Large Deviations Theory
Lecture notes
Master 2 in Fundamental Mathematics
Université de Rennes 1

Mathias Rousset

2021

Documents to be downloaded

Download pedagogical package at http://people.irisa.fr/Mathias.Rousset/****.zip where **** is the code given in class.

Main bibliography

Large Deviations


- Touchette [2009]: readable, formal/expository introduction to the link with statistical mechanics.

- Rassoul-Agha and Seppäläinen [2015]: the first part is somehow similar to the content of this course.

- Freidlin et al. [2012]: a well-known classic, reference for small noise large deviations of SDEs. The vocabulary differs from the modern customs.

- Feng and Kurtz [2015]: a rigorous, general treatment for Markov processes.

General references

- Bogachev [2007]: exhaustive on measure theory.


- Brézis [1987], Brezis [2010]: excellent master course on functional analysis.

- Billingsley [2013]: most classical reference for convergence of probability distributions.
Notations

- $C_b(E)$: set of continuous and bounded functions on a topological space $E$.
- $M_b(E)$: set of measurable and bounded functions on a measurable space $E$.
- $\mathcal{P}(E)$: set of probability measures on a measurable space $E$.
- $\mathcal{M}_\mathbb{R}(E)$: set of real valued (that is, finite and signed) measures on a measurable set $E$.
- i.i.d.: independent and identically distributed.
- $\mu(\varphi) = \int \varphi \, d\mu$ if $\mu$ is a measure, and $\varphi$ a measurable function.
- $X \sim \mu$ means that the random variable $X$ is distributed according to the probability $\mu$.
- Stars (**) or (*) indicates importance.
1 Introduction

1.1 The most important idea

The theory of large deviations may be summarized as "the asymptotic theory of rare events";

where "asymptotic" refers to the limit of a sequence of probability distributions

\[ \mu_n \overset{\text{def}}{=} \text{Law} \left( X_n \right) \in \mathcal{P}(E), \quad n \geq 1 \]

defined on a given (nice enough) measurable space \( E \) (\( X_n \in E \)).

We may switch between the notation \( (\mu_n)_{n \geq 1} \) or a random variables representation \( (X_n)_{n \geq 1} \) depending on notational convenience.

In this context, a "rare event" can be simply defined as any event \( A \) with vanishing probability:

\[ P(X_n \in A) \xrightarrow{n \to +\infty} 0. \]

Two classical examples are given by:

- (Small noise) \( X_n = f(U/n) \) is a small random perturbation of a deterministic state \( X_n \to_x x_\infty = f(0) \in E \), \( U \) being a given random vector and \( f \) a continuous function.

- (Large sample size) \( X_n \) is given as the average of \( n \) i.i.d. variables.

It turns out that quite generally, if one tries to look at the probability of such rare events using a given logarithmic scale (which is a **very rough** picture), then the rate of vanishing can be described by a (usually) explicit minimization problem. What mean by this is that for well-behaved subsets \( A \),

\[ -\log P(X_n \in A) \sim_{n \to +\infty} c_n \inf_{z \in A} I(z), \]

where \( c_n > 0 \) for \( n \geq 1 \) is a sequence called the speed verifying

\[ \lim_{n \to +\infty} c_n = +\infty, \]

and

\[ I : E \to [0, +\infty] \]

is a 'cost' function called the rate function.

Note that the speed and the rate function are well-defined separately only up to a conventional multiplicative constant, that is if we decide to multiply the reference speed by a constant \( c_0 \), the rate function has to be divided by the same value \( c_0 \).

The property (1), which is a simplified version of the so-called large deviation principle (LDP), has the following two key interpretations:

i) The probability of an asymptotically rare event \( \{ X_n \in A \} \) is determined by the least rare outcomes in \( A \), and only those. Those least rare outcomes are exactly those who minimizes the rate function \( I \).
ii) A probability conditioned by the rare event \( \{ X_n \in A \} \) is asymptotically concentrated on those **least rare outcomes in** \( A \) **that minimize** \( I \).

The specific speed \( c_n \) for \( n \geq 1 \) has to be chosen appropriately, in a way such that the rate function \( I \) is **non-trivial** that is

\[
\exists x \in E, \; 0 < I(z) < +\infty.
\]

In this notes, we will restrict to the case

\[ c_n = n, \quad n \geq 1, \]

for the following two reasons:

- The theory is completely similar for general speeds \( c_n, n \geq 1 \).
- We will mainly (if not only) discuss specific results when \( X_n \) is constructed as an average of \( n \) i.i.d. variables. The associated appropriate speed is given by \( c_n = n \).

We end this first subsection by the most fundamental lemma and corollary of the Large Deviations Theory.

**Lemma 1.1.1.** Consider a sequence \( \{ a_n \}_{n \geq 1} \) taking values in \( [0, +\infty]^k \) for some given \( k \in \mathbb{N}_* \). Assume that for each \( j, 1 \leq j \leq k \),

\[
\frac{1}{n} \ln a_n(j) \xrightarrow{n \to +\infty} l(j) \in [-\infty, +\infty]
\]

for some \( l \in [-\infty, +\infty]^k \). Then,

\[
\frac{1}{n} \ln \sum_{j=1}^{k} a_n(j) \xrightarrow{n \to +\infty} \max_{1 \leq j \leq k} l(j).
\]

**Proof.** Exercise below. \( \square \)

**Corollary 1.1.2.** Consider a sequence \( \mu_n = \text{Law}(X_n) \in \mathcal{P}(E), \; n \geq 1 \) on a **finite** state space \( E \). Then the Large Deviation Principle (LDP) (1) holds for any subset \( A \subset E \) if and only if

\[
\frac{1}{n} \ln \mathbb{P}(X_n = i) \xrightarrow{n \to +\infty} -I(i)
\]

for each \( i \in E \) and for some function \( I : E \to [0, +\infty] \).

**Proof.** Exercise below. \( \square \)

The next two exercises are fundamental to understand the robustness and generality of large deviations theory.

**Exercise 1.1.3 \((**)\).** Prove Lemma 1.1.1.

**Exercise 1.1.4 \((**)\).** Prove Corollary (1.1.2). Show that up to extraction, any sequence of probability distribution satisfies a LDP with (possibly trivial) rate function \( I \). Construct simple examples of non-trivial LDPs. Construct a simple example where a non-trivial LDP holds for two different speeds.
Exercise 1.1.5 (**). Consider the case a finite state space $E$. Make sense of the interpretation $ii$) above by showing that if $\inf_{A \cap B} I > \inf_A I$, then $\mathbb{P}(X_n \in B | X_n \in A)$ vanishes exponentially fast.

Exercise 1.1.6 (**). Write down all the formulas in the present section for the case where $X_n$ is the average of $n$ standard Gaussians in $\mathbb{R}^d$ and the rare event is described by an affine half-space $A = \{ z \in \mathbb{R}^d | z_1 > 1 \}$. Interpret geometrically.

Exercise 1.1.7. Prove that if (1) holds with speed $c_n$, then it still holds for any $c'_n \sim n \text{ cte} \times c_n$ with cte $> 0$. Prove that extracted subsequences of $(\mu_n)_{n \geq 1}$ satisfy (1) with arbitrarily fast speed.

1.2 The need for a state space topology and a precise definition of LDP

The simple form of LDP stated in (1) cannot be true for any Borel set $A$ if the state space $E$ is not discrete.

Exercise 1.2.1 (*). Using the example of Exercise 1.1.6, prove that the naïve LDP (1) is false for any countable subset $A$.

Intuitively, the rate function $I$ looks in a specific area in the state space and tells us how fast the probability mass of $X_n$ goes away. This is a notion which have similarity with the convergence in distribution, and requires the introduction of a topology on the state space $E$.

Definition 1.2.2. Assume that the state space $E$ in endowed with a (reasonable, i.e. Polish, see below) topology. When $n \to +\infty$, $(\mu_n)_{n \geq 1}$ is said to converge in distribution towards $\mu_\infty$ if $\mu_n(\varphi)$ converges to $\mu_\infty(\varphi)$ for any continuous and bounded $\varphi \in C_b(\mathcal{F})$.

We have then the Portmanteau theorem:

Theorem 1.2.3 (Portmanteau). Convergence in distribution is equivalent to

$$\limsup_n \mathbb{P}(X_n \in C) \leq \mathbb{P}(X_\infty \in C), \quad \forall C \text{ closed}$$

or equivalently, taking complementaries in $E$:

$$\liminf_n \mathbb{P}(X_n \in O) \geq \mathbb{P}(X_\infty \in O), \quad \forall O \text{ open}$$

It can be checked that if the limit $X_\infty$ has a strictly positive probability to belong to the boundary of $C$ or $O$, then the inequalities above may not be equalities.

Exercise 1.2.4 (**). Construct a sequence of probabilities for which the inequality in Pormanteau is not an equality.

In the same way, large deviations are rigorously defined using an upper bound for closed sets and a lower bound for open sets. Recall that the interior $A$ of a set $A$ is the largest open set contained in $A$, and the closure $\overline{A}$ is the smallest closed set containing $A$. The rate function $I$ must also have some form of continuity called lower-semi continuity.
**Definition 1.2.5** (Lower semi-continuity). A \([-\infty, +\infty]\)-valued function \(f\) on a topological space \(E\) is said to be lower semi-continuous if \(\{x \in E| f(x) \leq a\}\) is closed for any \(a \in \mathbb{R}\). If \(E\) is metric this is equivalent to
\[
f(\lim_{n \to +\infty} x_n) \leq \liminf_{n \to +\infty} f(x_n)
\]
for any converging sequence \((x_n)_{n \geq 1}\).

**Exercise 1.2.6** (**). Show that if \(f\) is lower semi-continuous, then \(\{x \in E| f(x) \leq a\}\) is closed also if \(a = +\infty\) or \(a = -\infty\).

**Definition 1.2.7** (Large Deviations Principle (LDP)). Let \((X_n)_{n \geq 1}\) a sequence of random variables on a measurable set \(E\). If there is a topology on \(E\) and a function \(I\) such that
i) \(I : E \to [0, +\infty]\) is lower semi-continuous.
ii) For any measurable set \(A\)
\[
-\inf_{x \in A} I(x) \leq \liminf_{n \to +\infty} \frac{1}{n} \log P\{X_n \in A\} \leq \limsup_{n \to +\infty} \frac{1}{n} \log P\{X_n \in A\} \leq -\inf_{x \in \mathcal{A}} I(x).
\]
Then we say that the sequence of distribution \(\mu_n = \text{Law}(X_n), n \geq 1\) satisfies a LDP with speed \(n\) and rate function \(I\).

(Nota Bene: In most cases \(E\) is a Polish topological space (e.g.: \(E = \mathbb{R}^d\)) and the measurable sets of \(E\) are the associated Borel subsets. However the definition of LDP makes sense in general, a feature required for various advanced results you may come across.)

Another recurrent element in the 'jargon' of LDP

**Definition 1.2.8** (Goodness). A rate function \(I\) is said to be good if the (closed) level sets \(\{x| I(x) \leq a\}\) are compact for all \(a \geq 0\).

**Exercise 1.2.9** (**). Prove that if \(I\) is a rate function then \(\inf I = 0\).

**Exercise 1.2.10** (**, Discrete case). Assume for simplicity \(E = \mathbb{N}_+\) (Nota Bene: this exercise can be easily generalized to any discrete space \(E\)). Simplify the definition of the LDP. Describe good rate functions functions. Show that the LDP with good rate function \(I\) is equivalent to the two following conditions:

- A condition called 'exponential tightness' that we will discuss in details later on:
\[
\limsup_{x \to +\infty} \limsup_{n \to +\infty} \frac{1}{n} \log P(X_n \notin \{1, 2, \ldots, x\}) = -\infty.
\]

- The identification of the rate function by
\[
\frac{1}{n} \log P(X_n = x) \underset{n \to +\infty}{\longrightarrow} -I(x), \forall x \in E
\]
Exercise 1.2.11 (**) Consider the rate function on \( \mathbb{N} \): \( I(0) = 0 \) and \( I(k) = +\infty \) for \( k \geq 1 \). Construct a sequence of distribution \((\mu_n)_n\) on \( \mathbb{N} \) such that \( \frac{1}{n} \ln \mu_n(\{k\}) \to -I(k) \) for all \( k \), but the LDP is not satisfied (Hint: transport mass towards infinity).

Exercise 1.2.12 (**) Prove (on metric spaces) that good lower semi-continuous functions always have a minimum on each closed set.

Exercise 1.2.13 (**) Prove that the constant distributions \( \mu_n = \mu_0, \forall n \geq 1 \) satisfy a LDP and describe the rate function, using the concept of support of a measure in a topological space: \( \text{supp} (\mu_0) \) is the (closed) set defined as the set of all points whose neighborhoods all have strictly positive measure. Compute the support of usual distributions in \( \mathbb{R}^d \).

Exercise 1.2.14 (**) Assume that \( E_0 \subset E \) is a closed subset such that \( \mu_n(E_0) = 1 \) for all \( n \) large enough. Check that the LDP in \( E_0 \) is equivalent to the LDP in \( E \) for the trace topology and the rate function extended to \( +\infty \) outside \( E_0 \).

Exercise 1.2.15 (**, Lower semi-continuity) Describe all the lower semi-continuous functions on \( \mathbb{R} \) that are continuous on \( \mathbb{R} \setminus \{0\} \).

Exercise 1.2.16 (*). Prove that if one relaxes the condition of lower semi-continuity of \( I \), then \( I \) may not be unique.

Exercise 1.2.17 (**, weak law of large number). Prove with Portmanteau theorem that if a LDP holds true and \( I \) has a unique minimizer: \( \{x | I(x) = 0\} = \{x_0\} \) for some \( x_0 \) in \( E \), then \( \mu_n \) converges in distribution towards \( \delta_{x_0} \).

Exercise 1.2.18 (*). Assume \( E \) is Polish. Prove that if the LDP holds true with a good rate function, then the sequence is tight: for any \( \varepsilon > 0 \), there is a compact \( K_\varepsilon \) such that
\[
\liminf_n \mathbb{P}(X_n \in K_\varepsilon) \geq 1 - \varepsilon.
\]

Exercise 1.2.19 (*). Prove rigorously the LDP in the case where \( X_n \) is distributed according to a Gaussian \( \mathcal{N}(0, \frac{1}{n}) \) in \( \mathbb{R} \) and \( A \) is an interval. (Hint: prove first that \( \frac{1}{n} \ln \mathbb{P}(X_n \in A) \) converges towards the infimum of \( x^2/2 \) over \( A \).

This last exercise is a particular example of LDP that can be generalized in various ways (see Laplace’s principle, Varadhan’s lemma and Cramér theorem). Those links will be discuss in more details during the course. For the moment we will generalize a bit the result of the exercise above:

Lemma 1.2.20 (Laplace’s principle in \( \mathbb{R} \). Let \( X_n \in \mathbb{R} \) for \( n \geq 1 \) distributed according to a density of the form:
\[
\text{Law}(X_n) = \frac{e^{-nI(x)} \, dx}{\int_{\mathbb{R}} e^{-nI(x)} \, dx}
\]
where \( I \) is a lower semi-continuous functions that is also good (that is it goes to \( +\infty \) at \( \pm \infty \)). Then \( X_n, n \geq 1 \) satisfies a LDP with rate function \( I \).

Proof. Section 4.
1.3 Cramér’s theorem in $\mathbb{R}$

Exercise can also be generalized as follows:

**Theorem 1.3.1** (Cramér). Consider a sequence of averages 

$$X_n := \frac{1}{n} \sum_{m=1}^{n} Z_m$$

where $Z_m$, $m \geq 1$ are i.i.d. taking value in $\mathbb{R}$. Assume that the cumulant generating function given by

$$\Lambda(l) \overset{\text{def}}{=} \ln \mathbb{E} \left[ e^{lZ} \right]$$

is finite on some open neighborhood\(^1\) of $l = 0$.

Then $(X_n)_{n \geq 1}$ satisfies a LDP with good convex rate function defined by

$$I(x) = \Lambda^*(x) = \sup_{l \in \mathbb{R}} (lx - \Lambda(l)),$$

called the convex dual of $\Lambda$.

*Proof.* Section 4. \hfill $\square$

1.4 The Gibbs conditioning principle

Assume $(X_n)_n$ satisfies a LDP with rate function $I$ and that $\mathbb{P}(X_n \in B) > 0$ for all $n$ large enough and a Borel set $B$. A natural question consists in studying the asymptotic behavior of the conditional probability,

$$\lim_{n \to +\infty} \text{Law} \left( X_n | X_n \in B \right).$$

Since large deviations quantifies the logarithmic cost of the least unlikely states, it is natural to expect that the conditional property above will **concentrate in the minima of $I$ on $B$.**

**Proposition 1.4.1.** Assume that $X_n$, $n \geq 1$ satisfies a LDP with good rate function $I$. Assume that $B$ is closed and that

$$\inf_{B} I = \inf_{B} I.$$

Then, exponentially fast,

$$X_n \overset{\mathbb{P}[|X_n \in B]|_{n \to +\infty}}{\longrightarrow} \text{arginf}_{B} I$$

in the sense that if $A$ is any open set with

$$\text{arginf}_{B} I \subset A$$

then $\mathbb{P}[X_n \notin A | X_n \in B]$ tends exponentially fast to 0 with $n$.

**Exercise 1.4.2** (**). Draw a picture in the Gaussian case. Prove the above lemma directly from the LDP.

\(^1\)N.B.: this condition could be relaxed, but this leads to degenerate cases
A possible refined version of the Gibbs conditioning principle is the following.

**Exercise 1.4.3.** Let $X_n \in E$, $n \geq 1$ satisfies a LDP with rate function $I$. Let $B \subset E$ measurable such that $\mathbb{P}(X_n \in B) > 0$ for each $n$. Assume there is a unique minimizer $x^*_n$ of $I$ in $B$ satisfying:

- $\inf_{B} I = I(x^*_n)$,
- The minimum $x^*_n$ in $B$ is attained locally only

$$\inf_{O_{x^*_n} \cap B} I > I(x^*_n) \quad \forall \text{ open } O_{x^*_n} \ni x^*_n.$$ 

Then prove that the conditional distribution $\text{Law}(X_n | X_n \in B)$ converges in law towards $\delta_{x^*_n}$ (that is $X^*_n \to x^*_n$ in probability where $\text{Law}(X^*_n) = \text{Law}(X_n | X_n \in B)$).

Hints: Check using Portmanteau theorem that convergence in law towards a deterministic $x^*_n$ is equivalent to convergence to 0 of the probability of being outside any neighborhood of $x^*_n$. Then apply the LDP to $\mathbb{P}(X_n \in O_{x^*_n} \cap B)$ and $\mathbb{P}(X_n \in B)$ to conclude.

### 1.5 Remarks on the state space $E$ and the large deviation topology

This section is psychological preparation to measure-valued LDPs. **See Section 2 for more details.** When the LDP is considered on $E = \mathbb{R}^d$ there is no difficulty: events are defined with Borel subsets and the topology of $\mathbb{R}^d$ is the usual one. It is no longer the case when considering empirical measures.

As usual in probability, we will assume that the measurable state space $E$ is ‘reasonable’: measurable sets of $E$ are given by the Borel sets of a Polish topology:

**Assumption (Standard Borel).** The measurable sets of the state space $E$ were random variables taking value

$$X_n \in E$$

are given by the Borel sets (that is the $\sigma$-algebra generated by open sets) of some Polish topology.

Note that such spaces are either countable, or measurably isomorphic to $]0,1[$ (see after).

A Polish topology is by definition separable (that is it has a dense countable subset) and completely metrizable. Polish spaces include many topological spaces used in modeling such as:

- Any countable set with the discrete topology.
- Any open or closed subset of $\mathbb{R}^d$.
- Any separable Banach space.

For instance, the space of bounded measurable function on $\mathbb{R}$ or $L^\infty(\mathbb{R},dx)$ are not separable hence not Polish.
For our purpose, an important example is the space of all probability distributions

\[ E = \mathcal{P}(F) \]

where \( F \) is a Polish state space. It possesses a natural and obvious \( \sigma \)-algebra of measurable sets, called the cylindrical \( \sigma \)-algebra, and defined as the smallest one making the maps \( \mu \mapsto \int \phi d\mu \) measurable for each measurable bounded test function \( \phi \). We will see that the latter are exactly the same as the Borel sets associated with convergence in distribution on \( \mathcal{P}(F) \).

It is important to remark that the topology considered in the Large Deviation Principle may be chosen appropriately depending on the problem at hand. In particular, one may be interested in looking for LDP in the finest possible topology.

**Exercise 1.5.1.** Check that if a LDP holds true for a given topology, then it is also true for any coarser topology, that is a topology with less open sets.

The type of topology that is required in order to carry out the general theory of LDPs is quite general.

**Assumption (Regularity).** The topology considered in LDPs are supposed to be at least regular: any point \( x \) and any closed set \( C \) can be separated by neighborhoods, that is they can respectively be included in two disjoint open sets.

Practically, all topologies considered in probability and analysis are regular, in particular all metric and locally convex Hausdorff topologies (and traces of such) are regular (in fact completely regular) – they are the only reasonable topologies considered on topological vector spaces.

**Exercise 1.5.2.** Check that if the topology is regular, then the rate function is unique.

However in most cases we will consider the standard case.

**Assumption.** Except otherwise mentioned, the state space \( E \) where LDP are considered is a Polish (and thus regular) space endowed with its Borel \( \sigma \)-field. For \( E = \mathcal{P}(F) \) the default case is the topology of convergence in distribution.

### 1.6 Sanov theorem

Sanov theorem is the LDP for empirical measures of i.i.d. random variables taking value in some state space \( F \). The LDP is given in the state space of probability distributions

\[ E \overset{\text{def}}{=} \mathcal{P}(F), \]

Assume \( F \) is Polish, and let \( \mathcal{P}(F) \) be endowed with convergence in distribution.

**Theorem 1.6.1** (Sanov). Let \( (Y_m)_{m \geq 1} \) denote a sequence of of i.i.d. variables in Polish \( F \) with Law\((Y_m) = \pi_0 \). The sequence of empirical distribution defined for \( n \geq 1 \) by

\[ X_n \overset{\text{def}}{=} \frac{1}{n} \sum_{m=1}^{n} \delta_{Y_m} \in E \]
satisfy a LDP on $\mathcal{P}(F)$ endowed with convergence in distribution with good rate function

$$I(\pi) \overset{\text{def}}{=} \text{Ent}(\pi|\pi_0),$$

where $\text{Ent}$ is the relative entropy (a.k.a. Kullback-Leibler divergence) between $\pi$ and $\pi_0$.

The relative entropy is defined by

$$\text{Ent}(\pi|\pi_0) \overset{\text{def}}{=} \begin{cases} \int_F \ln \frac{d\pi}{d\pi_0} d\pi & \pi \ll \pi_0 \\ +\infty & \text{else} \end{cases}$$

**Exercise 1.6.2.** Check that $\text{Ent}(\pi|\pi_0)$ is positive and vanishes if and only if $\pi = \pi_0$.

**Remark 1.6.3** (On the topology). In fact, Sanov can be strengthened to a stronger weak topology called the $\tau$-topology, also denoted the $\sigma(M_R,M_b)$-topology or the $M_b$-initial topology. This is the ‘weak’ i.e. the coarsest topology making all the maps $\mu \mapsto \int F \psi d\mu$ continuous, where $\psi$ is only measurable (and not continuous) and bounded. This is much finer topology than convergence in distribution (that is the convergence is stronger).

Sanov theorem is remarkable because it relates independence with (relative) entropy which is derived as a secondary concept. In statistical mechanics, the relative entropy above may not exactly be the Boltzmann entropy. Depending on cases or simplification choices, it can be for instance a non-interacting free energy or the opposite of (a variant) of the Boltzmann entropy. Those links will be discussed later on.

For now on the proof of Sanov is left as a mystery. The reader should however keep in mind that there are two main roads:

- **Let $F$ be a finite.** A direct combinatorial estimation of

  $$\frac{1}{n} \ln \mathbb{P}[\text{empirical density = given density}]$$

  (see Lemmas 2.1.2 and 2.1.9 in Dembo and Zeitouni [1998]) can be used as start. This is the road followed by Boltzmann (implicitly) and information theory.

- **Convex duality.** This is the classical road that was developed independently in statistics, and it will be the one we follow.
1.7 Problems

Problem 1.7.1 (A Statistical Physics Toy Model, **). Assume \( F = \{1, 2, \ldots, k\} \subset \mathbb{N} \) is a finite state space describing the possible energy values of a particle with given \( k \). We consider \( n \) particles that can exchange energy while the whole system conserves total energy. We assume that the particle system is distributed according to the uniform distribution over all possible states with given total energy given by \( n \times e \) where \( e \in F \). For technical reasons related to conditioning, it will be much simpler to condition the energy of the above particle system in the interval \([e - \delta e, e]\) with \( \delta e > 0 \) and \( e \in [0, k] \). The goal of this problem is to prove that the empirical distribution of particles conditioned by total mean energy in \([e - \delta e, e]\) converges towards the unique compatible density maximizing physical entropy.

\( \alpha \)-i) Check that the model can be described as \( n \) i.i.d. random variables \((Y_1, \ldots, Y_n) \in F^n\) with uniform distribution and conditioned by the event defined by constant mean energy:

\[
\frac{1}{n} \sum_{m=1}^{n} Y_m \in [e - \delta e, e].
\]  (4)

\( \alpha \)-ii) Recall Sanov theorem for the empirical measures

\[
\Pi_n = \frac{1}{n} \sum_{m=1}^{n} \delta_{Y_m} \in \mathcal{P}(F) \subset \mathbb{R}^k.
\]

We denote by \( I = \text{Ent}(., \text{Unif}) \) the rate function. Check that convergence in distribution in \( \mathcal{P}(F) \) is given by the usual (trace) topology of \( \mathbb{R}^k \), prove that \( I \) is continuous on \( \mathcal{P}(F) \), and that \( \mathcal{P}(F) \) is compact.

\( \alpha \)-iii) Consider the set of probabilities:

\[
B \overset{\text{def}}{=} \left\{ \pi \in \mathcal{P}(F) \mid \sum_{i=1}^{k} i\pi(i) \in [e - \delta e, e] \right\}.
\]

Check that \( B \) is closed as a subset of \( \mathbb{R}^k \), describe its interior, and that the closure of the interior \( \bar{B} \) is again \( B \). Check that the event \( \Pi_n \in B \) is equivalent to the event (4).

\( \alpha \)-iv) Assume that there exists a unique minimizer \( \pi_e \) of \( I \) on \( B \). Prove using a compactness argument that \( \inf_{A \cap B} I > I(\pi_e) \) for any open neighborhood \( A \) of \( \pi_e \). Check also using the continuity of \( I \) that \( \inf_{\bar{B}} I = I(\pi_e) \). Conclude using the Gibbs conditioning principle that \( \Pi_n \) conditioned by \( B \) converges in distribution towards \( \pi_e \).

We will now study the minimizers of \( I \) on the convex set:

\[
L_e \overset{\text{def}}{=} \left\{ \pi \in \mathcal{P}(F) \mid \sum_{i=1}^{k} i\pi(i) = e \right\}.
\]
\( \pi_e \) is called a critical point of \( I \) on \( L_e \) iff \( I \) is differentiable at \( \pi_e \)

\[
\frac{d}{dt}_{t=0} I(\pi_e + tv) = 0
\]

for any \( v \in \mathbb{R}^k \) such that \( \pi_e + tv \in L_e \) for all \( t \) small enough (we say that \( v \) is in tangent space of \( L_e \) at \( \pi_e \)).

\( \beta \)-i) Check that local minima are critical points.

\( \beta \)-ii) Prove using basic linear algebra that \( \pi_e \in L_e \) with \( \pi_e(i) > 0 \) is a critical point if and only if there are two real numbers \( \alpha, \beta \in \mathbb{R} \) (called Lagrange multipliers) such that for all \( i = 1 \ldots k \)

\[
\partial_i I(\pi_e(1), \ldots, \pi_e(k)) + \beta i + \alpha = 0.
\]

The above are called the Euler-Lagrange equations associated with the optimisation of \( I \) on \( L_e \).

\( \beta \)-iii) Prove that the Gibbs distribution

\[
\pi_\beta(i) = \frac{1}{\sum_i e^{-\beta i}} e^{-\beta i}
\]

is the unique solution to the Euler-Lagrange equation in \( L_{e_\beta} \) where \( e_\beta = \sum_i i\pi_\beta(i) \). Nota Bene: The Lagrange multiplier \( \beta \) associated with energy is called the inverse temperature \( \beta = 1/T \) (it can be negative here because the model is unphysical!).

\( \beta \)-iv) Check that \( e_{+\infty} = 1 \) and \( e_{-\infty} = k \) and conclude on the existence of a unique critical point of \( I \) on \( L_e \).

\( \beta \)-v) Check that \( I \) is strictly convex on \( \mathcal{P}(F) \) and smooth on the interior of \( \mathcal{P}(F) \).

\( \beta \)-vi) Prove that if a strictly convex smooth function has a critical point in the interior of a convex set of \( \mathbb{R}^k \), then this point is the unique minimizer (Hint: do the one dimensional case first). Conclude on the \( \pi_e \) in \( L_e \).

We can now study different formulas and verify that \( \pi_e \) is the unique minimizer on \( B \) for \( e \leq k/2 \) (other cases can be treated similarly).

\( \gamma \)-i) Compute \( \frac{d}{d\beta} e_\beta \) and remark it can be written as a strictly positive variance. Deduce that \( \beta \mapsto e_\beta \) is bijective and compute the derivative of its inverse. Denote \( \beta_e \) its inverse.

\( \gamma \)-ii) Compute \( \frac{d}{d\beta} I(\pi_\beta) \) and deduce that \( \frac{d}{d\beta} I(\pi_e) = -\beta e \) where \( \beta_e \) is the unique \( \beta \) such that \( \sum_i i\pi_{\beta_e}(i) = e \).

\( \gamma \)-iii) Conclude on the fact that \( \pi_e \) is the unique minimizer of \( I \) on \( B \).

Additional questions:

- Compute the rate function in Sanov theorem in the case where the \( n \) particles are i.i.d. but with distribution \( \mu_\beta \). Compare it to the case above.
Construct Markov chains having the distributions of this exercise as reversible distributions.

**Problem 1.7.2** (Cramér, **). The goal of this problem is to prove a simpler version of Cramér’s Large Deviations Principle using elementary arguments only.

Let \((Z_n)_{n \geq 1}\) an i.i.d. sequence in \(\mathbb{R}\), we assume without loss of generality and for simplicity that \(\mathbb{E} Z_1 = 0\), and that

\[
\Lambda : \lambda \mapsto \ln \mathbb{E} \left[ e^{\lambda Z_1} \right],
\]

is finite for any \(\lambda \in \mathbb{R}\). We denote the empirical mean \(X_n \overset{\text{def}}{=} \frac{1}{n} \sum_{m=1}^{n} Z_m\).

1. **Up-i)** Use elementary analysis to show that: 1) \(\Lambda\) is smooth \((C^\infty(\mathbb{R}))\), 2) convex, 3) \(\Lambda \geq 0\) with minimum \(\Lambda(0) = 0\).

2. **Up-ii)** The Legendre-Fenchel transform on \(\mathbb{R}\) is defined by \(\Lambda^*(x) \overset{\text{def}}{=} \sup_{\lambda \in \mathbb{R}} (\lambda x - \Lambda(\lambda))\). Prove that for \(x \geq 0\) the Legendre-Fenchel transform of \(\Lambda\) satisfies \(\Lambda^*(x) = \sup_{\lambda \geq 0} (\lambda x - \Lambda(\lambda))\).

3. **Up-iii)** Study the monotony of \(\Lambda^*\) on \(\mathbb{R}^+\).

4. **Up-iv)** Let \(x \geq 0\) be given. Determine the (sharp) upper bounds of the indicator function \(I_{y \geq x}\) by exponential functions of the form \(y \mapsto a e^{by}\). Deduce for each \(\lambda \geq 0\) a (sharp) upper bound of \(\ln \mathbb{P} [X_n \geq x]\) using \(\Lambda(\lambda)\).

5. **Up-v)** Compute an upper bound of \(\limsup_{n \to +\infty} \frac{1}{n} \ln \mathbb{P} [X_n \geq x]\) using iii) and compare it to the one in Cramér’s theorem.

We next consider for each \(\lambda \in \mathbb{R}\) a modified sequence \((Z_n^\lambda)_{n \geq 1}\) of random variables whose distribution is such that:

\[
\mathbb{E} \left[ \varphi \left( Z_1^\lambda, \ldots, Z_n^\lambda \right) \right] \overset{\text{def}}{=} \frac{\mathbb{E} \left[ \varphi \left( Z_1, \ldots, Z_n \right) e^{n\lambda X_n} \right]}{\mathbb{E} \left[ e^{n\lambda X_n} \right]},
\]

for each \(n \geq 1\) and any bounded measurable test function \(\varphi\). We define in the same way \(X_n^\lambda \overset{\text{def}}{=} \frac{1}{n} \sum_{m=1}^{n} Z_m^\lambda\).

1. **Low-i)** Describe simply the law of \((Z_n^\lambda)_{n \geq 1}\).

2. **Low-ii)** Let \(x, \varepsilon, \lambda \geq 0\) be given. Give the (sharp) lower bound of the form

\[
\mathbb{P} [x < X_n < x + \varepsilon] \geq \mathbb{P} \left[ x < X_n^\lambda < x + \varepsilon \right] e^{nA}
\]

where \(A\) depends only on \(\lambda, x, \varepsilon\), and \(\Lambda(\lambda)\).

3. **Low-iii)** Assume for simplicity that \(\mathbb{P}(Z_1 > z) > 0\) for any \(z \in \mathbb{R}\). What is the range of \(\lambda \mapsto \mathbb{E}_\lambda(Z_1)\) for \(\lambda \geq 0\)? Compute a lower bound of \(\liminf_{n \to +\infty} \frac{1}{n} \ln \mathbb{P} [x < X_n]\) by choosing carefully \(\lambda\) in the estimates above. Compare it to the one in Cramér’s theorem.
1.8 Varadhan lemmas

We will state Varadhan lemma as an extension of the large deviation upper bound and lower bound. Later in the course, we will give a more elegant, condensed and general form. But for practical purpose the following is better.

**Lemma 1.8.1** (Upper bound). Let $V$ be continuous and lower bounded. Assume $X_n$ satisfy a LDP with rate function $I$, then for any closed set $C$

$$\limsup_n \frac{1}{n} \ln \mathbb{E} \left( e^{-nV(X_n)} \mathbb{1}_{X_n \in C} \right) \leq - \inf_C (V + I).$$

**Lemma 1.8.2** (Lower bound). Let $V$ be continuous. Assume $X_n$ satisfy a LDP with rate function $I$, then for any open set $O$

$$\liminf_n \frac{1}{n} \ln \mathbb{E} \left( e^{-nV(X_n)} \mathbb{1}_{X_n \in O} \right) \geq - \inf_O (V + I).$$

**Exercise 1.8.3** (**). State and prove Varadhan’s lemmas in the case where $E$ is finite.

**Exercise 1.8.4.** What happens when $X_n$ is constant?

Varadhan’s lemma enables to obtain the following large deviation principle for a large class of measures;

**Corollary 1.8.5.** Let $V$ be continuous and lower bounded and assume $(\mu_n)_{n \geq 1}$ satisfy a LDP with rate function $I$. Then the sequence of probability (sometimes called 'Gibbs’) measures

$$\mu_n^V(dx) \overset{\text{def}}{=} \frac{1}{z_n} e^{-nV(x)} \mu_n(dx), \quad n \geq 1$$

where $z_n \overset{\text{def}}{=} \int_E e^{-nV} d\mu_n$ is the normalization, satisfies a LDP with rate function

$$I + V - \inf_E (I + V)$$

**Exercise 1.8.6.** Prove the above corollary from the Varadhan’s upper / lower bounds.

**Exercise 1.8.7** (**, Curie-Weiss model). Let $F$ be a Polish space and $\pi_0$ be a given probability on $F$. Let $Y_1, \ldots, Y_n$ be $n$ i.i.d random variables (called ‘particles’) with law $\pi_0$.

- Recall Sanov theorem for the empirical distribution

$$\Pi_n \overset{\text{def}}{=} \frac{1}{n} \sum_{m=1}^n \delta_{Y_m}.$$  

We now wish to apply Varadhan’s lemma in order to obtain a LDP in the case where the variables $Y_m$, $m \geq 1$ are no longer independent.

We next consider an interaction potential function

$$U : F^2 \mapsto \mathbb{R}$$
which is i) continuous and bounded, ii) symmetric $U(y, y') = U(y', y)$, and iii) $U(y, y) = 0$ for each $y \in F$ (no ‘self-interaction’). Assume that the variables $(Y_U^1, \ldots, Y_U^n) \in F^n$ (‘interacting particles’) are distributed according to the Gibbs probability measure:

$$
\frac{1}{Z} \exp \left( -\beta \frac{1}{n} \sum_{1 \leq l < m \leq n} U(y_l, y_m) \right) \pi_0(dy_1) \cdots \pi_0(dy_n)
$$

where in the above $Z$ is the normalization (so that the above is indeed a probability):

$$
Z \overset{\text{def}}{=} \int_{F^n} \exp \left( -\beta \frac{1}{n} \sum_{1 \leq l < m \leq n} U(y_l, y_m) \right) \pi_0(dy_1) \cdots \pi_0(dy_n)
$$

• Denote by $\mu_n$ the probability measure on $\mathcal{P}(F)$ of given by the law of $\Pi_n$. In which space belongs $\mu_n$ if $F = \{1, \ldots, k\}$ is finite? And in other cases?

• Denote

$$
\Pi_n^U \overset{\text{def}}{=} \frac{1}{n} \sum_{m=1}^n \delta_{Y_m^U}
$$

the empirical distribution of $(Y_U^1, \ldots, Y_U^n)$. Prove that the law of $\Pi_n^U$ is given by:

$$
d\mu_n^V(\pi) = \frac{1}{Z} e^{-n \beta V(\pi)} d\mu_n(\pi)
$$

for the measure valued function $V(\pi) \overset{\text{def}}{=} \frac{1}{2} \int_{F^2} U(y, y') \pi(dy) \pi(dy')$.

• Prove that $\Pi_n^U$ satisfies a LDP with good rate function

$$
I^V(\pi) \overset{\text{def}}{=} \beta V(\pi) + \text{Ent}(\pi|\pi_0) - \inf_{\pi} (\beta V(\pi) + \text{Ent}(\pi|\pi_0))
$$

Interpret in terms of competition between energy and entropy.

We will now study the Curie-Weiss model. We consider the setting of the previous exercise. Let $F = \{-1, +1\}$,

$$
U(y, y') = -y \times y' + 1,
$$

and $\pi_0(dy)$ is the uniform distribution.

• Check that $\mathcal{P}(F)$ is a one dimensional interval that can be parametrized by $p = \pi(1)$ if $\pi \in \mathcal{P}(F)$.

• Prove that

$$
I^V(p) = -\beta \frac{1}{2} (2p - 1)^2 + p \ln p + (1 - p) \ln(1 - p) + \ln 2
$$

Study the local minima depending on the values of $\beta$ (Answer: there is a phase transition: there is a unique minimum $1/2$ if $\beta \leq 1$, otherwise there are two, $p_*$ and $1 - p_*$).

• Construct a Markov chain having the Gibbs measure as an invariant distribution and interpret.
1.9 Contraction Principle

If a LDP is available for a sequence of random variables \((X_n)_{n \geq 1}\), one can obtain a LDP for any continuous image of the latter.

**Proposition 1.9.1** (Contraction Principle). Assume \((X_n)_{n \geq 1}\) satisfy a LDP in \(E\) with good rate function \(I\), and let

\[ f : E \to G \]

be a continuous function in topological \(G\). Then \((f(X_n))_{n \geq 1}\) satisfy a LDP in \(G\) with good rate function

\[ I_f(z) \overset{\text{def}}{=} \inf_{x \in E : f(x) = z} I(x). \]

**Exercise 1.9.2** (*). Show that \(I_f\) is a good semi-continuous function, and then prove the Contraction Principle. Interpret in terms of 'cost of the least unlikely states'.

**Exercise 1.9.3** (**, a variant of Cramérs Theorem). Let \(Z_n = \frac{1}{n} \sum_{m=1}^{n} \varphi(Y_m)\) were \(\varphi(Y_m)\) are i.i.d. bounded random variables. Using Sanov theorem and the Contraction Principle, prove a LDP for \(Z_n\) and compute the associated rate function.

The convex dual formulation of \(I_f\) will be detailed later on.

**Exercise 1.9.4**. Give a counterexample showing that if \(I\) is not good, then \(I_f\) may not be lower semi-continuous (Hint: \(I_f = 0\) on an open set, \(+\infty\) else).