t_*|n} \int_{\frac{x_1+\cdots+x_n}{n} \in B_\delta(0)} \mu_*(dx_1)\ldots\mu_*(dx_n) \\ &= e^{n\Lambda(t_*)} e^{-\delta|t_*|n} P\Big(\frac{S_n^*}{n} \in B_\delta(0)\Big)\,,\end{aligned}$$

where $S_n^* = X_1^* + \cdots + X_n^*$. By the Law of Large Numbers,

$$P\Big(\frac{S_n^*}{n} \in B_\delta(0)\Big) \to 1\,.$$

In particular, for all $\delta' < \delta$,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_\delta(0)\Big) &\geq \liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_{\delta'}(0)\Big) \\ &\geq \Lambda(t_*) - \delta'|t_*| \\ &\geq \inf_t \Lambda(t) - \delta'|t_*| \\ &= -\Lambda^*(0) - \delta'|t_*|\,.\end{aligned}$$

And (105) holds by taking $\delta' \to 0$.

b) <u>$\mu(-\infty, 0) > 0$ *and* $\mu(0,\infty) > 0$, *and* μ *has unbounded support*</u>: Let $M > 0$ be large enough so that $\mu(-M, 0) > 0$ and $\mu(0, M) > 0$. Let $\mu_M(\cdot) := \mu(\cdot||x| \leq M)$. Let $X_1^{(M)}, X_2^{(M)}, \ldots$ be i.i.d. with distribution μ_M. We write

$$\begin{aligned}P\Big(\frac{S_n}{n} \in B_\delta(0)\Big) &\geq P\Big(\frac{S_n}{n} \in B_\delta(0)\Big| \bigcap_{j=1}^n \{|X_j| \leq M\}\Big) P\Big(\bigcap_{j=1}^n \{|X_j| \leq M\}\Big) \\ &= P\Big(\frac{S_n^{(M)}}{n} \in B_\delta(0)\Big)\mu[-M, M]^n \qquad (106)\end{aligned}$$

Now since μ_M has finite support the previous argument applies. The logarithmic moment generating function of μ_M is given by

$$\log \int e^{tx} \mu_M(dx) = \underbrace{\log \int_{-M}^{M} e^{tx} \mu(dx)}_{=:\Lambda_M(t)} - \log \mu[-M, M]\,.$$

By (106) and a), we thus have that for all large enough $M > 0$,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in B_\delta(0)\Big) &\geq \liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n^{(M)}}{n} \in B_\delta(0)\Big) + \log \mu[-M, M] \\ &\geq -\Lambda_M^*(0) = \inf_t \Lambda_M(t)\,.\end{aligned}$$

We must then verify that

$$\liminf_{M\to\infty} \inf_t \Lambda_M(t) \geq \inf_t \Lambda(t) = -\Lambda^*(0)\,.$$

It is sufficient to find any t_0 for which

$$\lambda := \liminf_{M\to\infty} \inf_t \Lambda_M(t) \geq \Lambda(t_0)\,. \qquad (107)$$

Since $\Lambda_M(t) \leq \Lambda_{M+1}(t)$ we actually have that $\inf_t \Lambda_M(t) \nearrow \lambda$. Since $\inf_t \Lambda_M(t) \leq \Lambda_M(0) \leq \Lambda(0) = 0$, we have that $-\infty < \lambda \leq 0$. Consider the sets $C_M := \{t : \Lambda_M(t) \leq \lambda\}$. These are non-empty, closed (since clearly each Λ_M is lower semi-continuous), and $C_{M+1} \subset C_M$. Therefore, $\bigcap_M C_M \neq \emptyset$. Let t_0 be any element of this intersection. Then $\Lambda_M(t_0) \leq \lambda$ for all M. Moreover, dominated convergence implies $\lim_M \Lambda_M(t_0) = \Lambda(t_0)$. We have therefore shown (107).

c) Either $\mu(-\infty,0)=0$, or $\mu(0,\infty)=0$: Assume for example that $\mu(-\infty,0)=0$. Then in this case $t\mapsto\Lambda(t)$ is non-decreasing. In particular,

$$-\Lambda^*(0)=\inf_t\Lambda(t)=\lim_{t\to-\infty}\Lambda(t)=\log\mu(\{0\})\,.$$

Now

$$P\Big(\frac{S_n}{n}\in B_\delta(0)\Big)\geq P\Big(\frac{S_n}{n}=0\Big)=\mu(\{0\})^n\,,$$

which gives

$$\liminf_{n\to\infty}\frac{1}{n}\log P\Big(\frac{S_n}{n}\in B_\delta(0)\Big)\geq\log\mu(\{0\})=-\Lambda^*(0)\,.$$

This finishes the proof of the proposition. □

3. The Theorem of Cramér in $\mathbb{R}^d$

In higher dimensions, the logarithmic moment generating function of a random variable $\mathbb{R}^d$-valued random variable X is defined by

$$\Lambda:=\log E[e^{\langle t,X\rangle}]\,,\quad\forall t\in\mathbb{R}^d\,,$$

where $\langle t,x\rangle:=\sum_{i=1}^d t_ix_i$ is the Euclidian scalar product. The Legendre transform of Λ is therefore given by

$$\Lambda^*(x):=\sup_{t\in\mathbb{R}^d}\{\langle t,x\rangle-\Lambda(t)\}\,.$$

The extension of Theorem 7.1 to $\mathbb{R}^d$ is as follows. Its proof can be found in [**?**].

Theorem 7.2 (Theorem of Cramér on $\mathbb{R}^d$). *Let $X_1,X_2,\dots$ be an i.i.d. sequence with values in $\mathbb{R}^d$, such that $m=E[X_1]$ exists. Let $\Lambda(t)$ denote the logarithmic moment generating function of X_1 and Λ^* its Legendre transform. Then,*

(1) *for all closed set $F\subset\mathbb{R}^d$,*

$$\limsup_{n\to\infty}\frac{1}{n}\log P\Big(\frac{S_n}{n}\in F\Big)\leq-\inf_{x\in F}\Lambda^*(x)\,. \tag{108}$$

(2) *for all open set $G\subset\mathbb{R}^d$,*

$$\liminf_{n\to\infty}\frac{1}{n}\log P\Big(\frac{S_n}{n}\in G\Big)\geq-\inf_{x\in G}\Lambda^*(x)\,. \tag{109}$$

As an interesting consequence, one can derive the Theorem of Sanov for finite alphabets as a corollary. Namely, ... COMPLETER

CHAPTER 8

The Ising Model (COMPLETER)

This chapter follows the notes by Velenik [**?**].

COMPLETER...

The Ising model is defined on the integer d-dimensional ($d \geq 1$) lattice

$$\mathbb{Z}^d = \{x = (x_1, x_2, \dots, x_d) : x_i \in \mathbb{Z}\, \forall i = 1, \dots, d\}\,, \tag{110}$$

equipped with the L^1-distance $d(x,y) := \sum_{i=1}^d |x_i - y_i|$. Subsets of $\mathbb{Z}^d$ will often be denoted $\Lambda \subset \mathbb{Z}^d$, and when Λ is finite we write $\Lambda \subset\subset \mathbb{Z}^d$. At each site $x \in \mathbb{Z}^d$ lives a spin $\omega(x)$ taking values in the set $\{\pm 1\}$, which we denote by Ω_0. Everything below holds when Ω_0 is an arbitrary finite alphabet. The space of configurations for the Ising model in infinite model is the cartesian product

$$\Omega := \Omega_0^{\mathbb{Z}^d}\,, \tag{111}$$

i.e. the set of all maps $\omega : \mathbb{Z}^d \to \Omega_0$. We will nevertheless start by considering the model in a volume $\Lambda \subset\subset \mathbb{Z}^d$ and later take the limit $\Lambda \subset \mathbb{Z}^d$. The Ising model can also be defined directly on $\mathbb{Z}^d$ with the help of a Gibbs specification; this will be done in Chapter 9.

Define, for any subset $\Lambda \subset \mathbb{Z}^d$, the product space $\Omega_\Lambda := \Omega_0^\Lambda$. We equip Ω_Λ with the discrete σ-algebra containing all subsets of Ω_Λ, denoted $\mathcal{P}(\Omega_\Lambda)$. The set of probability distributions on $(\Omega_\Lambda, \mathcal{P}(\Omega_\Lambda))$ is abreviated $\mathcal{M}_1(\Omega_\Lambda)$.

We define a Gibbs distribution $\mu_\Lambda^\omega \in \mathcal{M}_1(\Omega)$ as follows. Consider a summable potential ϕ, i.e. any bounded function $\phi : \mathbb{Z}^d \to \mathbb{R}$ such that $\phi(-x) = \phi(x)$ and

$$\sum_{x \in \mathbb{Z}^d} |\phi(x)| < \infty\,. \tag{112}$$

The potential is called ferromagnetic if $\phi \geq 0$. Let $\sigma \in \Omega_\Lambda$, $\omega \in \Omega$. The hamiltonian is defined by

$$H_\Lambda^\omega(\sigma) := -\sum_{\substack{\{x,y\} \subset \Lambda \\ x \neq y}} \phi(x-y)\sigma_x\sigma_y - h\sum_{x \in \Lambda} \sigma_x - \sum_{\substack{x \in \Lambda \\ y \in \Lambda^c}} \phi(x-y)\sigma_x\omega_y\,. \tag{113}$$

H_Λ^ω is always well defined thanks to (112). The Gibbs distribution at inverse temperature $\beta > 0$ with boundary condition ω is given by

$$\mu_\Lambda^\omega(\sigma) := \frac{e^{-\beta H_\Lambda^\omega(\sigma)}}{Z_\Lambda^\omega}\,. \tag{114}$$

When necessary, we will indicate the dependence of μ_Λ^ω on β and ϕ, by $\mu_{\Lambda;\beta,\phi}^\omega$.

The Ising model corresponds to the nearest-neighbour ferromagnetic potential

$$\phi(x) := \begin{cases} 1 & \text{if } d(0,x) = 1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{115}$$

1. The FKG Inequality

2. The thermodynamic limit

There will be two ways of defining the Ising model in infinite volume. The first will be by taking the thermodynamic limit, the second by describing an infinite system within the DLR formalism (see Chapter 9).

By thermodynamic limit we mean considering a sequence of boxes $\Lambda_n := [-n,n]^d \cap \mathbb{Z}^d$, and to take the limit $n \to \infty$. We usually denote this by the simple symbol $\lim_{\Lambda \nearrow \mathbb{Z}^d}$. An example of the thermodynamic limit would be to study the limit

$$\lim_{\Lambda \nearrow \mathbb{Z}^d} \langle \sigma_0 \rangle_{\Lambda;\beta,h}^+ \,.$$

More generally, one can imagine that a fairly good description of the system in infinite volume is given if all the limits

$$\lim_{\Lambda \nearrow \mathbb{Z}^d} \langle f \rangle_{\Lambda;\beta,h}^+ \,.$$

are known, for any local function f, i.e. functions that depend only on a finite number of spins. Assuming all these limits exist (that is, for each local function), one can wonder is there exists a probability measure μ on the whole space Ω such that

$$\mu(f) = \lim_{\Lambda \nearrow \mathbb{Z}^d} \langle f \rangle_{\Lambda;\beta,h}^+$$

for all local function f. Implementing this program requires a few definition.

2.1. Random fields.

We will define various σ-fields on Ω associated to subsets $\Lambda \subset \mathbb{Z}^d$ by introducing the usual notion of cylinder. Let $\Lambda \subset \Lambda'$ and consider the canonical projection $\Pi_\Lambda^{\Lambda'} : \Omega_{\Lambda'} \to \Omega_\Lambda$, defined as follows: for each $\omega_{\Lambda'} \in \Omega_{\Lambda'}$, the configuration $\Pi_\Lambda^{\Lambda'}(\omega_{\Lambda'}) \in \Omega_\Lambda$ is defined by $\Pi_\Lambda^{\Lambda'}(\omega_{\Lambda'})(t) := \omega_{\Lambda'}(t)$, $\forall t \in \Lambda$. These maps are obviously measurable with respect to the discrete σ-algebras we defined on the sets $\Omega_{\Lambda'}$ and Ω_Λ. When $\Lambda' = \mathbb{Z}^d$, we abbreviate $\Pi_\Lambda^{\mathbb{Z}^d} \equiv \Pi_\Lambda$. We have the obvious identity

$$\Pi_\Lambda = \Pi_\Lambda^{\Lambda'} \circ \Pi_{\Lambda'} \,. \tag{116}$$

For each $\Lambda \subset\subset \mathbb{Z}^d$, consider the σ-algebra

$$\mathcal{C}(\Lambda) := \{\Pi_\Lambda^{-1}(A) : A \in \mathcal{P}(\Omega_\Lambda)\}\,. \tag{117}$$

Each set $\Pi_\Lambda^{-1}(A)$ is called a cylinder. Cylinders have the property that

$$\mathcal{C}(\Lambda) \subset \mathcal{C}(\Lambda') \quad \text{when} \quad \Lambda \subset \Lambda' \,. \tag{118}$$

Indeed, (116) gives, for all $A \in \mathcal{P}(\Omega_\Lambda)$, $\Pi_\Lambda^{-1}(A) = \Pi_{\Lambda'}^{-1}\big((\Pi_\Lambda^{\Lambda'})^{-1}(A)\big)$. But $(\Pi_\Lambda^{\Lambda'})^{-1}(A) \in \mathcal{P}(\Omega_{\Lambda'})$, and therefore $\Pi_\Lambda^{-1}(A) \in \mathcal{C}(\Lambda')$. For any subset $S \subset \mathbb{Z}^d$, consider the union

$$\mathcal{C}_S := \bigcup_{\Lambda \subset\subset S} \mathcal{C}(\Lambda)\,. \tag{119}$$

As can be seen easily using (118), $\mathcal{C}_S$ is an algebra of subsets of Ω called the algebra of cylinder events in S. As can be seen easily, $\mathcal{C}_S$ it has countably many elements. The σ-algebra generated by cylinder events in S is then denoted

$$\mathcal{F}_S := \sigma(\mathcal{C}_S)\,. \tag{120}$$

In words, $\mathcal{F}_S$ is the σ-algebra of events $A = \{\omega\}$ which depend only on the values $\omega(x)$, $x \in \Lambda$. The largest algebra is when $S = \mathbb{Z}^d$, in which case we write $\mathcal{C} := \mathcal{C}_{\mathbb{Z}^d}$, called the algebra of cylinders. The largest σ-algebra is thus $\mathcal{F} := \sigma(\mathcal{C})$. The set of measures on the measurable space $(\Omega, \mathcal{F})$ is denoted $\mathcal{M}(\Omega, \mathcal{F})$, and the set of probability measures on $(\Omega, \mathcal{F})$, which we call random fields, is denoted $\mathcal{M}_1(\Omega, \mathcal{F})$.

2.2. Metric Structure and Quasilocality. For any two configurations $\omega, \sigma \in \Omega$, consider the distance [1]

$$d(\omega, \sigma) := \sum_{x \in \mathbb{Z}^d} 2^{-\|x\|} 1_{\{\omega(x) \neq \sigma(x)\}}\,. \tag{121}$$

Denote the family of open subsets of Ω with respect to this topology by $\mathcal{T}$. The facts stated in the following lemma will considerably simplify the study of Ising Random Fields.

LEMMA 8.1. (1) *The space* $(\Omega, \mathcal{T})$ *is compact.*
(2) *We have* $\mathcal{C} \subset \mathcal{T} \subset \mathcal{F}$. *In particular, if* $\mathcal{B}$ *denotes the Borel* σ*-algebra generated by* $\mathcal{T}$, *then* $\mathcal{B} = \mathcal{F}$.

The set of everywhere continuous functions $f : \Omega \to \mathbb{R}$ (with respect to the topology $\mathcal{T}$) is denoted $C(\Omega)$. Since $(\Omega, \mathcal{T})$ is compact, each continuous function is also uniformly continuous, bounded, and has compact support. Moreover, $C(\Omega)$ is a Banach space for the norm

$$\|f\| := \sup_{\omega \in \Omega} |f(\omega)|\,. \tag{122}$$

Observe also that since $\mathcal{T} \subset \mathcal{F}$ (Lemma 8.1), each continuous function is measurable. We shall use the notation $\mu(f) = \int f\, d\mu$. If $\mu(f) = \nu(f)$ for all $f \in C(\Omega)$, then $\mu = \nu$. Namely, since the cylinders are open and closed, their indicators are continuous, and therefore μ and ν coincide on $\mathcal{C}$, which by Carathéodory's Extension Theorem implies that $\mu = \nu$.

Another feature of the space Ω being compact is that continuous functions can be uniformly approximated by local function. Namely, call a function $f : \Omega \to \mathbb{R}$ local if there exists some $\Lambda \subset\subset \mathbb{Z}^d$ such that f is $\mathcal{F}_\Lambda$-measurable. We denote by

[1] Here, $\sum_{x \in \mathbb{Z}^d} := \lim_{n \to \infty} \sum_{x \in \Lambda_n}$.

$\mathcal{L}_{oc}(\Omega)$ the space of local functions on Ω. Therefore, $C(\Omega)$ is the closure of $\mathcal{L}_{oc}(\Omega)$ with respect to $\|\cdot\|$.
Probability measures can be studied by testing them on local functions rather than on continuous functions. Namely, if $f \in C(\Omega)$, let $f_n \in \mathcal{L}_{oc}(\Omega)$ be any sequence converging uniformly to f, $\lim_n \|f - f_n\| = 0$. Then, since $\|f_n\| \leq \|f_n - f\| + \|f\| < \infty$, we have $\mu(f) = \lim_n \mu(f_n)$ by the Dominated Convergence Theorem.

THEOREM 8.1 (Riesz-Markov Theorem on $\Omega_0^{\mathbb{Z}^d}$). *Consider a functional $L : C(\Omega) \to \mathbb{R}$ with the following properties:*

1. *Linearity: $L(f_1 + \alpha f_2) = L(f_1) + \alpha L(f_2)$ $\forall f_1, f_2 \in C(\Omega), \alpha \in \mathbb{R}$.*
2. *Positivity: if $f \geq 0$ then $L(f) \geq 0$.*
3. *Normalization: $L(1) = 1$.*

Then there exists a unique probability measure $\mu \in \mathcal{M}$ such that $L(f) = \mu(f)$ for all $f \in C(\Omega)$.

3. Proofs

PROOF OF LEMMA 8.1: The proof is a standard diagonalization argument, similar to what was done in the proof of Theorem 9.1. Consider a sequence $(\omega_n)_{n\geq 1}$ in Ω. Enumerate the points of $\mathbb{Z}^d$ in an arbitrary manner $x_1, x_2, \ldots$. First consider the sequence $(\omega_n(x_1))_{n\geq 1}$ in Ω_0. Since Ω_0 is finite, there exists a subsequence $(\omega_{n_k^1}(x_1))_{k\geq 1} \subset (\omega_n(x_1))_{n\geq 1}$ which takes a fixed constant value, call it $\omega^*(x_1)$, for all large enough k. Then consider the subsequence $(\omega_{n_k^1}(x_2))_{k\geq 1}$. Again one can extract a subsequence $(\omega_{n_k^2}(x_2))_{k\geq 1} \subset (\omega_{n_k^1}(x_2))_{k\geq 1}$ such that $\omega_{n_k^2}(x_2)$ takes a fixed value, say $\omega^*(x_2)$, for all large enough k. Continuing this process one obtains for all $j \geq 1$ a subsequence $(\omega_{n_k^j}(x_j))_{k\geq 1}$ such that $\omega_{n_k^j}(x_j) = \omega^*(x_j)$ for k large enough. By considering the diagonal subsequence $(\omega_{n_k^k})_{k\geq 1}$, one easily sees that for all $j \geq 1$, $\omega_{n_k^k}(x_j) = \omega^*(x_j)$ for large enough k, and therefore $d(\omega_{n_k^k}, \omega^*) \to 0$ when $k \to \infty$, which finishes the proof of the first part.
Since $(\Omega, \mathcal{T})$ is a metric space, the set of open balls $B(\omega; r) := \{\sigma : d(\omega, \sigma) < r\}$ forms a base [2] of the topology $\mathcal{T}$. We note the following fact: *for each cylinder B and for each $\omega \in B$, there exists an $r > 0$ such that $B(\omega; r) \subset B$.* Namely, assume the cylinder B has the form $B = \Pi_\Lambda^{-1}(A)$, with $A \in \mathcal{P}(\Omega_\Lambda)$. Let $\omega \in B$. Then, if $r_\omega > 0$ is small enough, any $\sigma \in B(\omega; r_\omega)$ coincides with ω on Λ, and therefore $\Pi_\Lambda(\sigma) \in A$, i.e. $\sigma \in B$. Therefore, on can express B as $B = \bigcup_{\omega\in B} B(\omega, r_\omega)$, i.e. $B \in \mathcal{T}$. This implies $\mathcal{C} \subset \mathcal{T}$.
Now consider the following fact: *for each ball $B(\omega; r)$ and for each $\sigma \in B(\omega; r)$ there exists a cylinder $B \subset B(\omega; r)$ containing σ* (in other words, the cylinders also form a base for $\mathcal{T}$). Namely, fix any $\sigma \in B(\omega; r)$. Take Λ large enough such that if σ' coincides with σ on Λ then $d(\sigma, \sigma') < \frac{r - d(\sigma,\omega)}{2}$. Define the cylinder $B := \Pi_\Lambda^{-1}(\sigma_\Lambda)$ $(\sigma_\Lambda = \Pi_\Lambda(\sigma))$. Clearly, $\sigma \in B$ and if $\sigma' \in B$ then $\Pi_\Lambda(\sigma') = \sigma_\Lambda$ and therefore $d(\sigma', \omega) \leq d(\sigma', \sigma) + d(\sigma, \omega) < r$, i.e. $B \subset B(\omega; r)$. Therefore,

[2] A collection of subsets $\mathcal{S} \subset \mathcal{T}$ is a base for $\mathcal{T}$ if and only if each $O \in \mathcal{T}$ can be expressed as a union of elements of $\mathcal{S}$.

each open ball can be expressed as a union of cylinders, and so each open $A \in \mathcal{T}$ can be written as a union of cylinders: $A = \bigcup_\alpha B_\alpha$. But since the cylinders are countable, this union is (at most) countable: $A = \bigcup_n B_n$. This implies that $\mathcal{T} \subset \mathcal{F}$, by the definition of $\mathcal{F}$. □

Proof of the Riesz-Markov Theorem: CHANGER LES NOTATIONS Observe first that a functional satisfying the conditions of the theorem is bounded: since $-\|f\| \leq f \leq \|f\|$ and Λ is normalized, we have $|L(f)| \leq \|f\|$. Therefore, if $f_n \in C(\Omega)$ is such that $\|f_n - f\| \to 0$, then $|L(f_n) - L(f)| = |L(f_n - f)| \leq \|f_n - f\| \to 0$. Therefore, L is continuous. Now, define, for each thin cylinder $B \in \mathcal{C}$,

$$\mu(B) := L(1_B)\,.$$

This definition extends to an arbitrary cylinder by summation. Then $\mu(B) \in [0, 1]$ for all cylinder, and if B_1, B_2 are two disjoint cylinders then $\mu(B_1 \cup B_2) = \mu(B_1) + \mu(B_2)$. Let B_n be a decreasing sequence of cylinders, with $\bigcap_n B_n = \emptyset$. Since cylinders are closed and Ω is compact, the Finite Intersection Property implies that $B_n = \emptyset$ for all large enough n. Therefore, $\lim_{n\to\infty} \mu(B_n) = 0$. By the Extension Theorem of Carathéodory, μ extends uniquely to a probability measure on $(\Omega, \mathcal{F})$. To show that $\mu(f) = L(f)$ for all $f \in C(\Omega)$, we first show that finite linear combinations of indicators of thin cylinders are dense in $C(\Omega)$. Take $f \in C(\Omega)$ and fix $\epsilon > 0$. Since f is also uniformly continuous, there exists $\delta > 0$ such that $d(x, y) < \delta$ implies $|f(x) - f(y)| < \epsilon$. Let $K \geq 1$ be such that $\sum_{k>K} 2^{-k} < \delta$. Consider the partition of Ω into thin cylinders $C(x^{(i)})$, $x^{(i)} \in \Omega_0^K$, where $C(x^{(i)}) := \{y \in \Omega : y_k = x_k^{(i)}, k = 1, \dots, K\}$. If $y \in C(x^{(i)})$, then $d(y, x^{(i)}) < \delta$ and so $|f(y) - f(x^{(i)})| < \epsilon$. Consider the linear combination $h := \sum_i f(x^{(i)}) 1_{C(x^{(i)})}$. Clearly, $\|h - f\| < \epsilon$, which proves the above claim. Now consider a sequence of linear combinations h_n as above such that $\|h_n - f\| \to 0$. Clearly, $\mu(h_n) = L(h_n)$. By Dominated Convergence ($\|h_n\| \leq \|f\| + \|h_n - f\|$) and continuity of L,

$$\mu(f) = \lim_{n\to\infty} \mu(h_n) = \lim_{n\to\infty} L(h_n) = L(f)\,,$$

which proves the theorem. □

CHAPTER 9

The DLR Formalism

In this chapter, we consider the description of infinite spin systems, in the DLR (Dobrushin-Lanford-Ruelle) Formalism. Different in spirit from the procedure adopted in Section 2 of the previous chapter, this approach characterizes probability measures on $\Omega = \{\pm 1\}^{\mathbb{Z}^d}$ using the notion of *specification*.

The presentation of the general formalism on specifications found below is a simplification of the book of Georgii [**?**]. The results that are of a purely technical flavor are proven at the end of the section.

1. Random Fields via Kolmogorov's Extension Theorem

Before moving onto specifications, we remind of a classical way of constructing random fields.

Given a random field μ and a finite volume Λ, define the marginal distribution of μ on Λ, which is a probability $\mu_\Lambda \in \mathcal{M}_1(\Omega_\Lambda)$ given by

$$\mu_\Lambda(A) := \mu(\Pi_\Lambda^{-1}(A)), \quad \forall A \in \mathcal{P}(\Omega_\Lambda). \tag{123}$$

A natural problem is to know if a random field can be reconstructed from it's family of marginal distributions $\{\mu_\Lambda, \Lambda \subset\subset \mathbb{Z}^d\}$.

A family of distributions $\{\mu_\Lambda\}_{\Lambda \subset\subset \mathbb{Z}^d}$, $\mu_\Lambda \in \mathcal{M}_1(\Omega_\Lambda)$, is said to be a consistent system of marginal distributions if for all $\Lambda' \subset\subset \mathbb{Z}^d$,

$$\mu_\Lambda = \mu_{\Lambda'} \circ (\Pi_\Lambda^{\Lambda'})^{-1}, \forall \Lambda \subset \Lambda'. \tag{124}$$

The following is known as Kolmogorov's Extension Theorem.

THEOREM 9.1. *Let $\{\mu_\Lambda\}_{\Lambda \subset\subset \mathbb{Z}^d}$, $\mu_\Lambda \in \mathcal{M}_1(\Omega_\Lambda)$, be a consistent system of marginal distributions. Then there exists a unique random field $\mu \in \mathcal{M}_1(\Omega, \mathcal{F})$ such that for all $\Lambda \subset\subset \mathbb{Z}^d$,*

$$\mu \circ \Pi_\Lambda^{-1} = \mu_\Lambda. \tag{125}$$

The simplest application of the preceding theorem is the construction of a product measure. Namely, let λ_0 be a probability distribution on $(\Omega_0, \mathcal{P}(\Omega_0))$. For each $\Lambda \subset\subset \mathbb{Z}^d$, define the finite product measure on $(\Omega_\Lambda, \mathcal{P}(\Omega_\Lambda))$ by $\lambda_\Lambda := \lambda_0^{\otimes \Lambda} \equiv \lambda_0 \otimes \cdots \otimes \lambda_0$. Then clearly $\{\lambda_\Lambda, \Lambda \subset\subset \mathbb{Z}^d\}$ is a consistent system of marginal distributions. The unique random field obtained via Kolmogorov's Extension Theorem is denoted $\lambda_0^{\otimes \mathbb{Z}^d}$.

Unfortunately, this method is not well suited for the construction of statistical mechanical models in infinite volume, like the Ising model. Namely, let $H_\Lambda(\sigma)$ denote the hamiltonian of the Ising in the volume Λ. Then the family (μ_Λ) of Gibbs distributions with free boundary condition is not compatible in the sense of (124) (this can be verified for volumes $\Lambda \subset \Lambda'$ with for example $|\Lambda| = 1$, $|\Lambda'| = 2$).

2. Random Fields via Specifications

In the previous section we have seen that a random field is uniquely determined by it's family of marginal distributions in finite sets. In the present section we consider the problem of defining a random field by specifying its family of conditional distributions in finite sets, as was first done by Dobrushin in [**?**].

As a motivation, let us consider the behaviour of a random field μ when it is conditionned on some event living *outside* a finite region Λ. This is natural in statistical mechanics, where one fixes boundary conditions. For any $\Lambda \subset\subset \mathbb{Z}^d$ and any event $A \in \mathcal{F}$, consider the conditional probability

$$\mu(A|\mathcal{F}_{\Lambda^c})(\omega) := E_\mu(1_A|\mathcal{F}_{\Lambda^c})(\omega)\,. \tag{126}$$

By definition,

$$\omega \mapsto \mu(A|\mathcal{F}_{\Lambda^c})(\omega) \text{ is } \mathcal{F}_{\Lambda^c}\text{-measurable} \tag{127}$$

and is defined only up to a set of μ-measure zero. Moreover, we have the following property. Let $A \in \mathcal{F}$, $B \in \mathcal{F}_{\Lambda^c}$. Then the properties of conditional expectation imply that μ-a.s.,

$$\mu(A \cap B|\mathcal{F}_{\Lambda^c}) = E_\mu(1_A 1_B|\mathcal{F}_{\Lambda^c}) = E_\mu(1_A|\mathcal{F}_{\Lambda^c})1_B = \mu(A|\mathcal{F}_{\Lambda^c})1_B\,. \tag{128}$$

On the other hand, if $\Lambda \subset \Delta$, then $\mathcal{F}_{\Delta^c} \subset \mathcal{F}_{\Lambda^c}$ and therefore, μ-a.s.,

$$\mu(\mu(A|\mathcal{F}_{\Delta^c})|\mathcal{F}_{\Lambda^c}) = \mu(A|\mathcal{F}_{\Delta^c})\,. \tag{129}$$

Since the conditional distributions $\mu(\cdot|\mathcal{F}_{\Lambda^c})$ play the same role as the marginals of the previous section, we ask whether it is possible to *reconstruct* the random field starting from the family $(\mu(\cdot|\mathcal{F}_{\Lambda^c}))_{\Lambda\subset\subset\mathbb{Z}^d}$, assuming that properties (127)-(129) hold. The nuisance is that these properties hold μ-almost surely, and to start we must define objects $\pi_\Lambda(A|\cdot)$, playing the role of the distributions $\mu(A|\mathcal{F}_{\Lambda^c})(\cdot)$, whose properties do not depend on any a priori given measure. Concerning the two first properties (127)-(128), this is done by using the notion of *probability kernel*.

Definition 9.1. *Let $\Lambda \subset\subset \mathbb{Z}^d$. A* **probability kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$** *is a map $\pi_\Lambda : \mathcal{F} \times \Omega \to [0,1]$ with the following properties:*

- *For each $A \in \mathcal{F}$, $\pi_\Lambda(A|\cdot)$ is $\mathcal{F}_{\Lambda^c}$-measurable.*
- *For each $\omega \in \Omega$, $\pi_\Lambda(\cdot|\omega)$ is a probability measure on $(\Omega, \mathcal{F})$.*

If, moreover,

$$\pi_\Lambda(A \cap B|\omega) = \pi_\Lambda(A|\omega)1_B(\omega) \tag{130}$$

for each $\omega \in \Omega$ and for each $A \in \mathcal{F}$, $B \in \mathcal{F}_{\Lambda^c}$, the probability kernel π is called **proper**.

All the probability kernels we will consider in the sequel will be proper. Observe that the properties imposed on π_Λ are stronger than what is actually needed, in the sense that we ask that for *each* ω, $\pi_\Lambda(\cdot|\omega)$ be a probability measure, which is in general not the case for $\mu(\cdot|\mathcal{F}_{\Lambda^c})(\omega)$. Moreover, observe that, unlike in (128), we require the properness property (130) to hold *for all* ω rather than just for μ-almost all ω.

LEMMA 9.1. *Condition* (130) *is equivalent to the following:*

$$\pi_\Lambda(B|\omega) = 1_B(\omega)\,, \quad \forall\, B \in \mathcal{F}_{\Lambda^c}\,. \tag{131}$$

The following lemma shows that a probability kernel π_Λ need actually be defined only on the cylinders $\Pi_\Lambda^{-1}(\sigma_\Lambda)$, for $\sigma_\Lambda \in \Omega_\Lambda$:

LEMMA 9.2. *Let π_Λ be proper. Then for all $A \in \mathcal{F}$ and all $\omega \in \Omega$,*

$$\pi_\Lambda(A|\omega) = \sum_{\sigma_\Lambda \in \Omega_\Lambda} \pi_\Lambda(\Pi_\Lambda^{-1}(\sigma_\Lambda)|\omega) 1_A(\sigma_\Lambda \omega_{\Lambda^c})\,. \tag{132}$$

As a consequence, if $f : \Omega \to \mathbb{R}$ be measurable and bounded, then

$$\int f(\sigma)\pi_\Lambda(d\sigma|\omega) = \sum_{\sigma_\Lambda \in \Omega_\Lambda} \pi_\Lambda(\sigma_\Lambda|\omega) f(\sigma_\Lambda \omega_{\Lambda^c})\,.\,. \tag{133}$$

This lemma shows that a proper probability kernel can be defined only through the probabilities of the cylinders $\Pi_\Lambda^{-1}(\sigma_\Lambda)$. Since no confusion is possible, we will denote from now on

$$\pi_\Lambda(\Pi_\Lambda^{-1}(\sigma_\Lambda)|\omega) \equiv \pi_\Lambda(\sigma_\Lambda|\omega)\,.$$

A probability kernel π_Λ allows to transform measures and functions. For any measurable bounded function $f : \Omega \to \mathbb{R}$, let $\pi_\Lambda f : \Omega \to \mathbb{R}$ be defined by:

$$\pi_\Lambda f(\omega) := \pi_\Lambda(f|\omega) \equiv \int \pi_\Lambda(d\sigma|\omega) f(\sigma)\,, \tag{134}$$

and for each $\mu \in \mathcal{M}_1(\Omega)$, define $\mu\pi_\Lambda \in \mathcal{M}_1(\Omega)$ by

$$\mu\pi_\Lambda(A) := \int \pi_\Lambda(A|\omega)\mu(d\omega)\,. \tag{135}$$

As can be easily verified, we have $\mu(\pi_\Lambda f) = \mu\pi_\Lambda(f)$ for all bounded function.

As we said before, we wish to characterize those random fields whose conditional distributions are described, in any finite Λ, by some probability kernel π_Λ. Like in the case of Kolmogorov's Extension Theorem, this will be possible under some compatibility assumption. Define, for any two probability kernels π_Λ, $\pi_{\Lambda'}$,

$$\pi_\Lambda\pi_{\Lambda'}(A|\omega) := \pi_\Lambda\big(\pi_{\Lambda'}(A|\cdot)|\omega\big) = \int \pi_\Lambda(d\eta|\omega)\pi_{\Lambda'}(A|\eta) \tag{136}$$

$$= \sum_{\eta_\Lambda} \pi_\Lambda(\eta_\Lambda|\omega)\pi_{\Lambda'}(A|\eta_\Lambda\omega_{\Lambda^c}) \tag{137}$$

It is immediate to verify that $\pi_\Lambda\pi_{\Lambda'}$ is a probability kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$. As suggested by the third condition (129), our kernels should satisfy $\pi_\Delta\pi_\Lambda = \pi_\Delta$ when $\Lambda \subset \Delta$. We thus call a family of probability kernels $\pi = (\pi_\Lambda)$ (for simplicity we omit in the notation to always mention that $\Lambda \subset\subset \mathbb{Z}^d$) compatible if $\pi_\Lambda \pi_{\Lambda'} = \pi_\Lambda$ for all $\Lambda' \subset \Lambda$.

DEFINITION 9.2. *A compatible family $\pi = (\pi_\Lambda)$ of proper probability kernels is called a specification.*

DEFINITION 9.3. *Let $\pi = (\pi_\Lambda)$ be a specification. A random field $\mu \in \mathcal{M}_1(\Omega)$ is said to be specified by π if, for all $A \in \mathcal{F}$, $\Lambda \subset\subset \mathbb{Z}^d$,*

$$\mu(A|\mathcal{F}_{\Lambda^c})(\cdot) = \pi_\Lambda(A|\cdot)\,, \quad \mu - a.s. \tag{138}$$

The set of random fields specified by π is denoted by $\mathcal{G}(\pi)$.

After the previous definition, the first natural question is: what conditions should a specification π satisfy in order to guaranty that $\mathcal{G}(\pi) \neq \emptyset$? We will give an answer to this question in Section 2.3.

Let us first give a criterium defining the random fields specified by π, easier to handle than (138) in concrete situations.

LEMMA 9.3. *Let $\pi = (\pi_\Lambda)$ be a specification, μ a random field. The following are equivalent:*

- $\mu \in \mathcal{G}(\pi)$
- $\mu\pi_\Lambda = \mu$ *for all* $\Lambda \subset\subset \mathbb{Z}^d$.

PROOF. If $\mu \in \mathcal{G}(\pi)$, then by definition of $\mu\pi_\Lambda$

$$\mu\pi_\Lambda(A) = \int \pi_\Lambda(A|\omega)\mu(d\omega) = \int \mu(A|\mathcal{F}_{\Lambda^c})\mu(d\omega) = \mu(A)\,.$$

Now assume $\mu\pi_\Lambda = \mu$. Fix $A \in \mathcal{F}$. Then for all $B \in \mathcal{F}_{\Lambda^c}$ one has, using (130),

$$\mu(A \cap B) = \mu\pi_\Lambda(A \cap B) = \int \pi_\Lambda(A \cap B|\omega)\mu(d\omega) = \int_B \pi_\Lambda(A|\omega)\mu(d\omega)\,,$$

and therefore $\pi_\Lambda(A|\cdot) = \mu(A|\mathcal{F}_{\Lambda^c})(\cdot)$ μ-almost surely. □

2.1. Main example: Gibbs specifications. The hamiltonian of the Ising model (113) can be written

$$H_\Lambda(\sigma) = -\sum_{\substack{\{x,y\}\subset\Lambda \\ x\neq y}} \phi(x-y)\sigma_x\sigma_y - \sum_{\substack{x\in\Lambda \\ y\in\Lambda^c}} \phi(x-y)\sigma_x\sigma_y\,. \tag{139}$$

If we define the functions $\Phi_{\{x,y\}}(\sigma) := -\phi(x-y)\sigma_x\sigma_y$, we thus have

$$H_\Lambda(\sigma) = \sum_{\{x,y\}\cap\Lambda\neq\emptyset} \Phi_{\{x,y\}}(\sigma)\,. \tag{140}$$

A more general form is thus

$$H_\Lambda(\sigma) := \sum_{\substack{B\cap\Lambda\neq\emptyset \\ |B|<\infty}} \Phi_B(\sigma)\,, \tag{141}$$

where for each $B \subset \Lambda$, $\Phi_B : \Omega \to \mathbb{R}$ is $\mathcal{F}_B$-measurable, i.e. depends only on the spins at sites $x \in B$. The collection $\Phi = (\Phi_B)$ is called a potential. In order for (141) to be well defined, we always assume that the potential is uniformly absolutely summable, i.e. for all finite $\Lambda \subset \mathbb{Z}^d$,

$$c := \sup_{x\in\mathbb{Z}^d} \sum_{\substack{B\ni x\\|B|<\infty}} \|\Phi_B\| < \infty .$$

We have $\sup_\sigma |H_\Lambda(\sigma)| \leq c|\Lambda|$. We wish to define a specification $\pi^\Phi = (\pi^\Phi_\Lambda)$ such that π^Φ_Λ gives to each cylinder $\Pi^{-1}_\Lambda(\sigma_\Lambda)$ a probability proportional to the Boltzmann weight prescribed by equilibrium statistical mechanics. That is, for a boundary condition $\omega \in \Omega$,

$$\pi^\Phi_\Lambda(\sigma_\Lambda|\omega) := \frac{\exp(-H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c}))}{Z^\Phi_\Lambda(\omega_{\Lambda^c})} ,$$

where $Z^\Phi_\Lambda(\omega_{\Lambda^c})$ is the partition function, given by

$$Z^\Phi_\Lambda(\omega_{\Lambda^c}) = \sum_{\sigma_\Lambda\in\Omega_\Lambda} \exp(-H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c})) . \tag{142}$$

This leads to the following definition: for all $A \in \mathcal{F}$,

$$\pi^\Phi_\Lambda(A|\omega) := \sum_{\sigma_\Lambda\in\Omega_\Lambda} \pi^\Phi_\Lambda(\sigma_\Lambda|\omega)1_A(\sigma_\Lambda\omega_{\Lambda^c}) . \tag{143}$$

Lemma 9.4. *$\pi^\Phi = (\pi^\Phi_\Lambda)$ is a specification.*

2.2. The Topology of Local Convergence. The standard notion of convergence for random fields is the following (the terminology used is natural after the remarks made in the previous paragraph):

Definition 9.4. *A sequence of random fields $(\mu_n)_{n\geq 1}$ converges weakly to a random field μ (denoted $\mu_n \Rightarrow \mu$) if $\mu_n(f) \to \mu(f)$ for each $f \in C(\Omega)$. Weak convergence generates a topology on $\mathcal{M}_1(\Omega, \mathcal{F})$ called the topology of local convergence.*

The topology of local convergence has different equivalent characterizations, given in the following lemma. We call a sequence $\Lambda_n \subset\subset \mathbb{Z}^d$ invading if it is increasing (i.e. $\Lambda_n \subset \Lambda_{n+1}$) and if $\bigcup_n \Lambda_n = \mathbb{Z}^d$.

Lemma 9.5. *Let μ and μ_n be random fields for all $n \geq 1$. The following are equivalent:*

(1) $\mu_n \Rightarrow \mu$
(2) *$\mu_n(B) \to \mu(B)$ for all cylinder $B \in \mathcal{C}$*
(3) *$\rho(\mu_n, \mu) \to 0$, where ρ is the distance defined by*

$$\rho(\mu, \nu) := \sup_{n\geq 1} \frac{1}{n} \max_{\omega_{\Lambda_n}\in\Omega_{\Lambda_n}} |\mu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n})) - \nu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n}))| , \tag{144}$$

where Λ_n is any chosen invading sequence.

The last statement of the lemma says that the topology of local convergence is metrizable. In the sequel all the topological considerations about random fields will be with respect to any of the previous equivalent characterizations of local convergence. In particular, THEOREM 9.2. $\mathcal{M}_1(\Omega, \mathcal{F})$ *is compact.*

2.3. Quasilocal Specifications. In this section we answer the first basic question about random fields: what should be assumed about a specification π in order to guarantee that there exists at least one random field μ specified by π? The answer will be positive under some continuity assumption on each $\pi_\Lambda(A|\omega)$ with respect to ω.

We start by defining the notion of continuity of a specification $\pi = (\pi_\Lambda)$ with respect to its boundary condition, considered by Dobrushin. For each $\Lambda \subset\subset \mathbb{Z}^d$, consider the distribution on Ω_Λ with boundary condition ω inherited from π, which we as usual abreviate by $\pi_\Lambda(\cdot|\omega)$. Say that π is D-continuous if, for all Λ,

$$\lim_{\Delta \nearrow \mathbb{Z}^d} \sup_{\substack{\omega,\eta: \\ \omega_\Delta = \eta_\Delta}} \|\pi_\Lambda(\cdot|\omega) - \pi_\Lambda(\cdot|\eta)\|_1 = 0 \,.$$

By $\Delta \nearrow \mathbb{Z}^d$ we mean taking the limit $n \to \infty$ along any invading sequence Λ_n. In words, π is D-continuous when the distributions $\pi_\Lambda(\cdot|\omega)$ depend weakly on the values taken by ω far away from Λ.

LEMMA 9.6. *If Φ is uniformly absolutely summable, then π^Φ is D-continuous.*

PROOF. Fix ω, η, and define, for all $t \in [0,1]$,

$$h_t(\sigma_\Lambda) := tH_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c}) + (1-t)H_\Lambda^\Phi(\sigma_\Lambda \eta_{\Lambda^c}) \,,$$

and $z_t := \sum_{\tau_\Lambda} e^{-h_t(\tau_\Lambda)}$. By writing

$$\begin{aligned} \pi_\Lambda^\Phi(\sigma_\Lambda|\omega) - \pi_\Lambda^\Phi(\sigma_\Lambda|\eta) &= \frac{e^{-h_1(\sigma_\lambda)}}{z_1} - \frac{e^{-h_0(\sigma_\lambda)}}{z_0} \\ &= \int_0^1 \frac{d}{dt} \frac{e^{-h_t(\sigma_\lambda)}}{z_t} \, dt \,, \end{aligned} \tag{145}$$

and since $\frac{d}{dt} h_t(\sigma_\Lambda) = H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c}) - H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})$, we easily get

$$\|\pi_\Lambda^\Phi(\cdot|\omega) - \pi_\Lambda^\Phi(\cdot|\eta)\|_1 \leq 2 \sup_{\sigma_\Lambda} |H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c}) - H_\Lambda^\Phi(\sigma_\Lambda \eta_{\Lambda^c})| \,,$$

which is small if ω and η coincide on some large set Δ. □

Let's go back to the general case, i.e. when [**?**], Dobrushin showed that $\mathcal{G}(\pi) \neq \emptyset$ when π is D-continuous. D-continuity happens to be equivalent to another, more tractable notion of continuity for specifications.

DEFINITION 9.5. *A specification $\pi = (\pi_\Lambda)$ is called* **quasilocal** [1] *if $f \in C(\Omega)$ implies $\pi_\Lambda f \in C(\Omega)$ for all $\Lambda \subset\subset \mathbb{Z}^d$.*

The verification that quasilocality is equivalent to D-continuity is left as an exercise. Now goes the existence theorem.

THEOREM 9.3. *Let π be a quasilocal specification. Then*

[1] Such specifications are also said to possess the Feller Property. In general, the Feller property is slightly stronger but in the case we are considering, where the spin space is finite, it is equivalent.

(1) $\mathcal{G}(\pi) \neq \emptyset$.
(2) $\mathcal{G}(\pi)$ *is closed. (In particular,* $\mathcal{G}(\pi)$ *is compact.)*

PROOF. Fix any $\omega \in \Omega$, consider an increasing sequence $\Lambda_n \nearrow \mathbb{Z}^d$, and define $\nu_n^\omega(\cdot) := \pi_{\Lambda_n}(\cdot|\omega)$. By compactness (Theorem 9.2), there exist a subsequence of $(\nu_n^\omega)_{n\geq 1}$ (which we assume, for simplicity, to be the sequence itself) and a random field ν^ω such that $\nu_n^\omega \Rightarrow \nu^\omega$. We verify that $\nu^\omega \in \mathcal{G}(\pi)$. Take again $f \in C(\Omega)$, and write

$$\begin{aligned} \nu^\omega \pi_\Lambda(f) = \nu^\omega(\pi_\Lambda f) &= \lim_{n\to\infty} \nu_n^\omega(\pi_\Lambda f) \\ &= \lim_{n\to\infty} \pi_{\Lambda_n}\pi_\Lambda(f|\omega) \qquad (146) \\ &= \lim_{n\to\infty} \pi_{\Lambda_n}(f|\omega) \qquad (147) \\ &= \lim_{n\to\infty} \nu_n^\omega(f) \equiv \nu^\omega(f) \end{aligned}$$

In (146) we used the Theorem of Fubini and in (147) we used consistency. Therefore, $\nu^\omega \pi_\Lambda = \nu^\omega$ for all finite Λ, and therefore $\nu^\omega \in \mathcal{G}(\pi)$. For the second affirmation, assume that $\mu_n \Rightarrow \mu$, where $\mu_n \in \mathcal{G}(\pi)$. Take any $f \in C(\Omega)$. By hypothesis, $\pi_\Lambda f \in C(\Omega)$. We thus get, for all $\Lambda \subset\subset \mathbb{Z}^d$,

$$\mu\pi_\Lambda(f) = \mu(\pi_\Lambda f) = \lim_{n\to\infty} \mu_n(\pi_\Lambda f) = \lim_{n\to\infty} \mu_n \pi_\Lambda(f) = \lim_{n\to\infty} \mu_n(f) = \mu(f)\,,$$

which implies $\mu\pi_\Lambda = \mu$, i.e. $\mu \in \mathcal{G}(\pi)$. □

It is an interesting and nontrivial question of knowing what is exactly the dependence of ν^ω on ω. For example, can each $\mu \in \mathcal{G}(\pi)$ be obtained by a suitable limiting procedure as we did in the previous proof? Observe that if $\mathcal{G}(\pi) = \{\mu\}$, that is when there is a unique random field specified by π, then $\nu_n^\omega \Rightarrow \mu$ for all ω, which means that the limiting random field obtained via weak limits is independent of the boundary condition ω.

3. The Ising model, again

On the other hand, sensitivity with respect to the boundary condition was the main concern in the study of the Ising model in Chapter 8.

$$\overline{\mathrm{co}\mathcal{G}(\beta,h)} \subset \mathcal{G}(\gamma^{\beta,h})$$

COMPLETER

4. An Inhomogeneous Ising chain on $\mathbb{N}$

As mentioned in the previous section, it is in general a difficult task to describe completely the convex set of Gibbs states $\mathcal{G}(\gamma^\Phi)$ associated to an absolutey uniformly convergent potential, even when Φ has finite range. In the present section, we consider a simple model where this can be done explicitely.

COMPLETER

5. Uniqueness; Dobrushin's Condition of Weak Dependence

In the present section we consider the problem of giving conditions on a quasilocal specification $\pi = (\pi_\Lambda)$ guaranteing that $\mathcal{G}(\pi)$ is a singleton. This criterium was first introduced by Dobrushin [?] and is called Dobrushin's Condition of Weak Dependence.

The criterium is formulated in terms of the one-site probability kernels $\pi_{\{x\}}$. With a slight abuse of notation we will denote by $\pi_x(\cdot|\omega)$ the distribution of the spin σ_x, induced by the kernel $\pi_{\{x\}}(\cdot|\omega)$. We consider the dependence of $\pi_x(\cdot|\omega)$ on the values of $\omega(z)$, $z \neq x$. Define

$$c_{x,y}(\pi) := \sup_{\substack{\omega(z)=\eta(z) \\ \forall z \neq y}} \|\pi_x(\cdot|\omega) - \pi_x(\cdot|\eta)\|_1 \tag{148}$$

The number $c_{x,y}(\pi)$ measures the sensivity of the distribution $\pi_x(\cdot|\omega)$ under changes in $\omega(y)$. Since we assume that π is quasilocal, we must have $c_{x,y}(\pi) \to 0$ when $\|x-y\| \to \infty$. The theorem below shows that if the convergence of the coefficients $c_{x,y}(\pi)$ to 0 is sufficient in order to turn them summable, then uniqueness is guaranteed. Let

$$c(\pi) := \sup_{x \in \mathbb{Z}^d} \sum_{y \in \mathbb{Z}^d} c_{x,y}(\pi)\,. \tag{149}$$

THEOREM 9.4. *Let π be a quasilocal specification satisfying* Dobrushin's Condition of Weak Dependence:

$$c(\pi) < 1\,. \tag{150}$$

Then $\mathcal{G}(\pi)$ contains exactly one random field: $\mathcal{G}(\pi) = \{\mu\}$.

The proof of this result is based on a fixed-point argument. One will define an operator T (depending on π) acting on quasilocal functions, with two main properties: 1) $\mu(\mathsf{T}f) = \mu(f)$ for all $\mu \in \mathcal{G}(\pi)$, and 2) when iterated infinitely many times, T turns any quasilocal function into a constant.

5.1. The Total Oscillation of a Function. Let $f : \Omega \to \mathbb{R}$. Define the local oscillation of f at x by:

$$\delta_x(f) := \sup_{\substack{\omega(z)=\eta(z) \\ \forall z \neq x}} |f(\omega) - f(\eta)| \tag{151}$$

The local oscillation measures the variation of $f(\omega)$ when one changes ω into another configuration by successive spin flips. Namely, assume $\omega_{\Lambda^c} = \eta_{\Lambda^c}$. Then it is easy to see that

$$|f(\omega) - f(\eta)| \leq \sum_{x \in \Lambda} \delta_x(f)\,. \tag{152}$$

It is thus natural to define the total oscillation of f by

$$\Delta(f) := \sum_{x \in \mathbb{Z}^d} \delta_x(f)\,. \tag{153}$$

We denote the space of functions with finite total oscillation by $\mathcal{E}_{TO}$. Since obviously $\mathcal{L}_{\text{oc}}(\Omega) \subset \mathcal{E}_{TO}$, we have that $\mathcal{E}_{TO}$ is dense in $C(\Omega)$. Intuitively, $\Delta(f)$

measures how far f is from being constant. This is made clear in the following lemma.

LEMMA 9.7. *Let* $f \in C(\Omega)$*. Then* $\Delta(f) \geq \sup f - \inf f$*.*

PROOF. Let $f \in C(\Omega)$. Since Ω is compact, there exist, for all $\epsilon > 0$, two configurations ω^+, ω^- such that $\omega^+_{\Lambda^c} = \omega^-_{\Lambda^c}$ for some sufficiently large box Λ, and such that $\sup f \leq f(\omega^+) + \epsilon$, $\inf f \geq f(\omega^-) - \epsilon$. Then, using (152),

$$\sup f - \inf f \leq f(\omega^+) - f(\omega^-) + 2\epsilon \leq \sum_{x \in \Lambda} \delta_x(f) + 2\epsilon \leq \Delta(f) + 2\epsilon \,,$$

which finishes the proof since ϵ can be chosen arbitrarily small. □

Another property of the oscillation is the following:

LEMMA 9.8. *If* $f_n \in C(\Omega)$ *is any sequence converging to* $f \in C(\Omega)$*,* $\lim_n \|f - f_n\| = 0$*, then* $\lim_{n\to\infty} \delta_x(f_n) = \delta_x(f)$*. As a consequence,*

$$\Delta(f) \leq \liminf_{n\to\infty} \Delta(f_n) \,. \tag{154}$$

PROOF. Consider two configurations ω^*, η^* such that $\omega^*(y) = \eta^*(y)$ for $y \neq x$, and $\delta_x(f) = |f(\omega^*) - f(\eta^*)|$. Let $\epsilon > 0$. Then for n large enough we have

$$\delta_x(f) = |f(\omega^*) - f(\eta^*)| \leq |f_n(\omega^*) - f_n(\eta^*)| + 2\epsilon \leq \delta_x(f_n) + 2\epsilon \,,$$

which implies $\delta_x(f) \leq \liminf_n \delta_x(f_n)$. In the same way, consider two sequences ω^*_n, η^*_n such that $\omega^*_n(y) = \eta^*_n(y)$ for $y \neq x$ and

$$\begin{aligned}
\delta_x(f_n) &= |f_n(\omega^*_n) - f_n(\eta^*_n)| \\
&\leq |f_n(\omega^*_n) - f(\omega^*_n)| + |f(\omega^*_n) - f(\eta^*_n)| + |f(\eta^*_n) - f_n(\eta^*_n)| \\
&\leq \|f_n - f\| + \delta_x(f) + \|f - f_n\| \,,
\end{aligned}$$

which implies $\limsup_n \delta_x(f_n) \leq \delta_x(f)$. The second claim follows from Fatou's Lemma. □

5.2. The Operator T**.** Enumerate the points of $\mathbb{Z}^d$ in an arbitrary way: $x_1, x_2, \ldots$ Any such enumeration has the property that for any $\Lambda \subset\subset \mathbb{Z}^d$, $x_n \notin \Lambda$ for large enough n. For $f \in C(\Omega)$, let $T_0 f := f$ and for $n \geq 1$, define (remember (134)):

$$T_n f(\omega) := \pi_{x_1} \pi_{x_2} \ldots \pi_{x_n} f(\omega) \,. \tag{155}$$

Since π is quasilocal we have $T_n f \in C(\Omega)$. This operator is linear, and has the following property: if $\mu \in \mathcal{G}(\pi)$, then

$$\mu(T_n f) = \mu(f) \,. \tag{156}$$

Moreover, as can be easily verified, $\|T_n f\| \leq \|f\|$.

PROPOSITION 9.1. *Let* $f \in C(\Omega)$*. Then the limit*

$$\mathsf{T} f := \lim_{n\to\infty} T_n f \tag{157}$$

exists, $\|\mathsf{T} f - T_n f\| \to 0$*, and therefore* $\mathsf{T} f \in C(\Omega)$*. Moreover,* $\|\mathsf{T} f\| \leq \|f\|$ *and* $\mu(\mathsf{T} f) = \mu(f)$ *for all* $\mu \in \mathcal{G}(\pi)$*.*

PROOF. We first define T on local functions. Let $g \in \mathcal{L}_{\mathrm{oc}}(\Omega)$. Then

$$\begin{aligned}\|T_{n+m}g - T_n g\| &= \|T_n(\pi_{x_{n+1}} \dots \pi_{x_{n+m}} g - g)\| \\ &\leq \|\pi_{x_{n+1}} \dots \pi_{x_{n+m}} g - g\| \,. \end{aligned} \tag{158}$$

Since g is local, there exists a finite Λ such that g can be expressed as in (169): $g = \sum_{\sigma_\Lambda} g_{\sigma_\Lambda} 1_{\Pi_\Lambda^{-1}(\sigma_\Lambda)}$. Therefore, if $x \notin \Lambda$, then each cylinder $\Pi_\Lambda^{-1}(\sigma_\Lambda) \in \mathcal{F}_{\{x\}^c}$, so that by the properness of π_x:

$$\pi_x g(\omega) = \sum_{\sigma_\Lambda \in \Omega_\Lambda} g_{\sigma_\Lambda} \pi_x(\Pi_\Lambda^{-1}(\sigma_\Lambda)|\omega) = \sum_{\sigma_\Lambda \in \Omega_\Lambda} g_{\sigma_\Lambda} 1_{\Pi_\Lambda^{-1}(\sigma_\Lambda)}(\omega) = g(\omega)\,.$$

Therefore, the right-hand side of (158) is zero when n is large enough. As a consequence, $\mathsf{T}g := \lim_{n\to\infty} T_n g$ exists. We also have $\mathsf{T}g \in C(\Omega)$ since the convergence $T_n g \to \mathsf{T}g$ is uniform, and $\|\mathsf{T}g\| \leq \|g\|$. Then take $f \in C(\Omega)$. Consider any sequence $g_k \in \mathcal{L}_{\mathrm{oc}}(\Omega)$ such that $\lim_n \|f - g_k\| = 0$, and set $\mathsf{T}f := \lim_k \mathsf{T}g_k$. This limit exists since

$$\|\mathsf{T}g_k - \mathsf{T}g_l\| \leq \limsup_{n\to\infty} \|T_n g_k - T_n g_l\| \leq \|g_k - g_l\|\,,$$

and clearly $\mathsf{T}f$ does not depend on the choice of the sequence g_k. We verify that $\mathsf{T}f = \lim_n T_n f$. Fix $\epsilon > 0$. Take k large enough such that $\|f - g_k\| \leq \epsilon$. Then we also have $\|\mathsf{T}f - \mathsf{T}g_k\| \leq \|f - g_k\| \leq \epsilon$. Then, take n sufficiently large (depending on k and ϵ) such that $\|T_n g_k - \mathsf{T}g_k\| \leq \epsilon$. Then we have

$$\|\mathsf{T}f - T_n f\| \leq \|\mathsf{T}f - \mathsf{T}g_k\| + \|\mathsf{T}g_k - T_n g_k\| + \|T_n g_k - T_n f\| \leq 3\epsilon\,.$$

The last claim follows by writing

$$\mu(\mathsf{T}f) = \mu(T_n f) + \mu(\mathsf{T}f - T_n f) = \mu(f) + \mu(\mathsf{T}f - T_n f)\,.$$

This last term goes to zero by dominated convergence, since $\|\mathsf{T}f - T_n f\| \to 0$. □

The usefulness of T and its relation to the number $c(\pi)$ as a contraction coefficient is given in the following proposition, which is the central result of this section.

PROPOSITION 9.2. *Assume $c(\pi) \leq 1$. Then, for any $f \in C(\Omega)$,*

$$\Delta(\mathsf{T}f) \leq c(\pi)\Delta(f)\,. \tag{159}$$

This inequality (159) is useful, of course, when $c(\pi) < 1$ and $f \in \mathcal{E}_{TO}$. In such case, T has the effect of strictly *reducing* the total oscillation of f.

PROOF OF THEOREM 9.4. Since, π is quasilocal, $\mathcal{G}(\pi) \neq \emptyset$ by Theorem 9.3, and the operator T is well defined. Choose any $\mu \in \mathcal{G}(\pi)$. Take $g \in \mathcal{L}_{\mathrm{oc}}(\Omega)$. Then

$$\mu(g) = \mu(\mathsf{T}g) = \dots = \mu(\mathsf{T}^k g) = \dots = \lim_{k\to\infty} \mu(\mathsf{T}^k g)\,. \tag{160}$$

But, by Proposition 9.2,

$$\Delta(\mathsf{T}^k g) \leq c(\pi)\Delta(\mathsf{T}^{k-1} g) \leq \dots \leq c(\pi)^k \Delta(g)\,.$$

Since $\mathcal{L}_{\mathrm{oc}}(\Omega) \subset \mathcal{E}_{TO}$, we have $\Delta(g) < \infty$, and since we assume $c(\pi) < 1$, we get $\lim_{k\to\infty} \Delta(\mathsf{T}^k g) = 0$. This implies that there exists a finite constant α_g such that $\|\mathsf{T}^k g - \alpha_g\| \to 0$ when $k \to \infty$. This follows from the fact that $\mathsf{T}^k g$ is bounded in k (since $\|\mathsf{T}^k g\| \le \|g\| < \infty$) and from Lemma 9.7. Therefore, going back to (160), the Dominated Convergence Theorem gives

$$\lim_{k\to\infty} \mu(\mathsf{T}^k g) = \mu(\lim_{k\to\infty} \mathsf{T}^k g) = \mu(\alpha_g) = \alpha_g\,.$$

If $\nu \in \mathcal{G}(\pi)$ is another random field specified by π, then in the same way we obtain $\nu(g) = \alpha_g$. Therefore $\mu = \nu$, which finishes the proof. □

Using Lemma 9.8, Proposition 9.2 is an immediate consequence of the following result, often called "dusting lemma".

LEMMA 9.9. *Assume $c(\pi) \le 1$. Then, for any $f \in C(\Omega)$* [2]*,*

$$\Delta(T_n f) \le c(\pi) \sum_{j=1}^{n} \delta_{x_j}(f) + \sum_{j>n} \delta_{x_j}(f)\,, \quad \forall n \ge 0\,. \tag{161}$$

PROOF. We prove (161) by induction on n. For $n = 0$, (161) holds since its right-hand side is just $\Delta(f)$. Therefore, assume the result has been proved for $n-1$ and all $f \in C(\Omega)$. We compute

$$\begin{aligned}\Delta(T_n f) = \Delta(T_{n-1}(\pi_{x_n} f)) &\le c(\pi) \sum_{j=1}^{n-1} \delta_{x_j}(\pi_{x_n} f) + \sum_{j>n-1} \delta_{x_j}(\pi_{x_n} f) \\ &= c(\pi) \sum_{j=1}^{n-1} \delta_{x_j}(\pi_{x_n} f) + \sum_{j>n} \delta_{x_j}(\pi_{x_n} f)\,, \end{aligned} \tag{162}$$

since $\delta_{x_n}(\pi_{x_n} f) = 0$. We are thus lead to study the local variations $\delta_{x_j}(\pi_{x_n} f)$ for $j \ne n$.

LEMMA 9.10. *For any $x \ne y$,*

$$\delta_x(\pi_y f) \le \delta_x(f) + c_{y,x}(\pi)\delta_y(f)\,. \tag{163}$$

PROOF. Let ω, η be such that $\omega_z = \eta_z$ for all $z \ne x$. If $\widehat{f} := f - m$, where m will be chosen below, we have by Lemma 9.2,

$$\begin{aligned}|\pi_y f(\omega) - \pi_y f(\eta)| &= |\pi_y \widehat{f}(\omega) - \pi_y \widehat{f}(\eta)| \\ &\le \sum_{\sigma_y} \left|\pi_y(\sigma_y|\omega)\widehat{f}(\sigma_y\omega_{\{y\}^c}) - \pi_y(\sigma_y|\eta)\widehat{f}(\sigma_y\eta_{\{y\}^c})\right|.\end{aligned}$$

Since $|\widehat{f}(\sigma_y\omega_{\{y\}^c}) - \widehat{f}(\sigma_y\eta_{\{y\}^c})| = |f(\sigma_y\omega_{\{y\}^c}) - f(\sigma_y\eta_{\{y\}^c})| \le \delta_x(f)$, we get

$$\begin{aligned}|\pi_y f(\omega) - \pi_y f(\eta)| &\le \delta_x(f) + \|\pi_y(\cdot|\omega) - \pi_y(\cdot|\eta)\|_1 \sup_{\sigma_y} |\widehat{f}(\sigma_y\omega_{\{y\}^c})| \\ &\le \delta_x(f) + c_{y,x}(\pi) \sup_{\sigma_y} |\widehat{f}(\sigma_y\omega_{\{y\}^c})| \\ &\le \delta_x(f) + c_{y,x}(\pi)\delta_y(f)\,,\end{aligned}$$

where in the last line we made the choice $m := f(+_y\omega_{\{y\}^c})$. □

[2] The first sum on the right-hand side is defined as 0 when $n = 0$.

Using (163) two times in (162), rearranging the terms and using $c(\pi)\leq 1$, we get

$$\begin{aligned}\Delta(T_n f) &\leq \Big[c(\pi)\sum_{j=1}^{n-1}\delta_{x_j}(f)+\sum_{j>n}\delta_{x_j}(f)\Big]+\delta_{x_n}(f)\sum_{j\neq n}c_{x_n,x_j}(\pi)\\ &\leq \Big[c(\pi)\sum_{j=1}^{n-1}\delta_{x_j}(f)+\sum_{j>n}\delta_{x_j}(f)\Big]+\delta_{x_n}(f)c(\pi)\,,\end{aligned}$$

which is exactly (161). □

5.3. Application: a Criterion for Gibbsian Specifications. Let us consider the form taken by the Dobrushin Condition of Weak Dependence when the specification π is Gibbsian, with a uniformly absolutely summable potential. Define the variation of f by $\delta(f):=\sup_{\omega,\eta}|f(\omega)-f(\eta)|$.

THEOREM 9.5. *Let $\Phi=(\Phi_B)$ be a uniformly absolutely summable potential, and π^Φ the associated Gibbsian specification. If*

$$\sup_{x\in\mathbb{Z}^d}\sum_{B\ni x}(|B|-1)\delta(\Phi_B)<\frac{1}{2}\,, \tag{164}$$

then π^Φ satisfies the Dobrushin Condition of Weak Dependence: $c(\pi^\Phi)<1$.

PROOF. Let ω,η be such that $\omega_z=\eta_z$ for all $z\neq y$. Using (145) with $\Lambda=\{x\}$,

$$\|\pi_x^\Phi(\cdot|\omega)-\pi_x^\Phi(\cdot|\eta)\|_1 \leq 2\sup_{\sigma_x}|H_x(\sigma_x\omega_{\{x\}^c})-H_x(\sigma_x\eta_{\{x\}^c})| \tag{165}$$

$$\leq 2\sum_{B\ni x}\sup_{\sigma_x}|\Phi_B(\sigma_x\omega_{\{x\}^c})-\Phi_B(\sigma_x\eta_{\{x\}^c})|\,. \tag{166}$$

Observe that the difference $\Phi_B(\sigma_x\omega_{\{x\}^c})-\Phi_B(\sigma_x\eta_{\{x\}^c})=0$ if B does not contain y. Therefore,

$$c_{x,y}(\pi^\Phi)\leq 2\sum_{B\ni x}1_B(y)\delta(\Phi_B)\,.$$

By summing over $y\neq x$,

$$c(\pi^\Phi)\leq 2\sup_x\sum_{y\neq x}\sum_{B\ni x}1_B(y)\delta(\Phi_B)=2\sup_x\sum_{B\ni x}(|B|-1)\delta(\Phi_B)\,.$$

□

ISING, DONNER UNE BORNE EXPLICITE SUR $\beta_c(d)$.

5.4. Proofs.

PROOF OF THEOREM 9.1. Let $B\in\mathcal{C}$, i.e. $B=\Pi_\Lambda^{-1}(A)\in\mathcal{C}(\Lambda)$ for some finite Λ, and $A\in\mathcal{P}(\Omega_\Lambda)$. Define

$$\mu_0(B):=\mu_\Lambda(A)\,. \tag{167}$$

We first verify that μ_0 is well defined, i.e. does not depend on the representation of B. Namely, assume B can be written in two different ways (this can happen because of (118)): $B=\Pi_{\Lambda_1}^{-1}(A_1)=\Pi_{\Lambda_2}^{-1}(A_2)$. Let $\Lambda':=\Lambda_1\cup\Lambda_2$. By (118), there exists $A'\in\mathcal{P}(\Omega_{\Lambda'})$ such that $\Pi_{\Lambda'}^{-1}(A')=\Pi_{\Lambda_1}^{-1}(A_1)=\Pi_{\Lambda_2}^{-1}(A_2)$. By (116),

$\Pi_{\Lambda_1}^{-1}(A_1) = \Pi_{\Lambda'}^{-1}((\Pi_{\Lambda_1}^{\Lambda'})^{-1}(A_1))$, and therefore $A' = (\Pi_{\Lambda_1}^{\Lambda'})^{-1}(A_1)$. Therefore, using the consistency relation (124),

$$\mu_{\Lambda'}(A') = \mu_{\Lambda'}((\Pi_{\Lambda_1}^{\Lambda'})^{-1}(A_1)) = \mu_{\Lambda_1}(A_1)\,.$$

Similarly, one shows that $\mu_{\Lambda'}(A') = \mu_{\Lambda_1}(A_1)$, and so (167) is independent of the representation of B. We then verify that μ_0 is a probability. Clearly, $\mu_0(\Omega) = \mu_\Lambda(\Omega_\Lambda) = 1$. To verify that μ_0 is finitely additive, let $B_1, B_2 \in \mathcal{C}$ be such that $B_1 \cap B_2 = \emptyset$. We know that there exists Λ such that for $i = 1, 2$, $B_i = \Pi_\Lambda^{-1}(A_i)$ for some $A_i \in \mathcal{P}(\Omega_\Lambda)$. We necessarily have $A_1 \cap A_2 = \emptyset$, and $B_1 \cup B_2 = \Pi_\Lambda^{-1}(A_1 \cup A_2)$. Therefore,

$$\mu_0(B_1 \cup B_2) = \mu_\Lambda(A_1 \cup A_2) = \mu_\Lambda(A_1) + \mu_\Lambda(A_1) = \mu_0(B_1) + \mu_0(B_2)\,.$$

We then must verify that μ_0 is continuous at $\emptyset$, that is: if $B_1 \supset B_2 \supset \dots$ is a decreasing sequence $B_n \in \mathcal{C}$ such that $\bigcap_n B_n = \emptyset$, then $\lim_n \mu_0(B_n) = 0$. Equivalently, we will show that if $B_1 \supset B_2 \supset \dots$ is a decreasing sequence $B_n \in \mathcal{C}$ such that $\mu_0(B_n) \geq \lambda > 0$, then $\bigcap_n B_n \neq \emptyset$. Since Ω is compact and since the cylinders are closed, this follows directly by the finite intersection property. The extension of μ_0 to $\mathcal{F}$ thus follows from the Extension Theorem of Carathéodory. The validity of (125) is immediate. □

Proof of Lemma 9.1. Condition (130) clearly implies (131) by taking $A = \emptyset$. Then, assume (131) holds. Let $A \in \mathcal{F}$, $B \in \mathcal{F}_{\Lambda^c}$. Then

$$\begin{aligned} \pi_\Lambda(A \cap B|\cdot) &\leq \min\{\pi_\Lambda(A|\cdot), \pi_\Lambda(B|\cdot)\} \\ &= \min\{\pi_\Lambda(A|\cdot), 1_B(\cdot)\} = \pi_\Lambda(A|\cdot)1_B(\cdot)\,. \end{aligned} \tag{168}$$

Similarly, since $A\backslash B = A \cap B^c$ we have

$$\pi_\Lambda(A\backslash B|\cdot) \leq \pi_\Lambda(A|\cdot)1_{B^c}(\cdot)\,.$$

Using the identity

$$\pi_\Lambda(A \cap B|\cdot) + \pi_\Lambda(A\backslash B|\cdot) = \pi_\Lambda(A|\cdot) = \pi_\Lambda(A|\cdot)1_B(\cdot) + \pi_\Lambda(A|\cdot)1_{B^c}(\cdot)\,,$$

we get

$$\pi_\Lambda(A \cap B|\cdot) = \pi_\Lambda(A|\cdot)1_B(\cdot) + \pi_\Lambda(A|\cdot)1_{B^c}(\cdot) - \pi_\Lambda(A\backslash B|\cdot) \geq \pi_\Lambda(A|\cdot)1_B(\cdot)\,.$$

This finishes the proof. □

Proof of Lemma 9.2: We start the proof for cylinder events of the form $A = \Pi_\Delta^{-1}(\eta_\Delta)$, which can always be expressed as

$$\Pi_\Delta^{-1}(\eta_\Delta) = \Pi_{\Delta\cap\Lambda}^{-1}(\eta') \cap \Pi_{\Delta\cap\Lambda^c}^{-1}(\eta'')\,,$$

where $\eta' \in \Omega_{\Delta\cap\Lambda}$ (resp. $\eta'' \in \Omega_{\Delta\cap\Lambda^c}$) is the restriction of η_Δ to $\Delta \cap \Lambda$ (resp. $\Delta \cap \Lambda^c$). Since $\Pi_{\Delta\cap\Lambda^c}^{-1}(\eta'') \in \mathcal{F}_{\Lambda^c}$ and π_Λ is proper,

$$\begin{aligned} \pi_\Lambda(A|\omega) &= \sum_{\sigma_\Lambda \in \Omega_\Lambda} \pi_\Lambda(A \cap \Pi_\Lambda^{-1}(\sigma_\Lambda)|\omega) \\ &= \sum_{\sigma_\Lambda \in \Omega_\Lambda} \pi_\Lambda(\Pi_{\Delta\cap\Lambda}^{-1}(\eta') \cap \Pi_\Lambda^{-1}(\sigma_\Lambda)|\omega)1_{\{\omega_{\Delta\cap\Lambda^c}=\eta''\}}\,. \end{aligned}$$

Since

$$\Pi^{-1}_{\Delta\cap\Lambda}(\eta')\cap\Pi^{-1}_{\Lambda}(\sigma_\Lambda)=\begin{cases}\Pi^{-1}_{\Lambda}(\sigma_\Lambda) & \text{if } \sigma_\Lambda \text{ coincides with } \eta' \text{ on } \Delta\cap\Lambda,\\ \emptyset & \text{otherwise,}\end{cases}$$

we have proved (132) for cylinders. To extend this to any event $A\in\mathcal{F}$, it suffices to prove that the collection

$$\mathcal{D}:=\Big\{A\in\mathcal{F}:\pi_\Lambda(A|\omega)=\sum_{\sigma_\Lambda\in\Omega_\Lambda}\pi_\Lambda(\Pi^{-1}_\Lambda(\sigma_\Lambda)|\omega)1_A(\sigma_\Lambda\omega_{\Lambda^c})\Big\}$$

is a Dynkin system, which is easy to verify. The second claim follows at once from the first. Namely, if $f=1_A$, then (133) is (132). The general case follows by standard approximation. □

Proof of Lemma 9.4: Since $\omega\mapsto H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c})$ is clearly $\mathcal{F}_{\Lambda^c}$-measurable, $\omega\mapsto\pi^\Phi_\Lambda(A|\omega)$ is $\mathcal{F}_{\Lambda^c}$-measurable. It is immediate that $\pi^\Phi_\Lambda(\cdot|\omega)$ is a probability measure. By Lemma 9.1, properness follows from the following fact: if $B\in\mathcal{F}_{\Lambda^c}$,

$$\begin{aligned}\pi^\Phi_\Lambda(B|\omega)&=\sum_{\sigma_\Lambda\in\Omega_\Lambda}\pi^\Phi_\Lambda(\sigma_\Lambda|\omega)1_B(\sigma_\Lambda\omega_{\Lambda^c})\\&=\sum_{\sigma_\Lambda\in\Omega_\Lambda}\pi^\Phi_\Lambda(\sigma_\Lambda|\omega)1_B(\omega_\Lambda\omega_{\Lambda^c})\equiv 1_B(\omega)\,.\end{aligned}$$

Consistency is straightforward although a little boring to write down. We consider two finite volumes $\Lambda\subset\Delta\subset\mathbb{Z}^d$, and show that $\pi^\Phi_\Delta\pi^\Phi_\Lambda=\pi^\Phi_\Delta$. Using (137) and Lemma 9.2,

$$\begin{aligned}\pi^\Phi_\Delta\pi^\Phi_\Lambda(A|\omega)&=\sum_{\sigma_\Delta}\pi^\Phi_\Delta(\sigma_\Delta|\omega)\pi_\Lambda(A|\sigma_\Delta\omega_{\Delta^c})\\&=\sum_{\sigma_\Delta}\sum_{\eta_\Lambda}1_A(\eta_\Lambda\sigma_{\Delta\setminus\Lambda}\omega_{\Delta^c})\pi^\Phi_\Delta(\sigma_\Delta|\omega)\pi^\Phi_\Lambda(\eta_\Lambda|\sigma_{\Delta\setminus\Lambda}\omega_{\Delta^c})\end{aligned}$$

We split the first sum in two, writing $\sigma_\Delta=\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}$. Using the definition of the kernels π^Φ_Δ and π^Φ_Λ, the above thus becomes

$$\sum_{\sigma''_{\Delta\setminus\Lambda}}\sum_{\eta_\Lambda}1_A(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\frac{e^{-H^\Phi_\Lambda(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}}{Z^\Phi_\Delta(\omega_{\Delta^c})Z^\Phi_\Lambda(\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}\sum_{\sigma'_\Lambda}e^{-H^\Phi_\Delta(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}\,.$$

Decomposing the hamiltonian in this last sum,

$$\begin{aligned}H^\Phi_\Delta(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})&=\sum_{B\cap\Delta\neq\emptyset}\Phi_B(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\\&=H^\Phi_\Lambda(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})+\sum_{\substack{B\cap\Delta\neq\emptyset\\B\cap\Lambda=\emptyset}}\Phi_B(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\\&=H^\Phi_\Lambda(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})+\sum_{\substack{B\cap\Delta\neq\emptyset\\B\cap\Lambda=\emptyset}}\Phi_B(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\,,\end{aligned}$$

where we have used the $\mathcal{F}_B$-measurability of Φ_B, and the fact that those Bs don't intersect Λ and thus don't depend on σ'_Λ. Therefore,

$$H^\Phi_\Delta(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c}) - H^\Phi_\Lambda(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c}) = H^\Phi_\Delta(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c}) - H^\Phi_\Lambda(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\,,$$

which gives

$$\sum_{\sigma'_\Lambda} e^{-H^\Phi_\Delta(\sigma'_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})} = Z^\Phi_\Lambda(\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})e^{-H^\Phi_\Delta(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}e^{-H^\Phi_\Lambda(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}\,.$$

Inserting this in the above expression, and renaming $\eta_\Lambda\sigma''_{\Delta\setminus\Lambda} \equiv \tau_\Delta$,

$$\begin{aligned}\pi^\Phi_\Delta\pi^\Phi_\Lambda(A|\omega) &= \sum_{\sigma''_{\Delta\setminus\Lambda}}\sum_{\eta_\Lambda} 1_A(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})\frac{e^{-H^\Phi_\Delta(\eta_\Lambda\sigma''_{\Delta\setminus\Lambda}\omega_{\Delta^c})}}{Z^\Phi_\Delta(\omega_{\Delta^c})}\\ &= \sum_{\tau_\Delta} 1_A(\tau_\Delta\omega_{\Delta^c})\frac{e^{-H^\Phi_\Delta(\tau_\Delta\omega_{\Delta^c})}}{Z^\Phi_\Delta(\omega_{\Delta^c})}\\ &\equiv \pi^\Phi_\Delta(A|\omega)\,.\end{aligned}$$

This shows that $\pi^\Phi = (\pi^\Phi_\Lambda)$ is a specification. $\square$

Proof of Lemma 9.5: (1) *implies* (2): Each cylinder $B \in \mathcal{C}$ is open. But since $\mathcal{C}$ is an algebra, B^c is also a cylinder, and thus also open and therefore 1_B is continuous. This implies $\mu_n(B) \to \mu(B)$.
(2) *implies* (1): Let $f \in C(\Omega)$. Fix $\epsilon > 0$ and take $g \in \mathcal{L}_{\text{oc}}(\Omega)$ such that $\|f-g\| \leq \epsilon$. Since g is $\mathcal{F}_\Lambda$-measurable for some finite Λ, it can be written as

$$g = \sum_{\sigma_\Lambda} g_{\sigma_\Lambda} 1_{\Pi^{-1}_\Lambda(\sigma_\Lambda)}\,, \tag{169}$$

where for example $g_{\sigma_\Lambda} := g(\sigma_\Lambda +_{\Lambda^c})$. Now, for all n,

$$\begin{aligned}|\mu(f)-\mu_n(f)| &\leq |\mu(f)-\mu(g)| + |\mu(g)-\mu_n(g)| + |\mu_n(g)-\mu_n(f)|\\ &\leq 2\epsilon + \sum_{\sigma_\Lambda} |g_{\sigma_\Lambda}||\mu(\Pi^{-1}_\Lambda(\sigma_\Lambda)) - \mu_n(\Pi^{-1}_\Lambda(\sigma_\Lambda))|\,.\end{aligned}$$

We thus have $\limsup_n |\mu(f)-\mu_n(f)| \leq 2\epsilon$, which finishes the proof since ϵ is arbitrary.
(2) *implies* (3): Fix $\epsilon > 0$ and take n_0 large enough such that $\frac{1}{n} \leq \epsilon$ for all $n > n_0$. Then, take k large enough such that

$$\max_{\omega_{\Lambda_n}} |\mu_k(\Pi_{\Lambda_n}(\omega_{\Lambda_n})) - \mu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n}))| \leq \epsilon\,, \quad \forall n \leq n_0\,.$$

We thus have, for all n,

$$\frac{1}{n}\max_{\omega_{\Lambda_n}} |\mu_k(\Pi_{\Lambda_n}(\omega_{\Lambda_n})) - \mu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n}))| \leq \begin{cases}\epsilon & \text{if } n \leq n_0\,,\\ \epsilon & \text{if } n > n_0\,.\end{cases}$$

This implies $\rho(\mu_k,\mu) \leq \epsilon$.
(3) *implies* (2): Assume $\rho(\mu_n,\mu) \to 0$. Take $B \in \mathcal{C}$, of the form $B = \Pi^{-1}_\Lambda(A)$ for

some $A \subset \Omega_\Lambda$. Then

$$|\mu_n(B) - \mu(B)| \leq \sum_{\omega_\Lambda \in A} |\mu_n(\Pi_\Lambda^{-1}(\omega_\Lambda)) - \mu(\Pi_\Lambda^{-1}(\omega_\Lambda))| \,,$$

and each term of the sum converges to zero when $n \to \infty$. Namely, we have $\mathcal{C}(\Lambda) \subset \mathcal{C}(\Lambda_n)$ for large enough n. □

Proof of Theorem 9.2: Since the topology of local convergence is metrizable, we need only show that $\mathcal{M}_1(\Omega, \mathcal{F})$ is sequentially compact. Consider any sequence $(\mu_n)_{n\geq 1}$ in $\mathcal{M}_1(\Omega, \mathcal{F})$. Since the set of cylinders $\mathcal{C}$ is countable, we can proceed exactly as in the proof of Lemma 8.1 to construct a subsequence $(\mu_{n_k})_{k\geq 1}$ such that $\mu_{n_k}(\Pi_\Lambda^{-1}(A)) \to \mu_\Lambda^*(A)$ for each $A \in \mathcal{P}(\Omega_\Lambda)$. Let us show that $\{\mu_\Lambda^* : \Lambda \subset\subset \mathbb{Z}^d\}$ is a consistent system of marginal distributions. As can be easily verified, each $\mu_\Lambda^* \in \mathcal{M}_1(\Omega_\Lambda)$. Then, take $\Lambda \subset \Lambda'$, and compute

$$\begin{aligned} \mu_\Lambda^*(A) &= \lim_{k\to\infty} \mu_{n_k}(\Pi_\Lambda^{-1}(A)) \\ &= \lim_{k\to\infty} \mu_{n_k}(\Pi_{\Lambda'}^{-1}((\Pi_\Lambda^{\Lambda'})^{-1}(A))) = \mu_{\Lambda'}^*((\Pi_\Lambda^{\Lambda'})^{-1}(A)) \,. \end{aligned}$$

By Kolmogorov's Extension Theorem 9.1, there exists a (unique) random field μ^* such that $\mu^*(\Pi_\Lambda^{-1}(A)) = \mu_\Lambda^*(A) = \lim_{k\to\infty} \mu_{n_k}(\Pi_\Lambda^{-1}(A))$. By Lemma 9.5, this implies $\mu_{n_k} \Rightarrow \mu^*$, and finishes the proof of the Theorem. □

CHAPTER 10

The Variational Principle

In this chapter, we present a variational characterization of translation invariant Gibbs measures. As will be seen in Chapter 11, this variational principle has a deep link with Large Deviation Theory. Most of the material presented here is taken from [**?**].

1. Introduction

The DLR formalism exposed in the previous chapter characterizes equilibrium measures, describing infinite systems, by a collection of local conditions. Namely, a Gibbs measure is defined by fixing its conditional distributions on all finite regions, through the help of a specification.

The idea behind the variational principle is of a different nature, and of more thermodynamical flavor: a functional $\mathcal{W}(\cdot)$ is defined on the set of probability distributions describing the system, of the form

$$\mathcal{W}(\mu) = \text{Entropy}(\mu) - \beta \times \text{Energy}(\mu)\,. \tag{170}$$

In thermodynamics, $-\frac{1}{\beta}\mathcal{W}(\mu)$ is called the free energy [1] of the state μ. The variational principle states that *the probability distributions that describe a system in thermal equilibrium are those that maximize its functional* $\mathcal{W}$.

We illustrate this on a simple example, in the case where the phase space of the system is a finite set. Let Λ be a finite region of $\mathbb{Z}^d$, and let $\mathcal{M}_1(\Omega_\Lambda)$ denote the set of all probability distributions on Ω_Λ. Let $U_\Lambda : \Omega_\Lambda \to \mathbb{R}$ be an energy function (which can for example be defined with the help of a potential, but this has no importance here). Define the functional $\mathcal{W}_\Lambda : \mathcal{M}_1(\Omega_\Lambda) \to \mathbb{R}$ as

$$\mathcal{W}_\Lambda(\mu) := \mathcal{H}_\Lambda(\mu) - \beta\langle U_\Lambda, \mu\rangle\,, \tag{171}$$

where $\mathcal{H}_\Lambda(\mu)$ is the Shannon Entropy of μ, and $\langle U_\Lambda, \mu\rangle := \mu(U_\Lambda)$ the energy of μ. Before showing that the Gibbs distribution is the only maximizer of $\mathcal{W}(\cdot)$, it is interesting to consider limit cases. If $\beta \searrow 0$ (high temperatures), the dominant term is the entropy term, and $\mathcal{W}_\Lambda$ is maximal for the uniform distribution. On the other hand, in the limit $\beta \nearrow \infty$ (low temperatures), the dominant term is the energy term, and $\mathcal{W}_\Lambda$ is maximal for distributions with a minimal energy.

[1] In textbooks, the Helmoltz free energy is usually found in the form $F = U - TS$, where U is the internal energy, T the temperature, and S the entropy.

This mechanism has already been encountered in the study of the large deviations of the Curie-Weiss model, where an explicit study of the functional (see page 51)

$$\chi(z) := I(z) - \beta f(z)\,, \quad z \in [-1,1]$$

has lead to different macroscopic features of the model, depending strongly on the inverse temperature β. The point there was that the entropy I was concave and the energy f was convex, giving a non-trivial dependence of the maximizers of χ on β. The Peierls argument for the two-dimensional Ising model in Chapter 8 was similar in spirit.

Coming back to our example, consider the Gibbs distribution $\mu^{\text{Gibbs}} \in \mathcal{M}_1(\Omega_\Lambda)$,

$$\mu^{\text{Gibbs}}(\sigma) := \frac{e^{-\beta U_\Lambda(\sigma)}}{Z_\Lambda}\,.$$

As can be easily verified, $\mathcal{W}_\Lambda(\mu^{\text{Gibbs}}) = \log Z_\Lambda$. We claim that

$$\mathcal{W}_\Lambda(\mu^{\text{Gibbs}}) = \sup_\mu \mathcal{W}_\Lambda(\mu)\,. \tag{172}$$

Namely, for any $\mu \in \mathcal{M}_1(\Omega_\Lambda)$, by the concavity of $x \mapsto \log x$ and Jensen's Inequality,

$$\begin{aligned}\mathcal{W}_\Lambda(\mu) &= \sum_{\omega\in\Omega_\Lambda} \mu(\omega) \log \frac{e^{-\beta U_\Lambda(\omega)}}{\mu(\omega)}\\ &\leq \log \sum_{\omega\in\Omega_\Lambda} \mu(\omega) \frac{e^{-\beta U_\Lambda(\omega)}}{\mu(\omega)} = \log Z_\Lambda = \mathcal{W}_\Lambda(\mu^{\text{Gibbs}})\,,\end{aligned}$$

with equality if and only if $\mu = \mu^{\text{Gibbs}}$.

The purpose of this chapter is to develop an equivalent theory in infinite volume, and to show that the infinite volume Gibbs measures on $\mathbb{Z}^d$, associated to a potential Φ, are exactly those probability measures that maximize a functional $\mathcal{W}$ on the space of probability measures.

As we remembered above when mentioning the Curie-Weiss model, variational problems like (172) appeared naturally in Large Deviation Theory. In fact, the extremality property of the Gibbs distribution in (172) should not be surprising since we know from Proposition 3.1 of Chapter 3 that the relative entropy $\mu \mapsto D(\mu\|\mu^{\text{Gibbs}})$ attains its minimum at the unique distribution μ^{Gibbs} but also that, as a consequence of its definition,

$$\begin{aligned}D(\mu\|\mu^{\text{Gibbs}}) &= \log Z_\Lambda - \{\mathcal{H}_\Lambda(\mu) - \beta\langle U_\Lambda, \mu\rangle\} \\ &\equiv \mathcal{W}_\Lambda(\mu^{\text{Gibbs}}) - \mathcal{W}_\Lambda(\mu) \geq 0\,.\end{aligned} \tag{173}$$

Maximizing the functional $\mathcal{W}_\Lambda$ is thus equivalent to minimizing the relative entropy. As a consequence, most of the development presented below consists in introducing a definition of relative entropy in infinite volumes, and its link with the extension of the thermodynamic limit of the terms on the right hand side of (173) to infinite volume limits. The precise link with Large Deviation Theory will be explained in Chapter 11.

In what follows, the inverse temperature β will be absorbed into the energy function: $\beta U_\Lambda \equiv U_\Lambda$.

2. The Entropy of an Invariant Random Field

In the case of finite alphabets, we defined the Shannon Entropy of a probability distribution in Chapter 2, and the relative entropy between two probability distributions appeared naturally in Chapter 3 when studying the Theorem of Sanov. For these notions make sense in infinite volume, it is necessary to restrict our study to systems that are *invariance under translations*. Before defining translation invariance, we remind the important definition of relative entropy, adapted to our needs.

Consider a random field $\mu \in \mathcal{M}_1(\Omega)$. For a finite $\Lambda \subset \mathbb{Z}^d$, denote the marginal of μ on Λ by $\mu_\Lambda = \pi_\Lambda \mu := \mu \circ \pi_\Lambda^{-1} \in \mathcal{M}_1(\Omega_\Lambda)$. The marginals of μ are compatible: if $\Lambda \subset \Delta$, then $\mu_\Lambda = \mu_\Delta \circ (\pi_\Lambda^\Delta)^{-1}$. If $\mu, \nu \in \mathcal{M}_1(\Omega)$, the relative entropy of μ with respect to the reference measure ν in Λ is defined by [2]

$$\mathcal{H}_\Lambda(\mu|\nu) := \begin{cases} \sum_{\sigma_\Lambda \in \Omega_\Lambda} \mu_\Lambda(\sigma_\Lambda) \log \frac{\mu_\Lambda(\sigma_\Lambda)}{\nu_\Lambda(\sigma_\Lambda)} & \text{if } \mu_\Lambda \ll \nu_\Lambda\,, \\ +\infty & \text{otherwise.} \end{cases} \tag{174}$$

Here, $\mu_\Lambda \ll \nu_\Lambda$ means that $\mu_\Lambda(\sigma_\Lambda) = 0$ each time $\nu_\Lambda(\sigma_\Lambda) = 0$. With the notation of Chapter 3, $\mathcal{H}_\Lambda(\mu|\nu) \equiv D(\mu_\Lambda \| \nu_\Lambda)$.

LEMMA 10.1. *If $\Lambda, \Delta \subset \mathbb{Z}^d$ are finite, $\mu, \nu \in \mathcal{M}_1(\Omega)$,*

(1) *$\mathcal{H}_\Lambda(\mu|\nu) \geq 0$ with equality if and only if $\mu_\Lambda = \nu_\Lambda$.*
(2) *$\mu \mapsto \mathcal{H}_\Lambda(\mu|\nu)$ is convex.*
(3) *If $\Lambda \subset \Delta$, then $\mathcal{H}_\Lambda(\mu|\nu) \leq \mathcal{H}_\Delta(\mu|\nu)$.*

PROOF. (1) and (2) were proved in Proposition 3.1, so we need only prove (3). If $\mu_\Delta \not\ll \nu_\Delta$ then the inequality is trivial since $\mathcal{H}_\Delta(\mu|\nu) = \infty$. On the other hand, if $\mu_\Delta \ll \nu_\Delta$, then $\mu_\Lambda \ll \nu_\Lambda$. Therefore, since we can always assume that the sum is over those σ_Δ such that $\nu_\Delta(\sigma_\Delta) > 0$,

$$\mathcal{H}_\Delta(\mu|\nu) = \sum_{\sigma_\Delta} \psi\Big(\frac{\mu_\Delta(\sigma_\Delta)}{\nu_\Delta(\sigma_\Delta)}\Big) \nu_\Delta(\sigma_\Delta)\,,$$

where $\psi(x) = x \log x$. Since $\Lambda \subset \Delta$, we can split the sum over $\sigma_\Delta = \omega_\Lambda \eta_{\Delta\setminus\Lambda}$. By writing (observe the abuse of notation) $\nu_\Delta(\omega_\Lambda \eta_{\Delta\setminus\Lambda}) = \nu_\Delta(\eta_{\Delta\setminus\Lambda}|\omega_\Lambda)\nu_\Delta(\omega_\Lambda)$, and

[2] Here and in what follows, the conventions are: $\log 0 = -\infty$, $0 \log \frac{0}{0} := 0$, $0 \log 0 := 0$.

since ψ is convex,

$$\begin{aligned}\mathcal{H}_\Delta(\mu|\nu) &= \sum_{\omega_\Lambda}\Big\{\sum_{\eta_{\Delta\setminus\Lambda}}\psi\Big(\frac{\mu_\Delta(\omega_\Lambda\eta_{\Delta\setminus\Lambda})}{\nu_\Delta(\omega_\Lambda\eta_{\Delta\setminus\Lambda})}\Big)\nu_\Delta(\eta_{\Delta\setminus\Lambda}|\omega_\Lambda)\Big\}\nu_\Delta(\omega_\Lambda)\\ &\geq \sum_{\omega_\Lambda}\Big\{\psi\Big(\sum_{\eta_{\Delta\setminus\Lambda}}\frac{\mu_\Delta(\omega_\Lambda\eta_{\Delta\setminus\Lambda})}{\nu_\Delta(\omega_\Lambda\eta_{\Delta\setminus\Lambda})}\nu_\Delta(\eta_{\Delta\setminus\Lambda}|\omega_\Lambda)\Big)\Big\}\nu_\Delta(\omega_\Lambda)\\ &= \sum_{\omega_\Lambda}\psi\Big(\frac{\mu_\Delta(\omega_\Lambda)}{\nu_\Delta(\omega_\Lambda)}\Big)\nu_\Delta(\omega_\Lambda)\\ &\equiv \mathcal{H}_\Lambda(\mu|\nu)\,.\end{aligned}$$

Namely, $\mu_\Delta(\omega_\Lambda) \equiv \mu_\Delta((\pi_\Lambda^\Delta)^{-1}(\omega_\Lambda)) = \mu_\Lambda(\omega_\Lambda)$. □

Since the relative entropy is non-decreasing in the volume, it is tempting to define the specific relative entropy of μ with respect to ν by

$$h(\mu|\nu) := \lim_{n\to\infty}\frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu)\,. \tag{175}$$

Here and in the sequel, Λ_n always denotes the sequence of boxes $[-n,n]^d \cap \mathbb{Z}^d$. The above definition will be shown to make sense only when the two measures μ and ν are translation invariant (see below). We will then be able to show the existence of the limit (175) when the reference measure ν is 1) the uniform product measure, 2) a Gibbs measure associated to a potential. We now introduce the notation needed in order to proceed.

For $x \in \mathbb{Z}^d$, consider the translation $\theta_x : \mathbb{Z}^d \to \mathbb{Z}^d$ defined by $\theta_x y := x + y$. Observe that $\theta_x^{-1} = \theta_{-x}$, and that $\theta_x\theta_y = \theta_{x+y}$. The set of translations of the lattice $\mathbb{Z}^d$ (which actually forms a group) is denoted $\theta := (\theta_x)_{x\in\mathbb{Z}^d}$. We use the same symbol θ_x to denote the action of the translation on configurations as $(\theta_x\omega)_y := \omega_{y-x}$, on functions $f : \Omega \to \mathbb{R}$ as $\theta_x f(\omega) := f(\theta_x\omega)$, and on measures, as $\theta_x\mu(A) := \mu(\theta_x^{-1}A)$. A probability measure μ is called (translation) invariant if $\theta_x\mu = \mu$ for all $x \in \mathbb{Z}^d$. The set of invariant probability measures is denoted $\mathcal{M}_{1,\theta}(\Omega)$. As can be verified easily, μ is invariant if and only if $\mu(\theta_x f) = \mu(f)$ for all $f \in C(\Omega)$ and all x. Throughout the chapter, the topology considered on $\mathcal{M}_1(\Omega)$ is the topology of weak convergence. Observe that $\mathcal{M}_{1,\theta}(\Omega) \subset \mathcal{M}_1(\Omega)$ is closed, and therefore compact.

Let $\lambda = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ denote the uniform distribution on $\{\pm 1\}$. For all $S \subset \mathbb{Z}^d$, let $\lambda_S := \lambda^{\otimes S}$. For simplicity, we write $\lambda_{\mathbb{Z}^d} \equiv \lambda$. Observe that

$$\mathcal{H}_\Lambda(\mu|\lambda) = |\Lambda|\log 2 - \mathcal{H}_\Lambda(\mu)\,, \tag{176}$$

where $\mathcal{H}_\Lambda(\mu)$ is the Shannon entropy of μ in Λ,

$$\mathcal{H}_\Lambda(\mu) := -\sum_{\sigma_\Lambda\in\Omega_\Lambda}\mu_\Lambda(\sigma_\Lambda)\log\mu_\Lambda(\sigma_\Lambda)\,. \tag{177}$$

Therefore, when the reference measure is λ, showing the existence of (175) boils down to

THEOREM 10.1. *For all $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, the limit*

$$h(\mu) := \lim_{n\to\infty} \frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu) \tag{178}$$

exists, is finite, and is called the **specific entropy of** μ. *Moreover, the map h : $\mathcal{M}_{1,\theta}(\Omega) \to \mathbb{R}$ is affine and upper semicontinuous* [3], *and for all $c \in \mathbb{R}$, $\{h \geq c\}$ is compact.*

The proof of (178) relies on the two lemmas found below. The first shows that $\Lambda \mapsto \mathcal{H}_\Lambda(\mu)$ is subadditive, and the second gives a multi-dimensional version of the well known convergence property of subadditive sequences on $\mathbb{R}$ ($x_{n+m} \leq x_n + x_m$).

LEMMA 10.2. *If Λ, Δ are two disjoint finite subsets of $\mathbb{Z}^d$,*

$$\mathcal{H}_{\Lambda\cup\Delta}(\mu) \leq \mathcal{H}_\Lambda(\mu) + \mathcal{H}_\Delta(\mu)\,. \tag{179}$$

PROOF. Start by writing

$$\mathcal{H}_\Lambda(\mu) = -\sum_{\sigma_\Lambda} \mu_\Lambda(\sigma_\Lambda)\log\mu_\Lambda(\sigma_\Lambda) = -\sum_{\sigma_{\Lambda\cup\Delta}} \mu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta})\log\mu_\Lambda(\sigma_\Lambda)\,,$$

which gives

$$\begin{aligned}
\mathcal{H}_\Lambda(\mu) - \mathcal{H}_{\Lambda\cup\Delta}(\mu) &= \sum_{\sigma_{\Lambda\cup\Delta}} \mu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta})\log\frac{\mu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta})}{\mu_\Lambda(\sigma_\Lambda)}\\
&= \sum_{\sigma_{\Lambda\cup\Delta}} \mu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta})\log\frac{\mu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta})}{\mu_\Lambda(\sigma_\Lambda)\lambda_\Delta(\sigma_\Delta)}\lambda_\Delta(\sigma_\Delta)\,, \qquad (180)
\end{aligned}$$

Let $\nu := \mu_\Lambda \otimes \lambda_{\Lambda^c}$. Observe that $\nu_{\Lambda\cup\Delta}(\sigma_{\Lambda\cup\Delta}) = \mu_\Lambda(\sigma_\Lambda)\lambda_\Delta(\sigma_\Delta)$. As a consequence, $\mu_{\Lambda\cup\Delta} \ll \nu_{\Lambda\cup\Delta}$, and (180) can be expressed as

$$\begin{aligned}
\mathcal{H}_\Lambda(\mu) - \mathcal{H}_{\Lambda\cup\Delta}(\mu) &= \mathcal{H}_{\Lambda\cup\Delta}(\mu|\nu) - |\Delta|\log 2\\
&\geq \mathcal{H}_\Delta(\mu|\nu) - |\Delta|\log 2\\
&= -\mathcal{H}_\Delta(\mu)\,.
\end{aligned}$$

We used Lemma 10.1 in the second line, and $\nu_\Delta = \lambda_\Delta$ in the third. □

LEMMA 10.3. *Let $\mathcal{R}$ be the collection of all rectangular boxes* [4] *of $\mathbb{Z}^d$. Let $\{a(\Lambda)\}_{\Lambda\in\mathcal{R}}$ be a collection of real numbers satisfying the following properties:*

(1) *$a(\Lambda) = a(\theta_x\Lambda)$ for all $x \in \mathbb{Z}^d$,*
(2) *if $\Lambda, \Delta \in \mathcal{R}$ are disjoint, with $\Lambda \cup \Delta \in \mathcal{R}$, then $a(\Lambda\cup\Delta) \leq a(\Lambda) + a(\Delta)$.*

Then

$$\lim_{n\to\infty}\frac{a(\Lambda_n)}{|\Lambda_n|} = \inf_{\Lambda\in\mathcal{R}}\frac{a(\Lambda)}{|\Lambda|}\,. \tag{181}$$

---

[3] A map $f : \mathcal{M}_{1,\theta}(\Omega) \to \mathbb{R}$ is **affine** if $f(s\mu + (1-s)\nu) = sf(\mu) + (1-s)f(\nu)$ for all $0 < s < 1$ and all $\mu, \nu \in \mathcal{M}_{1,\theta}(\Omega)$, and **upper semi-continuous** if $\mu_n \Rightarrow \mu$ implies $\limsup_n f(\mu_n) \leq f(\mu)$.

[4] A **rectangular box** is a set of the form $R = \{x \in \mathbb{Z}^d : a_i \leq x_i \leq b_i\}$ where each $a_i \leq b_i$. Isolated points are therefore also considered as rectangular boxes.

Proof. Let $\alpha := \inf_{\Lambda\in\mathcal{R}} a(\Lambda)/|\Lambda|$. Take $\epsilon > 0$ and let $\Lambda \in \mathcal{R}$ be such that $a(\Lambda)/|\Lambda| \leq \alpha+\epsilon$. For large n, decompose Λ_n into a maximal union of K_n disjoint translates of Λ: $\Lambda_n = \big(\bigcup_{i=1}^{K_n} \Lambda^{(i)}\big) \cup R_n$, where $R_n := \Lambda_n \backslash \bigcup_{i=1}^{K_n} \Lambda^{(i)}$ is such that $|R_n|/|\Lambda_n| \to 0$. Then

$$a(\Lambda_n) \leq \sum_{i=1}^{K_n} a(\Lambda^{(i)}) + \sum_{x\in R_n} a(\{x\}) = K_n a(\Lambda) + |R_n| a(\{0\}) \,.$$

Since $|\Lambda_n| = K_n|\Lambda| + |R_n|$,

$$\alpha \leq \liminf_{n\to\infty} \frac{a(\Lambda_n)}{|\Lambda_n|} \leq \limsup_{n\to\infty} \frac{a(\Lambda_n)}{|\Lambda_n|} \leq \frac{a(\Lambda)}{|\Lambda|} \leq \alpha+\epsilon \,.$$

The proof finishes by taking $\epsilon \searrow 0$. □

Proof of Theorem 10.1: If $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, then $a_\mu(\Lambda) := \mathcal{H}_\Lambda(\mu)$ satisfies $a_\mu(\theta_x\Lambda) = a_\mu(\Lambda)$ for all $x \in \mathbb{Z}^d$. Therefore, the existence of the limit (178) follows at once from (179) and Lemma 10.3. In particular, $h(\mu) \leq a_\mu(\Lambda)/|\Lambda|$ for all finite Λ. If $\mu_k \Rightarrow \mu$, then $\limsup_k h(\mu_k) \leq \limsup_k a_{\mu_k}(\Lambda)/|\Lambda| = a_\mu(\Lambda)/|\Lambda|$. By taking the infimum over Λ we get that h is upper semicontinuous: $\limsup_k h(\mu_k) \leq h(\mu)$. Since $\mathcal{M}_{1,\theta}(\Omega)$ is compact, the sets $\{h \geq c\}$ being closed imply they are also compact, which proves the last claim. We then show that h is affine. Since h is concave by (176) and by (2) of Lemma 10.1, it suffices to show that it is also convex. Let therefore $\mu = s\mu^1 + (1-s)\mu^2$, with $\mu^1, \mu^2 \in \mathcal{M}_{1,\theta}(\Omega)$, and $0 < s < 1$. Since log is non-decreasing, we have $\log \mu_\Lambda(\sigma_\Lambda) \geq \log[s\mu^1_\Lambda(\sigma_\Lambda)]$, and $\log \mu_\Lambda(\sigma_\Lambda) \geq \log[(1-s)\mu^2_\Lambda(\sigma_\Lambda)]$. Therefore,

$$\begin{aligned}\mathcal{H}_\Lambda(\mu) &= -s\sum_{\sigma_\Lambda} \mu^1_\Lambda(\sigma_\Lambda)\log\mu_\Lambda(\sigma_\Lambda) - (1-s)\sum_{\sigma_\Lambda} \mu^2_\Lambda(\sigma_\Lambda)\log\mu_\Lambda(\sigma_\Lambda)\\ &\leq -s\sum_{\sigma_\Lambda} \mu^1_\Lambda(\sigma_\Lambda)\log[s\mu^1_\Lambda(\sigma_\Lambda)] - (1-s)\sum_{\sigma_\Lambda} \mu^2_\Lambda(\sigma_\Lambda)\log[(1-s)\mu^2_\Lambda(\sigma_\Lambda)]\\ &= s\mathcal{H}_\Lambda(\mu^1) + (1-s)\mathcal{H}_\Lambda(\mu^2) - s\log s - (1-s)\log(1-s)\,.\end{aligned}$$

Dividing by $|\Lambda|$ and taking $\Lambda \nearrow \mathbb{Z}^d$, we get that h is convex. □

Theorem 10.1 and (176) imply that for all $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, the specific relative entropy with respect to λ,

$$h(\mu|\lambda) := \lim_{n\to\infty} \frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\lambda)\,, \tag{182}$$

exists and equals $\log 2 - h(\mu)$. Our next step now is to show that the limit exists also when λ is replaced by a Gibbs measure specified by a Gibbsian specification.

3. Gibbs measures as reference measures

From now on, $\Phi = (\Phi_B)$ denotes a translation invariant potential. This means that for all finite set $B \subset \mathbb{Z}^d$, $\Phi_{\theta_x B} = \theta_x \Phi_B$ for all $x \in \mathbb{Z}^d$. We will also assume that Φ is uniformly absolutely summable. Therefore,

$$\infty > \sup_{x\in\mathbb{Z}^d} \sum_{B\ni x} \|\Phi_B\| = \sum_{B\ni 0} \|\Phi_B\| =: \|\Phi\|_0 \tag{183}$$

The convergence of the sum above is understood in the following sense:

$$\sum_{B\ni 0}\|\Phi_B\| := \lim_{n\to\infty}\sum_{\substack{B\ni 0\\ B\subset\Lambda_n}}\|\Phi_B\|\,. \tag{184}$$

Let γ^Φ be the Gibbsian specification associated to Φ. That is,

$$\gamma^\Phi_\Lambda(\sigma_\Lambda|\omega) = \frac{e^{-H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c})}}{Z^\Phi_\Lambda(\omega_{\Lambda^c})}\,,\quad \text{where } Z^\Phi_\Lambda(\omega_{\Lambda^c}) = \sum_{\sigma_\Lambda\in\Omega_\Lambda} e^{-H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c})}\,.$$

We denote by $\mathcal{G}(\Phi) := \mathcal{G}(\gamma^\Phi)$ the set of Gibbs measures specified by γ^Φ, and $\mathcal{G}_\theta(\Phi) := \mathcal{G}(\Phi)\cap\mathcal{M}_{1,\theta}(\Omega)$. As an exercise, the reader can verify that $\mathcal{G}_\theta(\Phi)\neq\emptyset$. The starting point for the study of the entropy of Gibbs measures is the following:

Proposition 10.1. *Let $\mu\in\mathcal{M}_{1,\theta}(\Omega)$ and (ν_n), (ν'_n) be two arbitrary sequences in $\mathcal{M}_1(\Omega)$. If either of the limits*

$$\lim_{n\to\infty}\frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu_n\gamma^\Phi_{\Lambda_n})\,,\quad \lim_{n\to\infty}\frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu'_n\gamma^\Phi_{\Lambda_n}) \tag{185}$$

exists, then the other one also exists, and they are equal.

Proof. Let Λ be finite. Observe that by definition of γ^Φ_Λ, there exists a constant $c^\Phi_\Lambda>0$ such that $\gamma^\Phi_\Lambda(\sigma_\Lambda|\omega)\geq c^\Phi_\Lambda$ for all σ_Λ,ω. Therefore, we always have $\mu_{\Lambda_n}\ll\nu_n\gamma^\Phi_{\Lambda_n}$, $\mu_{\Lambda_n}\ll\nu'_n\gamma^\Phi_{\Lambda_n}$. Now,

$$\mathcal{H}_\Lambda(\mu|\nu_n\gamma^\Phi_\Lambda) - \mathcal{H}_\Lambda(\mu|\nu_n\gamma^\Phi_\Lambda) = \sum_{\sigma_\Lambda}\mu_\Lambda(\sigma_\Lambda)\log\frac{\nu'_n\gamma^\Phi_\Lambda(\sigma_\Lambda)}{\nu_n\gamma^\Phi_\Lambda(\sigma_\Lambda)}\,.$$

Define

$$r^\Phi_\Lambda := \sup_{\sigma_\Lambda,\omega,\eta}|H^\Phi_\Lambda(\sigma_\Lambda\omega_{\Lambda^c}) - H^\Phi_\Lambda(\sigma_\Lambda\eta_{\Lambda^c})|\,.$$

Then clearly, uniformly in σ_Λ,

$$e^{-2r^\Phi_\Lambda}\leq\frac{\nu'_n\gamma^\Phi_\Lambda(\sigma_\Lambda)}{\nu_n\gamma^\Phi_\Lambda(\sigma_\Lambda)}\leq e^{2r^\Phi_\Lambda}\,.$$

Therefore, the two limits in (185) coincide if one can show that $r^\Phi_{\Lambda_n} = o(|\Lambda_n|)$, i.e. that $\lim_n r^\Phi_{\Lambda_n}/|\Lambda_n| = 0$. This follows by the uniform absolute summability of Φ, and can be seen as follows:

$$r^\Phi_\Lambda \leq \sum_{\substack{B\cap\Lambda\neq\emptyset\\ B\not\subset\Lambda}}\sup_{\sigma_\Lambda,\omega,\eta}|\Phi_B(\sigma_\Lambda\omega_{\Lambda^c}) - \Phi_B(\sigma_\Lambda\eta_{\Lambda^c})| \leq 2\sum_{\substack{B\cap\Lambda\neq\emptyset\\ B\not\subset\Lambda}}\|\Phi_B\| \leq 2\sum_{x\in\Lambda}\sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\|\Phi_B\|$$

For each x, let $\Lambda_k(x) := \Lambda_k + x$. For a fixed k, we decompose

$$\sum_{x\in\Lambda}\sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\|\Phi_B\| = \sum_{\substack{x\in\Lambda:\\ \Lambda_k(x)\subset\Lambda}}\sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\|\Phi_B\| + \sum_{\substack{x\in\Lambda:\\ \Lambda_k(x)\not\subset\Lambda}}\sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\|\Phi_B\|\,. \tag{186}$$

The first sum on the right-hand side is bounded by

$$|\Lambda|\sup_{x\in\mathbb{Z}^d}\sum_{\substack{B\ni x\\ B\not\subset\Lambda_k(x)}}\|\Phi_B\| \equiv |\Lambda|\sum_{\substack{B\ni 0\\ B\not\subset\Lambda_k}}\|\Phi_B\| \equiv |\Lambda|\alpha_k\,,$$

where $\alpha_k \to 0$ when $k \to \infty$ (see (184)). Then, observe that the only points $x \in \Lambda$ that contribute to the second sum are those for which $\Lambda_k(x) \cap \partial_k^- \Lambda \neq \emptyset$ ($\partial_k^- \Lambda := \{x \in \Lambda : d(x, \Lambda^c) \leq k\}$). Therefore,

$$\sum_{\substack{x\in\Lambda:\\ \Lambda_k(x)\not\subset\Lambda}} \sum_{\substack{B\ni x\\ B\not\subset\Lambda}} \|\Phi_B\| \leq |\partial_k^- \Lambda| \|\Phi\|_0 .$$

Altogether, $\limsup_n \frac{r^\Phi_{\Lambda_n}}{|\Lambda_n|} \leq \alpha_k$. The proof finishes by taking $k \to \infty$. □

We will now use Proposition 10.1 in two different ways, making different choices for μ and for the sequences (ν_n), (ν'_n) that appear in (185).

Let first μ in (185) be in $\mathcal{G}_\theta(\Phi)$, and take $\nu_n := \mu$, $\nu'_n := \delta_\omega$, where ω is any fixed configuration. Then $\nu_n \gamma^\Phi_{\Lambda_n} = \mu \gamma^\Phi_{\Lambda_n} = \mu$ and so $\mathcal{H}_{\Lambda_n}(\mu | \nu_n \gamma^\Phi_{\Lambda_n}) = 0$, and the limit on the left-hand side of (185) is zero. Therefore the limit on the right-hand side is also zero. But, as a distribution on Ω_{Λ_n},

$$\nu'_n \gamma^\Phi_{\Lambda_n}(\cdot) = \int \gamma^\Phi_{\Lambda_n}(\,\cdot\,|\sigma)\delta_\omega(d\sigma) = \gamma^\Phi_{\Lambda_n}(\,\cdot\,|\omega) = \frac{e^{-H^\Phi_{\Lambda_n}(\,\cdot\,\omega^c_{\Lambda_n})}}{Z^\Phi_{\Lambda_n}(\omega_{\Lambda_n^c})} ,$$

and therefore (remember (173))

$$\mathcal{H}_{\Lambda_n}(\mu | \nu'_n \gamma^\Phi_{\Lambda_n}) = \log Z^\Phi_{\Lambda_n}(\omega_{\Lambda_n^c}) - \big(\mathcal{H}_{\Lambda_n}(\mu) - \mu(H^\Phi_{\Lambda_n}(\,\cdot\,\omega_{\Lambda^c}))\big) . \tag{187}$$

Since $\mathcal{H}_{\Lambda_n}(\mu)/|\Lambda_n|$ has a limit when $n \to \infty$, we see that if either of the limits

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \mu(H^\Phi_{\Lambda_n}(\,\cdot\,\omega_{\Lambda_n^c})) , \quad \lim_{n\to\infty} \frac{1}{|\Lambda_n|} \log Z^\Phi_{\Lambda_n}(\omega_{\Lambda_n^c})$$

exists, then the other one also exists. We show existence of the first. Define

$$\varphi_\Phi := \sum_{B\ni 0} \frac{1}{|B|} \Phi_B , \tag{188}$$

which measures the energy of a configuration in the neighbourhood of the origin (the $1/|B|$ is to avoid over-counting). Observe that $\varphi_\Phi \in C(\Omega)$, and that $\|\varphi_\Phi\| \leq \|\Phi\|_0$. If $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, then $\langle \varphi_\Phi, \mu \rangle := \mu(\varphi_\Phi)$ is called the (specific) energy of μ.

LEMMA 10.4. *Let $\mu \in \mathcal{M}_{1,\theta}(\Omega)$. Then,*

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \mu(H^\Phi_{\Lambda_n}(\,\cdot\,\omega_{\Lambda_n^c})) = \langle \varphi_\Phi, \mu \rangle .$$

In particular, the limit doesn't depend on ω.

PROOF. Since $\theta_x \mu = \mu$ for all x,

$$\begin{aligned}\frac{1}{|\Lambda|}\mu(H^\Phi_\Lambda(\,\cdot\,\omega_{\Lambda^c})) &= \frac{1}{|\Lambda|}\sum_{x\in\Lambda} \mu(\theta_x\varphi_\Phi) + \frac{1}{|\Lambda|}\mu\Big(H^\Phi_\Lambda(\,\cdot\,\omega_{\Lambda^c}) - \sum_{x\in\Lambda}\theta_x\varphi_\Phi\Big)\\ &= \mu(\varphi_\Phi) + \frac{1}{|\Lambda|}\mu\Big(H^\Phi_\Lambda(\,\cdot\,\omega_{\Lambda^c}) - \sum_{x\in\Lambda}\theta_x\varphi_\Phi\Big)\end{aligned}$$

To estimate the difference, write

$$\begin{aligned}H^{\Phi}_{\Lambda}(\,\cdot\,\omega_{\Lambda^c}) &= \sum_{B\subset\Lambda}\Phi_B(\,\cdot\,\omega_{\Lambda^c}) + \sum_{\substack{B\cap\Lambda\neq\emptyset\\ B\not\subset\Lambda}}\Phi_B(\,\cdot\,\omega_{\Lambda^c})\\ &= \sum_{x\in\Lambda}\sum_{\substack{B\ni x\\ B\subset\Lambda}}\frac{1}{|B|}\Phi_B(\cdot) + \sum_{\substack{B\cap\Lambda\neq\emptyset\\ B\not\subset\Lambda}}\Phi_B(\,\cdot\,\omega_{\Lambda^c})\\ &= \sum_{x\in\Lambda}\Big\{\theta_x\varphi_\Phi(\cdot) - \sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\frac{1}{|B|}\Phi_B(\cdot)\Big\} + \sum_{\substack{B\cap\Lambda\neq\emptyset\\ B\not\subset\Lambda}}\Phi_B(\,\cdot\,\omega_{\Lambda^c})\end{aligned}$$

Therefore,

$$\Big|H^{\Phi}_{\Lambda}(\,\cdot\,\omega_{\Lambda^c}) - \sum_{x\in\Lambda}\theta_x\varphi_\Phi(\cdot)\Big| \leq 2\sum_{x\in\Lambda}\sum_{\substack{B\ni x\\ B\not\subset\Lambda}}\|\Phi_B\| = o(|\Lambda|)\,, \tag{189}$$

as was shown in (186). □

Remark 10.1. Observe that if Φ has finite range, i.e. if there exists $0 < R < \infty$ such that $\Phi_B \equiv 0$ if $\mathrm{diam}(B) > R$, then the difference in (189) is bounded by a term of order $O(|\partial^-_R\Lambda|)$.

Together with (187) and Theorem 10.1, Lemma 10.4 implies that

$$\psi(\Phi) := \lim_{n\to\infty}\frac{1}{|\Lambda_n|}\log Z^{\Phi}_{\Lambda_n}(\omega_{\Lambda_n^c}) \tag{190}$$

exists, and does not depend on $\omega \in \Omega$. Moreover, if $\mu \in \mathcal{G}_\theta(\Phi)$, then

$$h(\mu) - \langle\varphi_\Phi, \mu\rangle = \psi(\Phi)\,. \tag{191}$$

On the other hand, let now the measure μ in (185) be any $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, and let $\nu \in \mathcal{G}_\theta(\Phi)$. Take $\nu_n := \nu$, so that $\nu_n\gamma^{\Phi}_{\Lambda_n} = \nu\gamma^{\Phi}_{\Lambda_n} = \nu$, and $\nu'_n := \delta_\omega$, where ω is any fixed configuration, as before. Then Proposition 10.1 combined with the above comments imply that

$$h(\mu|\nu) := \lim_{n\to\infty}\frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu) \tag{192}$$

exists and equals

$$h(\mu|\nu) = \psi(\Phi) - \{h(\mu) - \langle\varphi_\Phi, \mu\rangle\}\,. \tag{193}$$

This last display is the infinite-volume equivalent of (173).

Notice that the right hand side of (193) does not depend on ν but only on the potential Φ which specifies ν. It is thus more natural to define $h(\mu|\Phi) := h(\mu|\nu)$. By (191), $h(\mu|\Phi) = 0$ if $\mu \in \mathcal{G}_\theta(\Phi)$. We have therefore proved the first part of the following theorem, called the Variational Principle:

Theorem 10.2. *Let Φ be a translation invariant, uniformly absolutely convergent potential. Then for all (compare with* (173)*)*

$$h(\mu|\Phi) := \psi(\Phi) - \{h(\mu) - \langle\varphi_\Phi, \mu\rangle\} \tag{194}$$

defines an affine, lower-semicontinuous mapping $h(\cdot|\Phi) : \mathcal{M}_{1,\theta}(\Omega) \to [0,\infty]$ *with compact level sets. Moreover,* $h(\mu|\Phi) = 0$ *if and only if* $\mu \in \mathcal{G}_\theta(\Phi)$.

PROOF. Since $\langle \varphi_\Phi, \cdot\rangle$ and $h(\cdot)$ are affine, $h(\cdot|\Phi)$ also is. The lower semicontinuity follows by the upper semicontinuity of h and from the obvious fact that if $\mu_n \Rightarrow \mu$, then $\lim_n \langle \varphi_\Phi, \mu_n\rangle = \langle \varphi_\Phi, \mu\rangle$, by the continuity of φ_Φ. We have already seen that $h(\cdot|\Phi) = 0$ on $\mathcal{G}_\theta(\Phi)$. The main point is thus to prove the *second half* of the variational principle, namely that if $\mu \in \mathcal{M}_{1,\theta}(\Omega)$ is such that $h(\mu|\Phi) = 0$, then μ is a Gibbs measure for Φ. This will be a consequence of the following proposition, which is slightly more general than what we need here. □

PROPOSITION 10.2. *Let* Φ *be translation invariant and uniformly absolutely convergent. Let* $\nu \in \mathcal{G}_\theta(\Phi)$*. If* $\mu \in \mathcal{M}_{1,\theta}(\Omega)$ *is such that*

$$\liminf_{n\to\infty} \frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu) = 0\,, \tag{195}$$

then $\mu \in \mathcal{G}_\theta(\Phi)$.

PROOF. We *fix* a finite $\Lambda \subset \mathbb{Z}^d$ and prove that $\mu\gamma_\Lambda^\Phi(g) = \mu(g)$ for all local function $g \in \mathcal{L}_{\mathrm{oc}}(\Omega)$. The proof contains three main steps.

Step 1: For all $\delta > 0$ *and all cube* $R \supset \Lambda$*, there exists* $\Delta \supset R$ *such that* $\mathcal{H}_\Delta(\mu|\nu) - \mathcal{H}_{\Delta\setminus\Lambda}(\mu|\nu) \le \delta$. By (195), n can be taken large enough so that $|\Lambda_n| \ge |R|$, and

$$\frac{1}{|\Lambda_n|}\mathcal{H}_{\Lambda_n}(\mu|\nu) \le \frac{\delta}{2^d|R|}\,.$$

Let then $m \ge 1$ be an integer satisfying $m^d|R| \le |\Lambda_n| < (2m)^d|R|$. Roughly, m^d is a lower bound on the number of disjoint translates of R that can be put into Λ_n. Choose then m^d points $x(1), \dots, x(m^d)$ in Λ_n such that the translates $R(i) := \theta_{x(i)}R$ are disjoint subsets of Λ_n. Let $\Delta(i) := R(1) \cup \cdots \cup R(i)$, and $\Lambda(i) := \theta_{x(i)}\Lambda$. Observe that $\Delta(i)\setminus R(i) \subset \Delta(i)\setminus\Lambda(i)$ and that $\Delta(m^d) \subset \Lambda_n$. Therefore, defining $\mathcal{H}_\emptyset(\mu|\nu) := 0$ and using twice (3) of Lemma 10.1,

$$\begin{aligned}
\frac{1}{m^d}\sum_{i=1}^{m^d}\big\{\mathcal{H}_{\Delta(i)}(\mu|\nu) - \mathcal{H}_{\Delta(i)\setminus\Lambda(i)}(\mu|\nu)\big\} &\le \frac{1}{m^d}\sum_{i=1}^{m^d}\big\{\mathcal{H}_{\Delta(i)}(\mu|\nu) - \mathcal{H}_{\Delta(i)\setminus R(i)}(\mu|\nu)\big\}\\
&= \frac{1}{m^d}\mathcal{H}_{\Delta(m^d)}(\mu|\nu)\\
&\le \frac{1}{m^d}\mathcal{H}_{\Lambda_n}(\mu|\nu) \le \delta\,.
\end{aligned}$$

Therefore, there exists at least one $i \in \{1, 2, \dots, m^d\}$ such that $\mathcal{H}_{\Delta(i)}(\mu|\nu) - \mathcal{H}_{\Delta(i)\setminus\Lambda(i)}(\mu|\nu) \le \delta$. Take $\Delta := \theta_{-x(i)}\Delta(i)$. Then, by the translation invariance of μ and ν, $\mathcal{H}_\Delta(\mu|\nu) - \mathcal{H}_{\Delta\setminus\Lambda}(\mu|\nu) = \mathcal{H}_{\Delta(i)}(\mu|\nu) - \mathcal{H}_{\Delta(i)\setminus\Lambda(i)}(\mu|\nu) \le \delta$.

As already seen before, the fact that ν is specified by a Gibbsian specification implies that for all finite region $\Delta \subset \mathbb{Z}^d$, $\nu_\Delta(\sigma_\Delta) > 0$ for all $\sigma_\Delta \in \Omega_\Delta$ and therefore

$\mu_\Delta \ll \nu_\Delta$. We can thus define the Radon-Nikodým densities

$$f_\Delta = f_\Delta(\sigma_\Delta) := \frac{\mu_\Delta(\sigma_\Delta)}{\nu_\Delta(\sigma_\Delta)}\,.$$

Step 2: For all $\epsilon > 0$, there exists $\delta > 0$ such that if $\mathcal{H}_\Delta(\mu|\nu) - \mathcal{H}_{\Delta\setminus\Lambda}(\mu|\nu) \leq \delta$, then $\nu(|f_\Delta - f_{\Delta\setminus\Lambda}|) \leq \epsilon$. Using the densities,

$$\begin{aligned}\mathcal{H}_\Delta(\mu|\nu) - \mathcal{H}_{\Delta\setminus\Lambda}(\mu|\nu) &= \mu\Big(\log\frac{f_\Delta}{f_{\Delta\setminus\Lambda}}\Big)\\ &= \nu\Big(f_\Delta \log\frac{f_\Delta}{f_{\Delta\setminus\Lambda}}\Big) = \nu\Big(f_{\Delta\setminus\Lambda}\phi\Big(\frac{f_\Delta}{f_{\Delta\setminus\Lambda}}\Big)\Big)\,, \qquad (196)\end{aligned}$$

where $\phi(x) := 1 - x + x\log x$. As can be easily verified, there exists $r > 0$ (depending on ϵ) such that $\phi(x) \geq r(|x-1| - \epsilon/2)$ for all $x \geq 0$. Therefore, if $\mathcal{H}_\Delta(\mu|\nu) - \mathcal{H}_{\Delta\setminus\Lambda}(\mu|\nu) \leq \delta$, (196) gives $\delta \geq r\big(\nu(|f_\Delta - f_{\Delta\setminus\Lambda}|) - \epsilon/2\big)$. The proof of the claim follows by taking $\delta := \epsilon/2r$.

Step 3: For all $g \in \mathcal{L}_{\text{oc}}(\Omega)$, $\mu\gamma^\Phi_\Lambda(g) = \mu(g)$. Fix some $\epsilon > 0$. Let $\delta > 0$ be as in Step 2. Since γ^Φ is quasilocal, $\gamma^\Phi_\Lambda g \in C(\Omega)$. Let $g_* \in \mathcal{L}_{\text{oc}}(\Omega)$ be such that $\|\gamma^\Phi_\Lambda g - g_*\| \leq \epsilon$. Let $R \supset \Lambda$ be a cube, large enough so that g is $\mathcal{F}_R$-measurable and g_* is $\mathcal{F}_{R\setminus\Lambda}$-measurable. Let Δ be chosen as in Step 1, in function of δ and R. By Step 2, $\nu(|f_\Delta - f_{\Delta\setminus\Lambda}|) \leq \epsilon$. Then write

$$\begin{aligned}|\mu\gamma^\Phi_\Lambda(g) - \mu(g)| \leq & \mu(|\gamma^\Phi_\Lambda g - g_*|) + |\mu(g_*) - \nu(f_{\Delta\setminus\Lambda}g_*)| + \nu(f_{\Delta\setminus\Lambda}|g_* - \gamma^\Phi_\Lambda g|)\\ &+ |\nu(f_{\Delta\setminus\Lambda}(\gamma^\Phi_\Lambda g - g))| + \|g\|\nu(|f_\Delta - f_{\Delta\setminus\Lambda}|) + |\nu(f_\Delta g) - \mu(g)|\,.\end{aligned}$$

The first and third terms are $\leq \epsilon$. The second and sixth terms are zero. The fifth is $\leq \|g\|\epsilon$. The fourth is zero. Namely, observe that $f_{\Delta\setminus\Lambda}\gamma^\Phi_\Lambda g = \gamma^\Phi_\Lambda(f_{\Delta\setminus\Lambda}g)$ due to the $\mathcal{F}_{\Lambda^c}$-measurability of $f_{\Delta\setminus\Lambda}$. But since by hypothesis ν is specified by γ^Φ, $\nu(f_{\Delta\setminus\Lambda}\gamma^\Phi_\Lambda g) = \nu(\gamma^\Phi_\Lambda[f_{\Delta\setminus\Lambda}g]) = \nu\gamma^\Phi_\Lambda(f_{\Delta\setminus\Lambda}g) = \nu(f_{\Delta\setminus\Lambda}g)$. Altogether,

$$|\mu\gamma^\Phi_\Lambda(g) - \mu(g)| \leq 3\epsilon + \epsilon\|g\|\,.$$

Since ϵ was arbitrary, this shows that $\mu\gamma^\Phi_\Lambda(g) = \mu(g)$. □

We have thus adapted the finite-volume considerations of the introduction to the Gibbsian description of translation invariant infinite systems, making a closer link with thermodynamics. For a given potential Φ, the functional we were looking for is thus

$$\mathcal{W}(\mu) := h(\mu) - \langle \varphi_\Phi, \mu\rangle\,,$$

and if μ^{Gibbs} is any Gibbs measure specified by γ^Φ, then

$$h(\mu|\Phi) = \mathcal{W}(\mu^{\text{Gibbs}}) - \mathcal{W}(\mu)\,.$$

This description of Gibbs measures is completely equivalent to the description in terms of specifications. It is interesting to notice that invariant Gibbs measures are sometimes (in ergodic theory, for example) *defined* using the variational characterization.

It is interesting to note that if $\mathcal{M}_{1,\theta}(\Omega)$ is considered together with the metric $\rho(\cdot,\cdot)$ of Lemma 9.5, then by Theorem 10.2, $h(\cdot|\Phi) : \mathcal{M}_{1,\theta}(\Omega) \to [0,\infty]$ satisfies all the properties of a good rate function, in the sense of Definition 5.2. In particular, by Lemma 5.4, there exists at least one $\mu_* \in \mathcal{M}_{1,\theta}(\Omega)$ such that $h(\mu_*|\Phi) = 0$. This gives an alternate proof of the existence of invariant Gibbs measures associated to Φ: $\mathcal{M}_{1,\theta}(\Omega) \neq \emptyset$.

The fact that $h(\cdot|\Phi)$ has the properties of a good rate function is of course not an accident. In the next chapter, we present a Large Deviation analysis of Gibbs measures, where the Variational Principle will appear naturally, as a consequence of a Large Deviation Principle.

CHAPTER 11

Gibbs Measures and Large Deviations

1. Introduction

In Chapter 3, we considered sequences of i.i.d. random variables X_k with distribution ν, and considered two random quantities associated to a finite sample of size n: the empirical mean $\frac{1}{n}\sum_{k=1}^n X_k = \frac{S_n}{n}$, and the empirical measure

$$L_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k} \,. \tag{197}$$

The Theorem of Sanov gave a Large Deviation Principle for the exponential concentration of L_n around the only zero of the rate function $D(\cdot\|\nu)$, given by ν. In this sense, the distribution ν could be read from the empirical measure. The contraction principle then allowed to derive a LDP for the empirical mean.

In this chapter, we consider a similar approach for the study of a Gibbs random field on $\mathbb{Z}^d$. Assume the collection of random variables $\sigma = (\sigma_x)_{x\in\mathbb{Z}^d}$ is described by a random field μ. The natural higher dimensional analogue of (197) is the sequence of empirical fields $L_n \in \mathcal{M}_1(\Omega)$, where [1]

$$L_n = L_n(\sigma) := \frac{1}{|\Lambda_n|}\sum_{x\in\Lambda_n} \delta_{\theta_x\sigma} \,. \tag{198}$$

The aim of this chapter is to study the concentration of L_n under a Gibbs measure μ associated to a potential Φ. Similarly to the setting of the Theorem of Sanov, a large deviation analysis of the distribution of L_n will lead to a rate function given exactly by the relative entropy $h(\cdot|\Phi)$ of Chapter 10. The interest lies in the fact that from the Variational Principle (Theorem 10.2), the set on which $h(\cdot|\Phi)$ attains its minimum zero on the set of translation invariant Gibbs measures, $\mathcal{G}_\theta(\Phi)$. And as we know from the study of the Ising model, $\mathcal{G}_\theta(\Phi)$ can contain more than one element.

[1] The reason for defining the empirical measure in this way is the following. Assume the distribution μ describing the system is unknown, and that one wants to guess the probability that the the spin at the origin is $+1$, say, by looking only at a given configuration ω that was drawn randomly. A natural way to do so is to estimate this probability by averaging the spins in a neighbourhood of the origin:

$$L_\Lambda(\{\sigma_0=+1\}) := \frac{1}{|\Lambda|}\sum_{x\in\Lambda} 1_{\{\omega_x=+1\}} \equiv \frac{1}{|\Lambda|}\sum_{x\in\Lambda}\delta_{\theta_x\omega}(\{\omega_0=+1\})$$

We will use two central results of Chapter 5, on the abstract theory of Large Deviations: the Varadhan Lemma and the Theorem on tilted measures. Note that since we are working on $\mathbb{Z}^d$, considering the thermodynamic limit along a sequence of boxes Λ_n, the large deviation principles aluded to are always of speed $a_n := |\Lambda_n|$ (see Remark 5.1).

The material presented presented here is taken from Section 3.4 of [?], which itself is largely inspired by Pfister's lecture notes [?].

2. The Free Gibbs Measure

Let Φ be a translation invariant uniformly absolutely convergent potential. To settle the large deviation study of the distribution of L_n under a Gibbs measure specified by Φ, we will make a simplification which consists essentially in removing boundary terms which don't contribute on the exponential scale, at which large deviations occur [2]. The resulting measure defined in (201) will be called the free Gibbs measure. A more general treatment of the full Gibbs measure, with boundary terms, can be found in [?].

First, we make explicit the appearance of the reference product measure λ in the definition of the Gibbs measure, as was done with the distribution of the magnetization for the Curie-Weiss model (see page 50). The sums over σ_Λ that appear in the kernel γ_Λ^Φ are changed into integrals with respect to the product reference measure λ. We start with the partition function:

$$\sum_{\sigma_\Lambda \in \Omega_\Lambda} e^{-H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})} = 2^{|\Lambda|} \int_{\Omega_\Lambda} e^{-H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})} \lambda_\Lambda(d\sigma_\Lambda) .$$

We can thus redefine $Z_\Lambda^\Phi(\omega_{\Lambda^c})$ without the factor $2^{|\Lambda|}$,

$$Z_\Lambda^\Phi(\omega_{\Lambda^c}) := \int_{\Omega_\Lambda} e^{-H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})} \lambda_\Lambda(d\sigma_\Lambda) \equiv \int_{\Omega} e^{-H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})} \lambda(d\sigma) ,$$

and do the same with the sum appearing in the numerator of the kernel, to obtain

$$\gamma_\Lambda^\Phi(\sigma_\Lambda|\omega) = \frac{e^{-H_\Lambda^\Phi(\sigma_\Lambda \omega_{\Lambda^c})}}{Z_\Lambda^\Phi(\omega_{\Lambda^c})} \lambda_\Lambda(\sigma_\Lambda) .$$

We then turn to the boundary terms. Since we will consider only the sequence of cubes Λ_n, We know from (189) in Chapter 10 that uniformly in σ and ω,

$$H_{\Lambda_n}^\Phi(\sigma_{\Lambda_n} \omega_{\Lambda_n^c}) - \sum_{x \in \Lambda_n} \theta_x \varphi_\Phi(\sigma) = o(|\Lambda_n|) . \tag{199}$$

[2] It is important to notice here that the construction of a Gibbs measure can indeed be very sensitive to boundary terms, as we have seen with the Ising model at low temperature. Nevertheless, our interest here is in *large deviations* properties of the measure, in which boundary terms play no role.

We can thus redefine the partition function without the boundary term $o(|\Lambda_n|)$, since it doesn't contribute to the volume in the thermodynamic limit:

$$Z_n^\Phi := \int_\Omega e^{-\sum_{x\in\Lambda_n}\theta_x\varphi_\Phi(\sigma)}\lambda(d\sigma) \tag{200}$$

Then, the free Gibbs measure μ_n^Φ is defined naturally by

$$\frac{d\mu_n^\Phi}{d\lambda} := \frac{e^{-\sum_{x\in\Lambda_n}\theta_x\varphi_\Phi(\sigma)}}{Z_n^\Phi} \equiv \frac{e^{-|\Lambda_n|\langle\varphi_\Phi, L_n\rangle}}{Z_n^\Phi}\,. \tag{201}$$

To define the distribution of L_n under μ_n^Φ, we endow $\mathcal{M}_1(\Omega)$ with the Borel σ-algebra generated by the open sets of the topology of local convergence. For all measurable $B \subset \mathcal{M}_1(\Omega)$,

$$Q_n^\Phi(B) := \mu_n^\Phi(L_n \in B)\,. \tag{202}$$

We also define the distribution of L_n under the reference product measure:

$$Q_n^0(B) = \lambda(L_n \in B)\,. \tag{203}$$

Transporting the measure μ_n^Φ onto $\mathcal{M}_1(\Omega)$ thus gives

$$Q_n^\Phi(B) = \int_B \frac{e^{-|\Lambda_n|\langle\varphi_\Phi,\mu\rangle}}{Z_n^\Phi} Q_n^0(d\mu)\,. \tag{204}$$

Q_n^Φ is thus a tilting of Q_n^0, in the sense of (90) in Chapter 5. Moreover, $\mu \mapsto -\langle\varphi_\Phi,\mu\rangle$ is bounded and continuous. We know that if a LDP holds for the sequence (Q_n^0), Theorem 5.2 allows to derive a LDP for (Q_n^Φ).

Remember that the same procedure was used for the Curie-Weiss model, in Chapter 6. There, Q_n^0 denoted the distribution of the empirical mean $\frac{S_n}{n}$ of a sequence of i.i.d. uniform $\{\pm 1\}$-valued random variables, and a LDP for (Q_n^0) followed from the Cramér Theorem for i.i.d. sequences. Here, a LDP for the distribution of L_n under the product measure λ will be more involved, due to the geometry of $\mathbb{Z}^d$ [3].

3. The LDP for L_n under the product measure

THEOREM 11.1. *Let Q_n^0 denote the distribution of L_n under the uniform product measure λ (see (203)). Then (Q_n^0) satisfies a LDP on $\mathcal{M}_1(\Omega)$ with speed $|\Lambda_n|$ and good affine rate function I_λ given by*

$$I_\lambda(\mu) = \begin{cases} h(\mu|\lambda) & \text{if } \mu \in \mathcal{M}_{1,\theta}(\Omega)\,,\\ +\infty & \text{if } \mu \in \mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)\,. \end{cases} \tag{205}$$

[3] A LDP for the empirical field of an i.i.d. sequence can be found in [**?**]. Observe that the technique used therein is based on the Eulerian paths we considered in the proof of Theorem 3.2, and does not generalize easily to higher dimensions.

Remember from Chapter 10 that $h(\mu|\lambda)$ equals $\log 2 - h(\mu)$, and is lower semi-continuous with compact level sets (Theorem 10.1). The strategy of the proof is the following. First, we will show in Section 3.1 that a LDP holds for each marginal $\pi_{\Lambda_k} L_n$, with a rate function $I_\lambda^{(k)}$. We then extend this to a LDP for L_n with the rate function $I_\lambda = \sup_{k\geq 1} I_\lambda^{(k)}$, without knowing its relation with the relative entropy given in (205). To identify $I_\lambda(\cdot)$ as $h(\cdot|\lambda)$, we will introduce in Section 3.2 a generalization of the free energy $\psi(\Phi)$: for all $f \in C(\Omega)$,

$$\psi(f|\lambda) := \lim_{n\to\infty} \frac{1}{|\Lambda_n|} \log \int \exp\Big(\sum_{x\in\Lambda_n} \theta_x f(\sigma)\Big)\lambda(d\sigma)\,. \tag{206}$$

$f \mapsto \psi(f|\lambda)$ is convex and lower semi-continuous. The relations between $I_\lambda(\cdot)$, $h(\cdot|\lambda)$ and $\psi(\cdot|\lambda)$ are as follows. On one hand, using the Varadhan Lemma in (206) will show that $\psi(\cdot|\lambda)$ is the Legendre transform of $I_\lambda(\mu)$ with respect to μ, which we denote by $I_\lambda^* = \psi$:

$$\psi(f|\lambda) = \sup_{\mu\in\mathcal{M}_{1,\theta}(\Omega)} \{\langle f,\mu\rangle - I_\lambda(\mu)\}\,. \tag{207}$$

On the other, we will see that $h(\cdot|\lambda)$ is the Legendre transform of $\psi(f|\lambda)$ with respect to f, $\psi^* = h(\cdot|\lambda)$: for all $\mu \in \mathcal{M}_{1,\theta}(\Omega)$,

$$\sup_{f\in C(\Omega)} \{\langle f,\mu\rangle - \psi(f|\lambda)\} = h(\mu|\lambda)\,. \tag{208}$$

Therefore, $h(\cdot|\lambda)$ equals the double Legendre transform of $I_\lambda(\cdot)$: $(I_\lambda^*)^* = h(\cdot|\lambda)$. Since I_λ is convex and lower semi-continuous, an classical duality theorem from convex analysis implies that $(I_\lambda^*)^* = I_\lambda$, i.e. that $I_\lambda(\cdot) = h(\cdot|\lambda)$.

3.1. The LDP for the marginals of L_n. Let $\pi_{\Lambda_k} L_n$ denote the marginal of L_n on $\mathcal{M}_1(\Omega_{\Lambda_k})$. We will denote a generic element of $\mathcal{M}_1(\Omega_{\Lambda_k})$ by α. The open ball with radius $\delta > 0$ centered at α is denoted $B_\delta(\alpha) := \{\alpha' : \|\alpha'-\alpha\|_1 < \delta\}$.

PROPOSITION 11.1. *Let $Q_n^{(k)}$ denote the distribution of $\pi_{\Lambda_k} L_n$ under λ. Then the sequence $(Q_n^{(k)})_{n\geq 1}$ satisfies a LDP with speed $|\Lambda_n|$ and with the good convex rate function $I_\lambda^{(k)} : \mathcal{M}_1(\Omega_{\Lambda_k}) \to [0,+\infty]$, given by*

$$I_\lambda^{(k)}(\alpha) := -\lim_{\epsilon\searrow 0} \limsup_{n\to\infty} \frac{1}{|\Lambda_n|} \log \lambda\big\{\pi_{\Lambda_k} L_n \in B_\epsilon(\alpha)\big\}\,, \tag{209}$$

$$= -\lim_{\epsilon\searrow 0} \liminf_{n\to\infty} \frac{1}{|\Lambda_n|} \log \lambda\big\{\pi_{\Lambda_k} L_n \in B_\epsilon(\alpha)\big\}\,. \tag{210}$$

The technical estimate needed for the proof is given in

LEMMA 11.1. *Let $f : \Omega \to \mathbb{R}^l$ be local, and $\phi : \mathbb{R}^l \to [0, \infty)$ be convex. For all $\epsilon > 0$ and $\Lambda \subset \mathbb{Z}^d$ finite, let $\mathcal{E}_\Lambda(\epsilon)$ be the event defined by*

$$\mathcal{E}_\Lambda(\epsilon) := \Big\{\sigma \in \Omega : \phi\Big(\frac{1}{|\Lambda|}\sum_{x\in\Lambda}\theta_x f(\sigma)\Big) < \epsilon\Big\}\,. \tag{211}$$

Then for all $\delta > 0$ and all $\epsilon' > \epsilon$, there exists $M_0 = M_0(\delta, \epsilon, \epsilon')$ and $N_0 = N_0(\epsilon, \epsilon', m, \delta)$ such that for all $m \geq M_0$ and all $n \geq N_0$,

$$\frac{1}{|\Lambda_n|}\log \lambda(\mathcal{E}_{\Lambda_n}(\epsilon)) \geq (1-\delta)\frac{1}{|\Lambda_m|}\log\lambda(\mathcal{E}_{\Lambda_m}(\epsilon'))\,. \tag{212}$$

PROOF. Let $r > 0$ be such that f is $\mathcal{F}_{\Lambda_r}$-measurable. Fix δ, ϵ and ϵ' as in the statement. For large n, we decompose Λ_n into a maximal union of disjoint translates of Λ_{m+r}: $\Lambda_n = \bigcup_{i=1}^K \Lambda^i_{m+r} \cup R_n$. For each i, let $a_i \in \mathbb{Z}^d$ denote the center of the box Λ^i_{m+r}, and let $\Lambda^{(i)}_m := \Lambda_m + a_i$. We write (see Figure 1)

$$\Lambda_n = \bigcup_{i=1}^K \Lambda^{(i)}_m \cup \widetilde{R}_{m,n}\,. \tag{213}$$

The corridors $\widetilde{R}_{m,n}$ satisfy $\lim_m \lim_n \frac{|\widetilde{R}_{m,n}|}{|\Lambda_n|} = \lim_m \frac{|\Lambda_{m+r}\setminus\Lambda_m|}{|\Lambda_m|} = 0$. Since ϕ is convex, it is continuous. Since f is local it is also bounded. Therefore, m and n can be taken large enough so that

$$\phi\Big(\frac{1}{|\Lambda_n|}\sum_{x\in\Lambda_n}\theta_x f(\sigma)\Big) \leq \phi\Big(\frac{1}{|\cup_{i=1}^K \Lambda^{(i)}_m|}\sum_{x\in\cup_{i=1}^K\Lambda^{(i)}_m}\theta_x f(\sigma)\Big) + (\epsilon - \epsilon')\,,$$

uniformly in σ. We also take m and n large enough to that $K|\Lambda_m| \geq (1-\delta)|\Lambda_n|$. By the convexity of ϕ we have, on $\bigcap_{i=1}^K \mathcal{E}_{\Lambda^{(i)}_m}(\epsilon')$,

$$\begin{aligned}\phi\Big(\frac{1}{|\cup_{i=1}^K \Lambda^{(i)}_m|}\sum_{x\in\cup_{i=1}^K\Lambda^{(i)}_m}\theta_x f(\sigma)\Big) &= \phi\Big(\frac{1}{K}\sum_{i=1}^K\frac{1}{|\Lambda^{(i)}_m|}\sum_{x\in\Lambda^{(i)}_m}\theta_x f(\sigma)\Big) \\ &\leq \sum_{i=1}^K\frac{1}{K}\phi\Big(\frac{1}{|\Lambda^{(i)}_m|}\sum_{x\in\Lambda^{(i)}_m}\theta_x f(\sigma)\Big) < \epsilon'\,,\end{aligned} \tag{214}$$

FIGURE 1. The decomposition of Λ_n in (213), in two dimensions. The gray part represents $\widetilde{R}_{m,n}$, the corridors between the boxes $\Lambda^{(i)}_m$.

which implies that $\bigcap_{i=1}^K \mathcal{E}_{\Lambda_m^{(i)}}(\epsilon') \subset \mathcal{E}_{\Lambda_n}(\epsilon)$. So, by independence of the events $\mathcal{E}_{\Lambda_m^{(i)}}(\epsilon')$ and translation invariance of λ,

$$\begin{aligned}\frac{1}{|\Lambda_m|}\log\lambda(\mathcal{E}_{\Lambda_m}(\epsilon')) &= \frac{1}{K|\Lambda_m|}\log\lambda\Big(\bigcap_{i=1}^K \mathcal{E}_{\Lambda_m^{(i)}}(\epsilon')\Big)\\ &\leq \frac{1}{K|\Lambda_m|}\log\lambda(\mathcal{E}_{\Lambda_n}(\epsilon)) \leq \frac{1}{(1-\delta)}\frac{1}{|\Lambda_n|}\log\lambda(\mathcal{E}_{\Lambda_n}(\epsilon)).\end{aligned}$$

□

Observe that the only place where we used the fact that λ is a product measure is in the last display. It happens that the lemma above is a simplification of a more general statement, valid when λ is replaced by any *asymptotically decoupled* measure. See [**?**] or [**?**] for more details.

Proof of Proposition 11.1: We first show the equivalence of the limits (209)-(210) (although both may be infinite). Fix $k \geq 1$ and consider temporarily the alphabet $\mathbb{A} := \Omega_{\Lambda_k}$. We identify $\mathbb{A}$ with a subset of $\mathbb{R}^l$, where $l := |\mathbb{A}_k|$, and write $\mathbb{A} = \{a_1, \dots, a_l\}$. Consider $f : \Omega \to \mathbb{R}^l$ defined by $f(\sigma)_j := 1_{\{\pi_{\Lambda_k}\sigma = a_j\}}$. Fix $\alpha \in \mathcal{M}_1(\mathbb{A})$, and define the convex function $\phi_\alpha : \mathbb{R}^l \to [0,\infty)$ by $\phi_\alpha(x) := \sum_{j=1}^l |\alpha(a_j) - x_j|$. We get

$$\{\pi_{\Lambda_k} L_n \in B_\epsilon(\alpha)\} = \{\phi_\alpha(\pi_{\Lambda_k} L_n) < \epsilon\} \equiv \mathcal{E}^\alpha_{\Lambda_n}(\epsilon)$$

Let $c_n(\epsilon) := |\Lambda_n|^{-1}\log\lambda(\mathcal{E}^\alpha_{\Lambda_n}(\epsilon))$, $\underline{c}(\epsilon) := \liminf_n c_n(\epsilon)$, $\overline{c}(\epsilon) := \limsup_n c_n(\epsilon)$. Fix $\delta > 0$, $\epsilon > \epsilon' > 0$, m and n as in Lemma 11.1. By (212), $c_n(\epsilon) \geq (1-\delta)c_m(\epsilon')$. This implies $\underline{c}(\epsilon) \geq \overline{c}(\epsilon') \geq \underline{c}(\epsilon')$. Taking $\epsilon' \searrow 0$ followed by $\epsilon \searrow 0$ proves that (209)=(210).

We show that $I_\lambda^{(k)}$ is lower semi-continuous. Let α be such that $I_\lambda^{(k)}(\alpha) > a$. Then there exists $\epsilon > 0$ such that

$$\limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\big\{\pi_{\Lambda_k}L_n \in B_\epsilon(\alpha)\big\} < -a\,.$$

If $\alpha' \in B_{\epsilon/2}(\alpha)$ and $\epsilon' < \epsilon/2$, then $B_{\epsilon'/2}(\alpha') \subset B_\epsilon(\alpha)$. Therefore,

$$\limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\big\{\pi_{\Lambda_k}L_n \in B_{\epsilon'}(\alpha')\big\} < -a\,.$$

Taking $\epsilon' \searrow 0$ gives $-I_\lambda^{(k)}(\alpha') < -a$, so $I_\lambda^{(k)}$ is lower semi-continuous.

We then show that

$$I_\lambda^{(k)}\Big(\frac{\alpha+\beta}{2}\Big) \leq \tfrac{1}{2}(I_\lambda^{(k)}(\alpha) + I_\lambda^{(k)}(\beta))\,, \tag{215}$$

which implies that $I_\lambda^{(k)}$ is convex. The proof is similar to that of Lemma 11.1. Consider the decomposition (213), where for simplicity we can assume that K is even. Divide the set of boxes $\Lambda_m^{(i)}$ in two groups: those with $i \in E$, and the others

$j \in D$, where $E = \{1, 2, \dots, K/2\}$, $D = \{K/2, \dots, K\}$. We then proceed as in (214), to decompose [4]

$$\phi_{\frac{\alpha+\beta}{2}}\Big(\frac{1}{|\cup_{i=1}^K \Lambda_m^{(i)}|} \sum_{x \in \cup_{i=1}^K \Lambda_m^{(i)}} \theta_x f\Big) \leq \frac{1}{2|E|} \sum_{i \in E} \phi_\alpha\Big(\frac{1}{|\Lambda_m^{(i)}|} \sum_{x \in \Lambda_m^{(i)}} \theta_x f\Big) + \frac{1}{2|D|} \sum_{j \in D} \phi_\beta\Big(\frac{1}{|\Lambda_m^{(j)}|} \sum_{x \in \Lambda_m^{(j)}} \theta_x f\Big)$$

This implies that

$$\Big(\bigcap_{i \in E} \mathcal{E}^\alpha_{\Lambda_m^{(i)}}(\epsilon') \cap \bigcap_{j \in D} \mathcal{E}^\beta_{\Lambda_m^{(j)}}(\epsilon')\Big) \subset \mathcal{E}^{\frac{\alpha+\beta}{2}}_{\Lambda_n}(\epsilon)\,,$$

from which (215) follows, after the proper limiting procedures. We then show that the sequence $(Q_n^{(k)})$ satisfies a LDP with rate function $I_\lambda^{(k)}$. We start with the lower bound. Let $G \subset \mathcal{M}_1(\mathbb{A})$ be open, non-empty, and take any $\alpha \in G$. If ϵ is small enough, then $G \supset B_\epsilon(\alpha)$. Therefore by (210),

$$\liminf_{n\to\infty} \frac{1}{|\Lambda_n|} \log Q_n^{(k)}(G) \geq \lim_{\epsilon \searrow 0} \liminf_{n\to\infty} \frac{1}{|\Lambda_n|} \log Q_n^{(k)}(B_\epsilon(\alpha)) = -I_\lambda^{(k)}(\alpha)\,.$$

For the upper bound, let $F \subset \mathcal{M}_1(\mathbb{A})$ be closed. Since $\mathcal{M}_1(\mathbb{A})$ is compact, F is also compact. Cover F with a finite set of open balls $B_\epsilon(\alpha_i)$, $\alpha_i \in F$, $i = 1, \dots, N$. By Lemma 5.5 of Chapter 5,

$$\limsup_{n\to\infty} \frac{1}{|\Lambda_n|} \log Q_n^{(k)}(F) \leq \max_{1\leq i\leq N} \limsup_{n\to\infty} \frac{1}{|\Lambda_n|} \log Q_n^{(k)}(B_\epsilon(\alpha_i))$$

After taking $\epsilon \searrow 0$, (209) gives

$$\limsup_{n\to\infty} \frac{1}{|\Lambda_n|} \log Q_n^{(k)}(F) \leq -\min_{1\leq i\leq N} I_\lambda^{(k)}(\alpha_i) \leq -\inf_{\alpha\in F} I_\lambda^{(k)}(\alpha)\,.$$

□

We are now ready to prove the first part of Theorem 11.1, i.e. the validity of a LDP, without yet identifying the rate function as in (205). Remember the metric $\rho(\cdot,\cdot)$ we used to characterize the weak topology on $\mathcal{M}_1(\Omega)$:

$$\rho(\mu,\nu) := \sup_{n\geq 1} \frac{1}{n} \max_{\omega_{\Lambda_n} \in \Omega_{\Lambda_n}} |\mu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n})) - \nu(\Pi^{-1}_{\Lambda_n}(\omega_{\Lambda_n}))|$$

Denote again by $B_\delta(\mu) := \{\mu' : \rho(\mu',\mu) < \epsilon\}$ the open ball of radius δ centered at μ.

Corollary 11.1. *Let Q_n^0 denote the distribution of L_n under the uniform product measure λ. The sequence (Q_n^0) satisfies a LDP on $\mathcal{M}_1(\Omega)$ with speed $|\Lambda_n|$ and good*

[4] Observe that $\phi_{\frac{\alpha+\beta}{2}}(x) \leq \frac{1}{2}\phi_\alpha(x) + \frac{1}{2}\phi_\beta(x)$.

rate function $I_\lambda : \mathcal{M}_1(\Omega) \to [0, +\infty]$*, given by* $I_\lambda(\mu) := \sup_{k\geq 1} I^{(k)}(\pi_{\Lambda_k}\mu)$*.* I_λ *is convex and satisfies:*

$$I_\lambda(\mu) = -\lim_{\epsilon\searrow 0}\limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\bigl\{L_n \in B_\epsilon(\mu)\bigr\}\,, \tag{216}$$

$$= -\lim_{\epsilon\searrow 0}\liminf_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\bigl\{L_n \in B_\epsilon(\mu)\bigr\}\,. \tag{217}$$

Moreover, $I_\lambda = \infty$ *on* $\mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)$*.*

Proof. We first show that there exists $k \geq 1$ such that

$$\lim_{\epsilon\searrow 0}\liminf_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\{L_n \in B_\epsilon(\mu)\} \geq -I_\lambda^{(k)}(\pi_{\Lambda_k}\mu)\,. \tag{218}$$

Namely, fix $\epsilon > 0$ and take $k \geq 0$ such that $1/k < \epsilon$. If $\|\pi_{\Lambda_k}L_n - \pi_{\Lambda_k}\mu\|_1 < \epsilon 2^{-|\Lambda_k|}$, then for all $j \leq k$,

$$\begin{aligned}\frac{1}{j}\max_{\omega_{\Lambda_j}}|\pi_{\Lambda_j}L_n(\omega_{\Lambda_j}) - \pi_{\Lambda_j}\mu(\omega_{\Lambda_j})| &\leq 2^{|\Lambda_k|}\max_{\omega_{\Lambda_k}}|\pi_{\Lambda_k}L_n(\omega_{\Lambda_k}) - \pi_{\Lambda_k}\mu(\omega_{\Lambda_k})|\\ &\leq 2^{|\Lambda_k|}\|\pi_{\Lambda_k}L_n - \pi_{\Lambda_k}\mu\|_1 < \epsilon\,,\end{aligned}$$

and therefore $\rho(L_n, \mu) < \epsilon$. We thus have

$$\liminf_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\{L_n \in B_\epsilon(\mu)\} \geq \liminf_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\Bigl\{\pi_{\Lambda_k}L_n \in B_{\frac{\epsilon}{2^{|\Lambda_k|}}}(\pi_{\Lambda_k}(\mu))\Bigr\}\,,$$

and (218) follows by taking $\epsilon \searrow 0$. On the other hand, $\rho(L_n, \mu) < \epsilon$ implies $\|\pi_{\Lambda_k}L_n - \pi_{\Lambda_k}\mu\|_1 < \epsilon k 2^{|\Lambda_k|}$ for all k, and so

$$\limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\{L_n \in B_\epsilon(\mu)\} \leq \limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\bigl\{\pi_{\Lambda_k}L_n \in B_{\epsilon k 2^{|\Lambda_k|}}(\pi_{\Lambda_k}\mu)\bigr\}\,.$$

By taking the limit $\epsilon \searrow 0$ followed by the infimum over $k \geq 0$, this concludes the proof that (216) = (217) = $\sup_{k\geq 1} I^{(k)}(\pi_{\Lambda_k}\mu)$. The lower semi-continuity of I_λ follows from the lower semi-continuity of each $I_\lambda^{(k)}$ and from the continuity of each projection π_{Λ_k}. The convexity of I_λ also follows from the convexity of $I_\lambda^{(k)}$. Once the above properties of I_λ are established, the upper and lower bounds of the LDP for (Q_n^0) are obtained exactly as for the sequences $(Q_n^{(k)})$. In particular, the upper bound for closed sets is obtained using the compacity of $\mathcal{M}_1(\Omega)$.
For the last statement, let $\mu \in \mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)$. Then $\delta := \rho(\mu, \theta_x\mu) > 0$ for some $x \in \mathbb{Z}^d$. By the triangle inequality, $\delta \leq \rho(\mu, L_n) + \rho(L_n, \theta_x L_n) + \rho(\theta_x L_n, \theta_x\mu)$. Since $\rho(L_n, \theta_x L_n) \to 0$ when $n \to \infty$, uniformly on Ω, we have $\{\rho(\mu, L_n) < \delta/3\} \cap \{\rho(\theta_x\mu, \theta_x L_n) < \delta/3\} = \emptyset$ when n is large enough. If there existed a subsequence L_{n_k} such that $\rho(\mu, L_{n_k}) \to 0$, then we would also have $\rho(\theta_x\mu, \theta_x L_{n_k}) \to 0$, which is impossible. So there exists $\epsilon > 0$ such that $\liminf_n \rho(\mu, L_n) \geq 2\epsilon > 0$. As a consequence, $L_n \notin B_\epsilon(\mu)$ for all large enough n, and so

$$I_\lambda(\mu) = -\lim_{\epsilon\searrow 0}\limsup_{n\to\infty}\frac{1}{|\Lambda_n|}\log\lambda\{L_n \in B_\epsilon(\mu)\} = +\infty\,. \qquad \square$$

3.2. The Free Energy and the Varadhan Lemma. The key to the identification of the rate function I_λ as the relative entropy studied in the previous section is to introduce a generalization of the free energy studied in Chapter 10.

THEOREM 11.2. *For all $f \in C(\Omega)$, $n \geq 1$, let*

$$\psi_n(f|\lambda) := \frac{1}{|\Lambda_n|} \log \int_\Omega \exp\Big(\sum_{x\in\Lambda_n} \theta_x f(\sigma)\Big)\lambda(d\sigma)\,. \tag{219}$$

Then $\psi(f|\lambda) := \lim_{n\to\infty} \psi_n(f|\lambda)$ exists, and equals

$$\psi(f|\lambda) = \sup_{\mu\in\mathcal{M}_1(\Omega)} \{\langle f,\mu\rangle - I_\lambda(\mu)\}\,. \tag{220}$$

The mapping $f \mapsto \psi(f|\lambda)$ is convex. Since $|\psi(f|\lambda) - \psi(g|\lambda)| \leq \|f-g\|$, it is also continuous.

PROOF. By writing $\sum_{x\in\Lambda_n} \theta_x f = |\Lambda_n|\langle f, L_n\rangle$, transporting the measure onto $\mathcal{M}_1(\Omega)$ gives

$$\psi_n(f|\lambda) = \frac{1}{|\Lambda_n|} \log \int_\Omega e^{|\Lambda_n|\langle f,L_n\rangle}\lambda(d\sigma) = \frac{1}{|\Lambda_n|} \log \int_{\mathcal{M}_1(\Omega)} e^{|\Lambda_n|\langle f,\mu\rangle} Q_n^0(d\mu)\,.$$

Since the sequence (Q_n^0) satisfies a LDP with speed $|\Lambda_n|$ and rate function I_λ and since $\mu \mapsto \langle f,\mu\rangle$ is continuous and bounded, the Varadhan Lemma 5.1 implies (220). Convexity of $f \mapsto \psi_n(f|\lambda)$ (and thus of $f \mapsto \psi(f|\lambda)$) follows by the Hölder Inequality, as was done for the Curie-Weiss and Ising models. For the last claim, let $f, g \in C(\Omega)$. Set $f_t := tf + (1-t)g$. Then $f_t' = f - g$ and so

$$|\psi_n(f|\lambda) - \psi_n(g|\lambda)| = \frac{1}{|\Lambda_n|}\Big|\int_0^1 dt \frac{d}{dt}\log\int_\Omega \exp\Big(\sum_{x\in\Lambda_n}\theta_x f_t\Big)d\lambda\Big| \leq \|f-g\|\,. \quad \square$$

The difference between $\psi_n(-\varphi_\Phi|\lambda) \equiv \frac{1}{|\Lambda_n|}\log Z_n^\Phi$ and the finite volume free energy $\frac{1}{|\Lambda_n|}\log Z_{\Lambda_n}^\Phi(\omega_{\Lambda_n^c})$ (see (190) of Chapter 10) is, as we have seen, that the latter contains a boundary term, which we estimated in (199). Since this term vanishes in the thermodynamic limit, $\psi(\Phi) = \log 2 + \psi(-\varphi_\Phi|\nu)$ (the $\log 2$ comes from the fact that $\psi(\cdot|\lambda)$ is defined with the help of the normalized reference measure λ).

We now turn to the link between $\psi(f|\lambda)$ and the relative entropy. Let $\psi^*(\cdot|\lambda) : \mathcal{M}_1(\Omega) \to \mathbb{R}$ denote the Legendre transform of $\psi(f|\lambda)$ with respect to f:

$$\psi^*(\mu|\lambda) := \sup_{f\in C(\Omega)} \{\langle f,\mu\rangle - \psi(f|\lambda)\}\,. \tag{221}$$

PROPOSITION 11.2. *For all $\mu \in \mathcal{M}_{1,\theta}(\Omega)$, $\psi^*(\mu|\lambda) = h(\mu|\lambda)$.*

We need the following result, which was essentially proved in the Introduction of Chapter 10, and which gives the finite volume equivalent of Proposition 11.2.

LEMMA 11.2. *Let $\Lambda \subset \mathbb{Z}^d$ be finite. Then for all $\mu \in \mathcal{M}_1(\Omega)$,*

$$\mathcal{H}_\Lambda(\mu|\lambda) = \sup_U \Big\{\langle U,\mu\rangle - \log\int e^{U(\sigma)}\lambda(d\sigma)\Big\}\,, \tag{222}$$

where the supremum is over all $\mathcal{F}_\Lambda$-measurable functions $U : \Omega \to \mathbb{R}$.

FIGURE 2. The maximal decomposition of Λ_n into translates of Λ_m, with a rest R_n.

PROOF. By Jensen's Inequality,

$$\begin{aligned}\mathcal{H}_\Lambda(\mu|\lambda) - \langle U, \mu\rangle &= -\sum_{\sigma_\Lambda} \mu_\Lambda(\sigma_\Lambda) \log\Big[\frac{\lambda_\Lambda(\sigma_\Lambda)}{\mu_\Lambda(\sigma_\Lambda)} e^{U(\sigma_\Lambda)}\Big] \\ &\geq -\log \sum_{\sigma_\Lambda} e^{U(\sigma_\Lambda)} \lambda_\Lambda(\sigma_\Lambda) \\ &\equiv -\log \int e^{U(\sigma)} \lambda(d\sigma)\,.\end{aligned}$$

On the other hand, if $U_*(\sigma) := \log \frac{\mu_\Lambda(\sigma_\Lambda)}{\lambda_\Lambda(\sigma_\Lambda)}$, then

$$\langle U_*, \mu\rangle - \log \int e^{U_*(\sigma)} \lambda(d\sigma) \equiv \mathcal{H}_\Lambda(\mu|\lambda)\,. \qquad \square$$

PROOF OF PROPOSITION 11.2: We first show that

$$\sup_{f \in \mathcal{L}_{\text{oc}}(\Omega)} \{\langle f, \mu\rangle - \psi(f|\lambda)\} \leq h(\mu|\lambda)\,. \tag{223}$$

Let $f \in \mathcal{L}_{\text{oc}}(\Omega)$ be $\mathcal{F}_{\Lambda_r}$-measurable for some $r \geq 1$. Define $U := \sum_{x \in \Lambda_n} \theta_x f$, which is $\mathcal{F}_{\Lambda_{n+r}}$-measurable. Using Lemma 11.2 in the box Λ_{n+r} and the invariance of μ,

$$\begin{aligned}\mathcal{H}_{\Lambda_{n+r}}(\mu|\lambda) &\geq \langle U, \mu\rangle - \log \int e^{U(\sigma)} \lambda(d\sigma) \\ &= |\Lambda_n| \{\langle f, \mu\rangle - \psi_n(f|\lambda)\}\,.\end{aligned}$$

Dividing by $|\Lambda_n|$ and taking $n \to \infty$ gives $h(\mu|\lambda) \geq \langle f, \mu\rangle - \psi(f|\lambda)$. This proves (223). We then show that

$$\sup_{f \in \mathcal{L}_{\text{oc}}(\Omega)} \{\langle f, \mu\rangle - \psi(f|\lambda)\} \geq h(\mu|\lambda)\,. \tag{224}$$

Let $k \geq 1$ and let f be $\mathcal{F}_{\Lambda_k}$-measurable. For large n, decompose Λ_n into a maximal disjoint union of translates of Λ_k, as we did in the first steps of the proof of Lemma 11.1: $\Lambda_n = \bigcup_{i=1}^K \Lambda_k^i \cup R_n$ (see Figure 2). Let a_i denote the center of Λ_k^i. Then,

$$\Big|\sum_{x \in \Lambda_n} \theta_x f - \sum_{x \in \Lambda_k} \sum_{i=1}^K \theta_{x+a_i} f\Big| = \Big|\sum_{x \in \Lambda_n} \theta_x f - \sum_{i=1}^K \sum_{x \in \Lambda_k^i} \theta_x f\Big| \leq |R_n| \|f\|\,.$$

Now,

$$\begin{aligned}
\psi_n(f|\lambda) &\leq \frac{1}{|\Lambda_n|}\log\int \exp\Big(\sum_{x\in\Lambda_k}\sum_{i=1}^{K}\theta_{x+a_i}f\Big)\lambda(d\sigma)+\frac{|R_n|\|f\|}{|\Lambda_n|}\\
&\leq \frac{1}{|\Lambda_n|}\log\int \exp\Big(|\Lambda_k|\sum_{i=1}^{K}\theta_{a_i}f\Big)\lambda(d\sigma)+\frac{|R_n|\|f\|}{|\Lambda_n|}\\
&= \frac{K}{|\Lambda_n|}\log\int \exp\big(|\Lambda_k|f\big)\lambda(d\sigma)+\frac{|R_n|\|f\|}{|\Lambda_n|}\\
&\leq \frac{1}{|\Lambda_k|}\log\int \exp\big(|\Lambda_k|f\big)\lambda(d\sigma)+\frac{|R_n|\|f\|}{|\Lambda_n|}
\end{aligned}$$

In the second inequality we used the Hölder's Inequality repeatedly, in the equality we used the independence of the functions $\theta_{a_i}f$ and the invariance of λ. Defining $U := |\Lambda_k|f$,

$$\langle f,\mu\rangle-\psi_n(f|\lambda)\geq \frac{1}{|\Lambda_k|}\Big\{\langle U,\mu\rangle-\log\int e^U d\lambda\Big\}-\frac{|R_n|\|f\|}{|\Lambda_n|}\,.$$

Taking $n\to\infty$, the term $\frac{|R_n|\|f\|}{|\Lambda_n|}$ disappears. This shows that

$$\sup_{f\in\mathcal{L}_{\mathrm{oc}}(\Omega)}\{\langle f,\mu\rangle-\psi(f|\lambda)\}\geq \frac{1}{|\Lambda_k|}\Big\{\langle U,\mu\rangle-\log\int e^U d\lambda\Big\}$$

Taking then the sup over U gives, using Lemma 11.2 in the box Λ_k,

$$\sup_{f\in\mathcal{L}_{\mathrm{oc}}(\Omega)}\{\langle f,\mu\rangle-\psi(f|\lambda)\}\geq \frac{1}{|\Lambda_k|}\mathcal{H}_{\Lambda_k}(\mu|\lambda)\,.$$

We then get (224) by taking $k\to\infty$. Since $\mathcal{L}_{\mathrm{oc}}(\Omega)$ is dense in $C(\Omega)$, and since $f\mapsto\langle f,\mu\rangle-\psi(f|\lambda)$ is continuous, $\mathcal{L}_{\mathrm{oc}}(\Omega)$ in (223) and (224) can then be replaced by $C(\Omega)$, which proves that $\psi^*(\mu|\lambda)=h(\mu|\lambda)$. □

Proof of Theorem 11.1: Since we know that $I_\lambda=\infty$ on $\mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)$, we only need to show that $I_\nu(\mu)=h(\mu|\nu)$ for all $\mu\in\mathcal{M}_{1,\theta}(\Omega)$ to conclude the proof of Theorem 11.1. We know that $I_\lambda^*=\psi(\cdot|\lambda)$, and we have just seen that $\psi^*(\mu|\lambda)=h(\mu|\nu)$. That is, $(I_\lambda^*)^*=h(\cdot|\lambda)$ on $\mathcal{M}_1(\Omega)$. Since I_λ is convex and lower semi-continuous, a well-known theorem of convex analysis (see next section) implies that $(I_\lambda^*)^*=I_\lambda$, thus showing that $I_\lambda(\cdot)=h(\cdot|\lambda)$. □

3.3. Parenthesis: Convex duality. The conclusion of the proof of Theorem 11.1 was based on a convex duality argument, saying under which conditions can a function be recuperated from its Legendre transform.

Theorem 11.3. *Let $\mathcal{X}$ be a locally convex Hausdorff topological vector space. Let $F:\mathcal{X}\to(-\infty,\infty]$ be convex and lower semi-continuous. Let $\mathcal{X}^*$ denote the topological dual of $\mathcal{X}$, i.e. the set of all continuous linear functionals $x_*:\mathcal{X}\to\mathbb{R}$ (we write $x_*(x)\equiv\langle x,x_*\rangle$). Let $F^*:\mathcal{X}^*\to\mathbb{R}$ be the Legendre transform of F:*

$$F^*(x_*):=\sup_{x\in\mathcal{X}}\{\langle x,x_*\rangle-F(x)\}\,. \tag{225}$$

Then F is the Legendre transform of F^*: for all $x \in \mathcal{X}$,

$$F(x) = \sup_{x_* \in \mathcal{X}^*} \{\langle x, x_* \rangle - F^*(x_*)\} . \tag{226}$$

For our needs, $\mathcal{X} \equiv \mathcal{M}(\Omega)$, the space of signed measures on Ω, whose dual is $C(\Omega)$, and $F := I_\lambda$ (extended by setting $I_\lambda := \infty$ on $\mathcal{M}(\Omega)\backslash\mathcal{M}_1(\Omega)$).
Below, we give a proof of this result in the case where $\mathcal{X} \equiv \mathbb{R}$, whose dual is $\mathbb{R}$ itself, and $\langle x, x_* \rangle := xx_*$, but we continue using the notation with brackets $\langle \cdot, \cdot \rangle$.

Theorem 11.4. *Let $f : \mathbb{R} \to (-\infty, \infty]$ be convex, lower semi-continuous, and such that $f \not\equiv \infty$. Then $(f^*)^* = f$.*

Let the epigraph of f be defined by $\mathcal{E} := \{(x, y) : y \geq f(x)\}$. It is easy to show that $\mathcal{E} \subset \mathbb{R}^2$ is closed (resp. convex) if and only if f is lower semi-continuous (resp. convex). Of course, $\mathcal{E} \neq \emptyset$ if and only if $f \not\equiv \infty$. The epigraph of f^*, denoted $\mathcal{E}^*$, is defined in the same way.

Proof. First, write $f(x) = \inf_{y \geq f(x)} y$, and the Legendre transform as

$$f^*(x_*) = \sup_{x \in \mathbb{R}} \sup_{y \geq f(x)} \{\langle x, x_* \rangle - y\} \equiv \sup_{(x,y) \in \mathcal{E}} \{\langle x, x_* \rangle - y\} .$$

Therefore, $f^*(x_*) \geq \langle x, x_* \rangle - y$ for all x_* and for all $(x, y) \in \mathcal{E}$. As a consequence, $y \geq \sup_{x_*} \{\langle x, x_* \rangle - f^*(x_*)\}$ for all $(x, y) \in \mathcal{E}$. In particular,

$$f(x) \geq \sup_{x_*} \{\langle x, x_* \rangle - f^*(x_*)\} , \quad \forall x . \tag{227}$$

To show the converse inequality, i.e. $f(x) \leq \sup_{x_*} \{\langle x, x_* \rangle - f^*(x_*)\}$, we must show that for all $y < f(x)$, there exists x_* such that $\langle x, x_* \rangle - f^*(x_*) \geq y$. To do so, we will show the following stronger statement:

$$\forall (x, y) \notin \mathcal{E} , \ \exists \, (x_*, y_*) \in \mathcal{E}_* \text{ such that } \langle x, x_* \rangle - y_* \geq y .$$

Since $\mathcal{E}$ is convex and closed, there exists a unique $(x_0, y_0) \in \mathcal{E}$ which realizes the distance from (x, y) to $\mathcal{E}$ (see Figure 3). Let π denote the straight line

Figure 3. The separating line π.

separating (x, y) from $\mathcal{E}$, orthogonal and passing through the midpoint of the segment joinging (x, y) to (x_0, y_0). Let $(x_*, y_*) \in \mathbb{R}^2$ be the pair characterizing the inclination and abcisse at the origin of π, in the sense that π has equation $x \mapsto \langle x, x_* \rangle - y_*$. Since (x, y) lies below π, we have $\langle x, x_* \rangle - y_* \geq y$. Then, since $\mathcal{E}$ lies above π, we have $f(x) \geq \langle x, x_* \rangle - y_*$ for all x, and so $y_* \geq \langle x, x_* \rangle - f(x)$, i.e. also $y_* \geq f^*(x_*)$. That is, $(x_*, y_*) \in \mathcal{E}_*$. □

Observe that the first half of the proof goes through in the case treated in Theorem 11.1 giving the equivalent of the lower bound (227) for all $\mu \in \mathcal{M}_1(\Omega)$:

$$I_\lambda(\mu) \geq \sup_{f \in C(\Omega)} \{\langle f, \mu \rangle - \psi(f|\lambda)\} .$$

For the upper bound, a geometric argument similar to the one used on Figure 3 must be used, but in an infinite-dimensional setting. There, the existence of separating hyperplanes follows by an abstract functional-analytic argument (the Geometric Hahn-Banach Theorem). We refer the reader to Chapter 6 of [?] for a complete proof of Theorem 11.3.

4. The LDP for L_n under the Free Gibbs measure

We can finally make the link between the Variational Principle of the previous chapter and the Large Deviations of the empirical field.

THEOREM 11.5. *Let Φ be a translation invariant uniformly absolutely summable potential. Let Q_n^Φ denote the distribution of L_n under μ_n^Φ. The sequence (Q_n^Φ) satisfies a Large Deviation Principle with a good rate function $J^\Phi : \mathcal{M}_1(\Omega) \to [0, +\infty]$ given by*

$$J^\Phi(\mu) := \begin{cases} h(\mu|\Phi) & \text{if } \nu \in \mathcal{M}_{1,\theta}(\Omega)\,, \\ \infty & \text{if } \nu \in \mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)\,. \end{cases} \tag{228}$$

PROOF. We know that $\mu \mapsto \langle -\varphi_\Phi, \mu\rangle$ is continuous and bounded. By the representation of Q_n^Φ in (204), the LDP for (Q_n^0) in Theorem 11.1, and from Theorem 5.2 on tilted measures, (Q_n^Φ) satisfies a LDP with the rate function

$$\begin{aligned} J^\Phi(\mu) &= \sup_{\mu\in\mathcal{M}_1(\Omega)} \{\langle -\varphi_\Phi, \mu\rangle - I_\lambda(\mu)\} - \{\langle -\varphi_\Phi, \mu\rangle - I_\lambda(\mu)\}\,, \\ &= \psi(-\varphi_\Phi|\lambda) - \{\langle -\varphi_\Phi, \mu\rangle - I_\lambda(\mu)\}\,. \end{aligned}$$

If $\mu \in \mathcal{M}_1(\Omega)\backslash\mathcal{M}_{1,\theta}(\Omega)$, then $J^\Phi(\mu) = \infty$. Remembering that $\psi(-\varphi_\Phi|\lambda) = \psi(\Phi) - \log 2$ and that $I_\lambda(\mu) = h(\mu|\lambda) = \log 2 - h(\mu)$ on $\mathcal{M}_{1,\theta}(\Omega)$ we get, for $\mu \in \mathcal{M}_{1,\theta}(\Omega)$,

$$J^\Phi(\mu) = \psi(\Phi) - \{h(\mu) - \langle \varphi_\Phi, \mu\rangle\} \equiv h(\mu|\Phi)\,,$$

by the Variational Principle. □