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

Mathias Rousset

2019-2020

Documents to be downloaded

PLEASE download refs at `http://people.irisa.fr/Mathias.Rousset/****.zip` where **** is the code given in class.

Main bibliography

Large Deviations

- Dembo and Zeitouni [1998]: classic, main reference of the course. General, complete, not easy yet readable book.
- Touchette [2009]: readable, formal/expository introduction with links to statistical mechanics.
- Rassoul-Agha and Seppäläinen [2015]: the first part is similar to the content of this course. Advanced examples of Ising-like models.
- Freidlin et al. [2012]: the classic reference for small noise large deviations of SDEs.
- Feng and Kurtz [2015]: good for the maths basics related to the rigorous general treament in a Markovian setting.

General references

- Bogachev [2007]: exhaustive on measure theory.
- Kechris [1994, 2012]: short and long about descriptive set theory and advanced results on Polish spaces and co.
- Brézis [1987], Brezis [2010]: excellent master course on functional analysis.
- Billingsley [2013]: most classical ref for convergence of probability distributions.

Interesting blogs

- Djalil Chafaï.
- George Lowther.

Usual notations

- $C_b(E)$: continuous and bounded functions on topological E.
- $M_b(E)$: measurable and bounded functions on measurable E.
- $\mathcal{P}(E)$: probability measures on measurable E.
- $\mathcal{M}_{\mathbb{R}}(E)$: real valued (finite, signed) measures on measurable E.

Exercises: Stars ($**$ or $*$) indicates *importance*.

1 Introduction

1.1 General remarks

The theory of large deviations may be summarized as the "**asymptotic theory of rare events**". "Asymptotic" will refer to a sequence of probability distributions

$$(\mu_n = \text{Law}\,(X_n))_{n\geq 1}$$

defined on a given (nice enough) measurable space E; for presentation prupose, the sequence $(\mu_n)_{n\geq 1}$ is assumed to be the distribution of a sequence of random variables $(X_n)_{n\geq 1}$.

In this context, a "rare event" may be simply defined as an event A with vanishing probability

$$\mathbb{P}(X_n \in A) \xrightarrow{n\to+\infty} 0. \tag{1}$$

Two classical examples:

- *(Small noise) X_n is a small random perturbation of a deterministic state $x_\infty \in E$.*
- *(Large population) $X_n = S_n$ is the average of n i.i.d. variables.*

It turns out that quite generally, if one tries to look at the probability of rare events using a **logarithmic scale** (a rough picture), then the associated *rate of vanishing* can be described by a fairly *explicit minimization problem*. Formally:

$$\log \mathbb{P}(X_n \in A) \underset{n\to+\infty}{\sim} -c_n \inf_{z\in A} I(z), \tag{2}$$

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

$$\lim_{n\to+\infty} c_n = +\infty, \tag{3}$$

and

$$I : S \to [0, +\infty] \tag{4}$$

is a 'cost' function called the *rate function*. The property (2), which is a simplified version of the so-called *large deviation principle (LDP)*, has the following key interpretation:

i) The probability of an asymptotically rare event $\{X_n \in A\}$ is determined by the *least rare* outcomes in A, that is those who minimizes the rate function I.

ii) A probability conditioned by the rare event $\{X_n \in A\}$ is asymptotically concentrated on the *least rare* outcomes in A.

The specific speed c_n for $n \geq 1$ has to be chosen appropriately, up to a > 0 multiplicative constant; typically in a way such that the rate function I is *non-trivial* that is

$$\exists x \in F,\, 0 < I(z) < +\infty$$

it can be arbitrarily increased. In these notes, we will simplify notation and always consider the case

$$c_n = n, \quad n \geq 1,$$

which is the only non-trivial speed for i.i.d. samples of size n.

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

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

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

Prove that

$$\frac{1}{n} \ln \sum_j a_n(j) \to \max_j l(j)$$

Exercise 1.1.2 $(**)$**.** Show that the LDP holds in the case where E is finite if and only if $\frac{1}{n} \ln \mathbb{P}(X_n = i) \sim -I(i)$ for each $i \in E$ for some function $I : E \to [0, +\infty]$. 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 1.1.3 $(**)$**.** Make sense of *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. Interpret.

Exercise 1.1.4 $(**)$**.** 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 $A = \left\{z \in \mathbb{R}^d | z_1 > 1\right\}$. Interpret geometrically.

Exercise 1.1.5. Prove that if (2) holds with speed c_n, then it still holds for any $c'_n \sim_n \text{cte} c_n$ woth cte > 0. Prove that extracted subsquences of $(\mu_n)_{n\geq 1}$ satisfy (2) with arbitrarily fast speed.

1.2 The need for a topology and a precise definition

The simple form of LDP stated in (2) 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 1.1.4, prove that (2) is false for any countable set 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 E.

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

Portmanteau theorem:

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

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

or

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

It can be checked that if the limit X_∞ 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 $\mathring{A}$ 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. *A $[-\infty, +\infty]$-valued function f on a topological space is said to be lower semi-continuous if $\{x | f(x) \leq a\}$ is closed for any $a \in \mathbb{R}$.*

Definition 1.2.6. *Let $(X_n)_{n\geq 1}$ a sequence of random variables on 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

$$-\inf_{x\in\mathring{A}} I(x) \leq \liminf_n \frac{1}{n} \log \mathbb{P}\{X_n \in A\} \leq \limsup_n \frac{1}{n} \log \mathbb{P}\{X_n \in A\} \leq -\inf_{x\in\overline{A}} I(x). \tag{5}$$

Then we say that the sequence of distribution $\mu_n = \mathrm{Law}(X_n)$, $n \geq 1$ satisfies a LDP with speed n and rate function I.

Another recurrent element in the 'jargon' of LDP

Definition 1.2.7 (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 1.2.8 (Discrete case)**.** Simplify the definition of the LDP in the case where E is discrete. Describe simply good functions. Assume $E = \mathbb{N}_*$ Check that if I is good then

$$\limsup_{x \to +\infty} \limsup_{n \to +\infty} \frac{1}{n} \ln \mathbb{P}(X_n \geq x) = -\infty.$$

Show that the LDP is then equivalent to the convergence of $\frac{1}{n} \ln \mathbb{P}(X_n = x)$, $x \in E$.

Exercise 1.2.9 $(*)$**.** What means goodness on locally compact spaces ? Check (on metric spaces) that good lower semi-continuous functions always have a minimum on each closed set.

Exercise 1.2.10 $(*)$**.** Prove that the constant distributions $\mu_n = \mu_0$, $\forall n \geq 1$ satisfy a LDP and descibe 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 mass. Compute the support of usual distributions in $\mathbb{R}^d$.

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 $(**)$**.** Show that if $E_0 \subset E$ 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 1.2.13 $(**$, Lower semi-continuity$)$**.** Describe all the lower semi-continuous functions on $\mathbb{R}$ that are continuous on $\mathbb{R} \setminus \{0\}$.

Exercise 1.2.14 $(*)$**.** Prove that if I is a rate function then $\inf I = 0$.

Exercise 1.2.15. Prove that if one relaxes the condition of lower semi-continuity of I, then I may not be unique.

Exercise 1.2.16 $(*)$**.** 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 μ_n converges in distribution towards δ_{x_0}.

Exercise 1.2.17. 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_ε such that

$$\liminf_n \mathbb{P}\left(X_n \in K_\varepsilon\right) \geq 1 - \varepsilon.$$

Exercise 1.2.18 $(**)$**.** Prove rigorously the LDP in the case where X_n is distributed according to a Gaussian $\mathcal{N}(0, \frac{1}{n})$ (Hint: prove first that $\frac{1}{n}\mathbb{P}(X_n \in A)$ converges towards the Lebesgue essential infimum of $x^2/2$ over A).

This last exercise is a particular example of a large class of LDPs that follows from Varadhan's lemma.

Lemma 1.2.19 (∗∗)**.** *Let π denote a probability on topological E (say, $E = \mathbb{R}^d$). Let $V : E \to \mathbb{R}$ be continuous and bounded from below. Then the sequence of probability*

$$\mu_n(dx) \stackrel{\text{def}}{=} \frac{\mathrm{e}^{-nV(x)}}{\int_E \mathrm{e}^{-nV(x)}\pi(dx)}\pi(dx)$$

satisfies a LDP with rate function:

$$I(x) = \begin{cases} V(x) - \inf_{\mathrm{supp}(\pi)} V & x \in \mathrm{supp}(\pi), \\ +\infty \qquad \text{else.} \end{cases}$$

This lemma will proven as a consequence of Varadhan's lemma below.

Exercise: Prove in the lemme above that for a general bounded from below V, $\frac{1}{n}\ln \mathbb{P}(X_n \in A)$ converges towrds $-\pi \cdot \mathrm{essinf}_A V + \pi \cdot \mathrm{essinf}_E V$.

1.3 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,

$$\mathrm{Law}\left(X_n | X_n \in B\right), n \,\text{large}.$$

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.

Lemma 1.3.1. *If $\inf_{\overline{A\cap B}} I > \inf_{\mathring{B}} I$ then $\mathbb{P}[X_n \in A | X_n \in B]$ tends exponentially fast to 0 with n.*

Exercise 1.3.2 (∗∗)**.** Draw a picture in the Gaussian case for which $\inf_{\mathring{B}} I = \inf_{\overline{B}} I$. Prove the above lemma directly from the LDP.

The Gibbs conditioning principle then states that:

Proposition 1.3.3 (Gibbs Conditioning Principle)**.** *Let $X_n \in E$, $n \geq 1$ satisfies a LDP with rate function I. Let $B \subset E$ measurable such that $\mathbb{P}\left(X_n \in B\right) > 0$ for each n. Assume there is a unique minimizer x_* of I in $\overline{B}$ satisfying:*

- $\inf_{\mathring{B}} I = I(x_*)$,
- *The minimum x_* in $\overline{B}$ is attained locally only*

$$\inf_{O^c_{x_*}\cap \overline{B}} I > I(x_*) \qquad \forall\, \textit{open}\, O_{x_*} \ni x_*.$$

Then the conditional distribution $\mathrm{Law}\left(X_n | X_n \in B\right)$ *converges in law towards δ_{x_*} ($X_n^B \to x_*$ in probability where* $\mathrm{Law}(X_n^B) = \mathrm{Law}\left(X_n | X_n \in B\right)$*).*

Proof. Check using Portmanteau theorem that convergence in law towards a deterministic x_* is equivalent to convergence to 0 of the probability of being outside any neighborhood of x_*.

Then apply the LDP to $\mathbb{P}\left(X_n \in O^c_{x_*} \cap B\right)$ and $\mathbb{P}\left(X_n \in B\right)$ to conlude. □

1.4 Remarks on the state space E and the large deviation topology

This section is *psychological preparation to measure-valued* LDPs.

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. *The measurable sets of the state space E are given by the Borel sets (that is the σ-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 σ-algebra of measurable sets, called the cylindrical σ-algebra, and defined as the smallest one making the maps $\mu \mapsto \int \phi d\mu$ measurable for each measurable bounded test function ϕ. 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 (5) *may be chosen appropriately* depending on the problem at hand. In particular, one interested in looking for LDP in the *finest possible* topology.

Exercise 1.4.1. Check that if a LDP holds true for a given topology, then it is also true for any coarser topology.

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

Assumption. *The topology considered in LDPs are supposed to be at least regular: any point x and any closed set C can be separated by neighborhood, 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 locally convex topologies (and traces of such) are regular (in fact completely regular) – they are the only reasonable topologies considered on topological vector spaces. See below.

Exercise 1.4.2. Check that if the topology is regular, then the rate function is unique.

1.5 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

$$E \stackrel{\text{def}}{=} \mathcal{P}(F),$$

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

Theorem 1.5.1 (Sanov)**.** *Let $(Y_m)_{m\geq 1}$ denote a sequence of of i.i.d. variables in Polish F with* $\mathrm{Law}(Y_m) = \pi_0$*. The sequence of empirical distribution defined for $n \geq 1$ by*

$$X_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{m=1}^{n} \delta_{Y_m} \in E$$

satisfy a LDP on $\mathcal{P}(F)$ endowed with convergence in distribution with good rate funcion

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

where Ent *is the* relative entropy *(a.k.a.* Kullback-Leibler divergence*) between π and π_0.*

The relative entropy is defined by

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

Exercise 1.5.2. Check that $\mathrm{Ent}(\pi|\pi_0)$ is positive and vanishes if and only if $\pi = \pi_0$.

Remark 1.5.3 (On the topology)**.** In fact, Sanov can be strengthed to a stronger weak topology called the τ-topology, also denoted the $\sigma(\mathcal{M}_{\mathbb{R}}, M_b)$-topology or the M_b-intial topology. This is the 'weak' i.e. the coarsest topology making all the maps $\mu \mapsto \int_F \psi d\mu$ continuous, where ψ is *only measurable* (and not continuous) and bounded. This is much stronger than convergence in distribution.

Sanov theorem is remarkable because it relates independence with (relative) entropy which is derived as a secondary concept. Nota Bene: 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
$$\frac{1}{n}\ln \mathbb{P}[\text{empirical density} = \text{given density}]$$
(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 theorists.
- Convex duality. This is the classical road that was developed independently in statistics, and it will be the one we follow.

Exercise 1.5.4 (∗∗). Assume $F = \{1, 2, \ldots, k\} \subset \mathbb{N}$ is a finite state space describing the possible energy of a particle. 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 $e \in F$.

- 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:
$$\frac{1}{n} \sum_{m=1}^{n} Y_m = e. \tag{6}$$

For technical reasons, we condition the energy of the above particle system in the interval $[e - \delta e, e]$ with $\delta e > 0$:

$$\frac{1}{n} \sum_{m=1}^{n} Y_m \in [e - \delta e, e]. \tag{7}$$

- Interpret the uniform distribution conditioned by (7) as an average on $[e-\delta e, e]$ of the uniform distribution conditioned by (6) for various mean energy e'.

- 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 $I = \mathrm{Ent}\,(\,.\,|\mathrm{Unif})$ the rate function. Prove that convergence in distribution in $\mathcal{P}(F)$ is given by the usual trace topology on $\mathbb{R}^k$, that I is continuous on $\mathcal{P}(F)$, and that $\mathcal{P}(F)$ is compact.

- Consider the set of probabilities:
$$B \stackrel{\mathrm{def}}{=} \left\{ \pi \in \mathcal{P}(F) | \sum_{i=1}^{k} i\pi(i) \in [e - \delta e, e] \right\}.$$
Check that B is closed as a subset of $\mathbb{R}^k$, describe its interior, and that the closure of the interior $\mathring{B}$ is again B. Check that the event $\Pi_n \in B$ is equivalent to the event (7).

- Assume that there is a unique minimizer π_e of I on B. Prove using a compacity argument that $I(\pi) > I(\pi_e)$ for all $\pi \in B$ outside any neighborhood of π_e. Check also using the continuity of I that $\inf_{\mathring{B}} I = I(\pi_e)$. Conclude using the Gibbs conditioning principle that Π_n conditioned by B converges in distrbution towards π_e.

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

$$L_e \stackrel{\mathrm{def}}{=} \left\{ \pi \in \mathcal{P}(F) | \sum_{i=1}^{k} i\pi(i) = e \right\}.$$

π_e is called a critical point of I on L_e iff

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

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 π_e.

- 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 $i = 1 \dots k$

  $$\partial_{\pi_i} I(\pi_e) + \beta i + \alpha = 0.$$

  The above are called the Euler-Lagrange equations.

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

- 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 minimzer (Hint: do the one dimensional case first).

- Prove that the Gibbs distribution

  $$\pi_\beta(i) = \frac{1}{\sum_i \mathrm{e}^{-\beta i}} \mathrm{e}^{-\beta i}$$

  is solution to the Euler-Lagrange equation in L_{e_β} where $e_\beta = \sum_i i\pi_\beta(i)$. Nota Bene: The Lagrange multiplier β associated with energy is called the inverse temperature $\beta = 1/T$.

- Check that $e_{+\infty} = 1$ and $e_{-\infty} = k$ and $e_0 = k/2$ and conclude on the existence of a unique minimzer of I on L_e. Check that it is positive if $e > k/2$.

We can now study different formulas and verify that π_e is the unique minimizer on B.

- 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 β_e its inverse.

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

- Conclude on the fact that π_e is the unique minimizer of I on B.

Additional problems:

- Compute the rate function in Sanov theorem in the case where the n particles are i.i.d. but with distribution μ_β. Compare it to the case above.

- Construct Markov chains having the distributions of this exercise as reversible distributions.

1.6 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 1.6.1 (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*

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

Lemma 1.6.2 (Lower bound)**.** *Let V be continuous. Assume X_n satisfy a LDP with rate function I, then for any open set O*

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

Exercise 1.6.3 (∗∗)**.** State and prove Varadhan's lemmas in the case where E is finite.

Exercise 1.6.4. 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 1.6.5. *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 proability (sometimes called 'Gibbs') measures*

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

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

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

Exercise 1.6.6. Prove the above corollary from the Varadhan's upper / lower bounds.

Exercise 1.6.7 (∗∗, Curie-Weiss model)**.** Let F be a Polish space and π_0 be a given probability on F. Let $Y_1, \ldots, Y_n$ be n i.i.d random variables (called 'particles') with law π_0.

- Recall Sanov theorem for the empirical distribution

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

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 $(Y_1^U, \ldots, Y_n^U) \in F^n$ ('interacting particles') are distributed according to the Gibbs probability measure:

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

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

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

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

- Denote

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

  the empirical distribution of $(Y_1^U, \ldots, Y_n^U)$. Prove that the law of Π_n^U is given by:

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

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

- Prove that Π_n^U satisfies a LDP with good rate function

$$I^V(\pi) \stackrel{\text{def}}{=} \beta V(\pi) + \mathrm{Ent}(\pi | \pi_0) = \frac{1}{2} \int_{F^2} U(y,y') \pi(dy) \pi(dy')$$

  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 β (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.

1.7 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 1.7.1 (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

$$I_f(z) \stackrel{\text{def}}{=} \inf_{x\in E:\, f(x)=z} I(x).$$

Exercise 1.7.2 $(*)$**.** 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 1.7.3 $(*$,Cramers Theorem$)$**.** Let $Z_n = \frac{1}{n}\sum_{m=1}^{n} \varphi(Y_m)$ were $\phi(Y_m)$ 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.

A dual formulation of I_f will be detailed later on.

Exercise 1.7.4. 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).

2 Background material: measure theory and functional analysis

2.1 Measure theory

A measurable space is the data of a set E and a collection of sets $\mathcal{E}$, called the its *measurable sets*, that must form a σ-algebra: it is stable by taking complements, and by taking any *countable* union and/or intersection.

For any collection of sets, one can consider the *smallest* σ-algebra containing this collection – the collection is said to generate the σ-algebra.

Exercise 2.1.1. Check that the smallest σ-algebra containing a collection of sets exists and is unique.

σ-algebra are required in order to restrict the sets on which one can define properly measures μ satisfying the usual axiom $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$ as well as its countable generalizations.

The famous monotone class theorems gives various conditions for a class of subsets to generate a given σ-algebra. A functional version is particularly useful and easy to remember. Recall that integration theory enables to construct

$$\langle \mu, \psi \rangle \stackrel{\text{def}}{=} \int_F \psi d\mu$$

for any bounded mesurable ψ and any measure μ, and that if $f_n \to f$ pointwise and monotonically for f_n measurable and bounded from above then g is measurable and $\int f_n d\mu \to \int f d\mu$ (monotone convergence theorem). A functional variant of the monotone class theorem states that in fact all measurable functions can be approximated pointwise and monotonically by a generating algebra of functions.

Theorem 2.1.2. *Let $H_0 \subset H$ where H is a vector space of real valued bounded functions. Assume:*

- *H_0 is stable by product of functions.*
- *H contains constant functions and is stable by increasing pointwise limits.*

Then H contains all the bounded functions measurable with respect to $\sigma(H_0)$.

Exercise 2.1.3. Check that the monotone class theorem enables to recover that any positive measurable function on $\mathbb{R}$ is the pointwise increasing limit of smooth functions with compact support.

In what follows, we will denote $\mathcal{P}(F)$ the space of probability measures, and more generally $\mathcal{M}_{\mathbb{R}}(F)$ the vector space of all *real (or signed)* measures, that is measures of the form $a\mu - b\nu$ for $\mu, \nu \in \mathcal{P}(F)$ – the vector space generated by probabilities. The Hahn-Jordan decomposition theorem states that the latter decomposition in positive and negative parts can be uniquely chosen in a disjoint way: for any $\mu \in \mathcal{M}_{\mathbb{R}}(F)$, then there exists a unique pair of non-negative finite measures μ^+, μ^- such that $\mu = \mu^+ - \mu^-$ and $\mu^+(A) > 0 \Leftrightarrow \mu^-(A) = 0$ for each measurable A.

The space $\mathcal{M}_{\mathbb{R}}(F)$ can be endowed with the operator norm associated with the supremum norm of measurable test functions, and called the *total variation norm*:

$$\|\mu\|_{\text{tv}} \stackrel{\text{def}}{=} \sup_{\phi \in M_b(F)} \frac{\int \phi d\mu}{\|\phi\|_\infty} = \mu^+(\mathbb{1}_F) + \mu^-(\mathbb{1}_F)$$

Exercise 2.1.4. For $F = \mathbb{R}$, prove that the total variation norm is still the operator norm associated with continuous bounded functions. Check that convergence in distribution can be seen as a weak topology on the space of finite measures for the duality given by integration.

2.2 Topology

A topology is collection of subsets of E called *open sets* that are stable by *any union* and any *finite intersection*. Complements of open sets are called *c*losed sets.

For any collection of sets, one can consider the smallest topology containing this collection. The collection is called a *subbase* and is said to generate the topology. In $\mathbb{R}^d$ for instance, a subbase is given by all open affine half-spaces.

a function is continuous iff the pull-back of open (or closed) sets is again open.

Finite intersections of a subbase defines a *base* (of open sets) for the topology, which is an intuitive concept. N_x is a neighborhood of a point x if it contains an element of the base O_x that contains x: $x \in O_x \subset N_x$. Open sets are then exactly sets that are neighborhoods of each of their points.

Moreover a function is continuous at x if for any O_x in the base (containing x, as small as one wishes), there is a $O_{f(x)}$ (containing $f(x)$, as small as necessary) in a base of the arrival topological space such that the pull-back is included in the initial O_x: $f^{-1}(O_{f(x)}) \subset O_x$.

Exercise 2.2.1. Check that there is only one topology generated by a subbase of a topology, and then that N_x is a neighborhood of x if and only if there is a *finite intersection* of subsets in the subbase containing x and contained in N.

The σ-algebra generated by open sets is called the Borel σ-algebra.

Exercise 2.2.2. Check that if a topology τ is generated by a *countable* subbase $\mathcal{B}$ then $\sigma(\tau) = \sigma(\mathcal{B})$.

Exercise 2.2.3. Recall the definition of: i) product topology, ii) compacity (in general and sequential in metric spaces). Give an explicit metric for $F^{\mathbb{N}}$ is F is metric. Recall Tychonoff theorem.

Comparison of topologies If a topology τ_s contains more open sets (or equivalently more closed sets) than a topology τ_w, then τ_s may be called 'stronger', 'finer', or 'richer'. Inversely τ_w may be called 'weaker', 'coarser', or 'poorer'.

As a consequence the weaker the topology is, the more difficult it is for a set to be open/closed and for a real valued function to be continuous. On the contrary, the weaker the topology, the easier it is for a set to be compact.

2.3 Polish topologies

A Polish topology is by definition *metrizable with a complete metric and separable* (that is it has a dense countable subset). Polish spaces include most usual separable topological spaces such as:

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

Non-separable metric spaces include for instance bounded measurable or even continuous functions on $\mathbb{R}$, or measures endowed with total variation norm.

Polish topologies are quite robust. The following are again a Polish spaces if E is one:

- Open or closed subsets of E.
- Sequences in E, endowed with convergence on finite sub-sequences, that is $E^{\mathbb{N}}$.
- $\mathcal{P}(E)$ endowed with convergence in distribution (see below).

Exercise 2.3.1. Construct explicit complete metric making i) $]0,1]$; ii) $\mathbb{R}^{\mathbb{N}}$ Polish. Propose a metric for the trace topology of an open subset of a Polish space which is complete and separable.

As in any metric (or more generally 'normal') space, Urysohn's lemma easily applies: for any two disjoint closed sets F and G there exists a continuous function (a 'cut-off' function) which is identically 0 on F and 1 on G, and thus indicator of open/closed sets are pointwise increasing limits of continuous functions.

Exercise 2.3.2. Exhibit the cut-off function proving Urysohn lemma. Check that in metric spaces the indicator function of any closed set is the bounded decreasing pointwise limit of continuous functions.

Using the monotone class theorem above, this implies (Exercise):

Lemma 2.3.3. *On a metric space, monotone pointwise limits of continuous functions generates all Borel measurable functions.*

In the same spirit

Lemma 2.3.4. *On a metric space all probability measure μ is regular: for any Borel set B and $\varepsilon > 0$ there is $O \subset B \subset C$ with C closed and O open and $\mu(O \setminus C) < \varepsilon$.*

Exercise 2.3.5. Prove the above theorem. Hint: check it when C is closed, and then check that Borel sets verifying the property is a σ-algebra.

Exercise 2.3.6. Prove that on metric spaces, the total variation norm is also the operator norm associated with *continuous* test functions.

Polish spaces also have nice properties related to compacity and covering with small balls.

Lemma 2.3.7. *A set in a complete metric space is said be* totally bounded *if for any $\varepsilon > 0$, it can be covered by a finite number of balls of size ε. This is equivalent to pre-compactness (Exe: why ?).*

Obviously if the space is moreover separable, then there is a countable covering of it by such small balls.

Lemma 2.3.8. *Any probability measure μ on a Polish space is tight: the supremum of $\mu(K)$ over compact K is one.*

Exercise 2.3.9. Proof. Hint: use a countable covering of the space by small balls of size $1/k$ for each k, and construct a totally bounded set which contains almost all the mass.

2.4 Remarks on standard Borel spaces

If E is a Polish space, and one 'forgets' the original topology of E, one obtains a measurable space sometimes called a *standard Borel* measurable space.

Most state spaces used in practice in probability theory are standard Borel. Descriptive Set Theory studies various remarkable (and simplifying properties) of standard Borel spaces and of their Borel subsets.

A first important class of results show that, up to refining the topology, any Borel set (respectively any measurable function) can be made open (respectively continous).

Theorem 2.4.1 (Refining Polish topologies)**.** *Let E be a Polish space.*

i) Let $B \subset E$ a Borel subset. There is a finer Polish topology with the same Borel sets making B open.

ii) Let $\phi : E \to \mathbb{R}$ be bounded and measurable. There is a finer Polish topology with the same Borel sets making ϕ continuous.

In the present notes, we will only use *ii*) in the strengthening of Sanov theorem.

A second class results (quite difficult to prove) shows that there is, up to isomorphism, only one standard Borel space, and only one standard probability space (a standard Borel space with an atomless probability distribution).

Theorem 2.4.2 (Isomorphisms theorems)**.** *i) (Kuratowski) Two standard Borel spaces E_1 and E_2 with the same cardinality are isomorphic as measurable spaces: there is a measurable bijection $F : E_1 \to E_2$ with measurable inverse (only possible cases: a finite set, a countable set, or $\mathbb{R}$).*

ii) Two measured standard Borel spaces E_1 and E_2 with respective atomless probability distributions μ_1, μ_2 are isomorphic as measured space: there is a measurable map $F : E_1 \to E_2$ such that the measure image by F of μ_1 is μ_2. Moreover, F can be chosen one-to-one with measurable inverse, up to sets of measure zero.

i) means that the structure of usual (standard Borel) measurable state spaces is unique. ii) implies that any probability on a Polish state space can be represented with a random variable X in the form:

$$X = F(U)$$

where U is a uniform distribution on $]0, 1[$.

It turns out that many usual tools and intuitions in probability only apply nicely when F is a standard Borel measurable space. This is a customary assumption that is almost always made on state spaces.

2.5 Convergence in distribution

The total variation norm generates a non-separable, very strong topology on measures (and probabilities). Much more useful in practice is the convergence in distribution, but the latter is dependent on a *choice of topolgy on the state space* F. Let us recall the Portmanteau theorem.

Theorem 2.5.1. *If a metric and its Borel sets are given on F then we say that $\mu_n \to \mu$ in distribution in $\mathcal{P}(F)$ if one of the following equivalent conditions hold.*

i) $\int_F \varphi\mu_n \to \int_F \varphi\mu$ *for any φ continuous and bounded.*

ii) $\limsup \mu_n(F) \leq \mu(F)$ *for any closed set F.*

iii) $\liminf \mu_n(O) \geq \mu(O)$ *for any open set O.*

iv) $\lim \mu_n(A) = \mu(A)$ *for any Borel set A with $\mu(\partial A) = 0$, $\partial A \stackrel{\text{def}}{=} \overline{A} \setminus \mathring{A}$.*

Exercise 2.5.2. Sketch the proof of Portmanteau theorem. Hints: define the open set

$$F^\varepsilon = \{x | d(x, F) < \varepsilon\}. \tag{8}$$

i) $\Rightarrow$ ii) follows from constructing a continuous function that is 1 on F and 0 outside F^ε. Check ii) $\Rightarrow$ iii) $\Rightarrow$ iv). Write integral of a function using its level sets to sketch iv) $\Rightarrow$ i).

It turns out that convergence in distribution endows $\mathcal{P}(F)$ with a Polish topology if F is itself Polish. The first complete and separable metric usually introduced is the so-called Levy-Prohorov metric $\text{dist}_P(\mu, \nu)$, which depends on a metric d defined on F, and is defined as follows.

Definition 2.5.3 (Levy-Prohorov distance)**.** *Let $\mu, \nu \in \mathcal{P}(F)$. The Levy-Prohorov metric $\text{dist}_P(\mu, \nu)$ is defined as the infimum over all the $\varepsilon > 0$ such that*

$$\mu(A) \leq \nu(A^\varepsilon) + \varepsilon \qquad \forall A \text{ Borel},$$

where A^ε is defined by (8).

Theorem 2.5.4. *If F is Polish, then the Levy-Prohorov metric $\text{dist}_P(\mu, \nu)$ is separable, complete, and metrizes convergence in distribution.*

Exercise 2.5.5.

- Check that the Levy-Prohorov metric has a symmetric expression and is indeed a distance.
- Check that convergence with the latter implies convergence in distribution.
- Check that for all $\varepsilon, \delta > 0$ and all μ we can cover a set of mass $1-\varepsilon$ with finitely many δ-small balls, hence partitioning it with finitely many δ-small subsets.
- Sketch the proof of: convergence in distribution implies convergence for the Levy-Prohorov metric. (Hint: take $\delta = \varepsilon$ and pick a n_ε such that for $n \geq n_\varepsilon$, $\mu_n(A^\varepsilon)$ is below its limsup up to ε for all A generated by the finite partition).
- Sketch a possible proof of separability of the Levy-Prohorov metric.
- Admit that the following property implies pre-compactness of a set of probabilities for the Levy-Prohorov metric: for all $\varepsilon, \delta > 0$, there is a finite union of δ-small balls that have a mass uniformly greater than $1-\varepsilon$ (N.B.: this is Prohorov theorem see below). Sketch a proof of the completeness of the Prohorov metric.

Exercise 2.5.6. Research the Ky-Fan metric and the coupling interpretation of the Levy-Prohorov metric.

Exercise 2.5.7. Let d be a metric for Polish F. Check that $d \wedge 1$ is again a metric and check that in that case L^1 convergence of random variables is equivalent to convergence in probability. Research the so-called Wasserstein metrics on $\mathcal{P}(F)$ (Villani [2008]). It is a remarkable fact that the latter are complete metrics for convergence in distribution on $\mathcal{P}(F)$ when F is Polish. Do some research on Kantorovitch duality in the W_1 case and check that the dual norm of Lipschitz test functions defined up to a constant metrizes convergence in distribution. Compare to Levy-Prohorov.

2.6 The Borel sets of $\mathcal{P}(F)$

Note that if one 'forgets' the topology of F, $\mathcal{P}(F)$ still possesses a natural σ-algebra of measurable sets, called the cylindrical σ-algebra, defined as the smallest one making the maps $\mu \mapsto \int \phi d\mu$ measurable for each measurable bounded test function ϕ.

Lemma 2.6.1. *Let F be a Polish space. The Borel sets of $\mathcal{P}(F)$ endowed with convergence in distribution coincide with the cylindrical σ-algebra.*

Exercise 2.6.2. Prove the above lemma. For Borel $\subset$ cylindrical, use the fact that the topology is Polish and thus countably generated. For cylindrical $\subset$ Borel, consider the following space of functions

$$H_0 \stackrel{\text{def}}{=} \left\{ \phi \,\text{mesurable} | \mu \mapsto \int \phi d\mu \,\text{Borel measurable} \right\}.$$

Check that H_0 contains continuous function and apply the monotone class theorem. Alternatively, you can prove that the map $\mu \mapsto \int \phi d\mu$ for $\phi \in M_b(F)$ is the pointwise limit of $\mu \mapsto \int \phi_n d\mu$ with ϕ_n continuous, and thus is Borel measurable.

The latter result means that when F is Polish, we do not have to worry about the interplay between measurability and topology on the space $\mathcal{P}(F)$: i) $\mathcal{P}(F)$ inherits from F a natural Polish topology, ii) the Borel sets of $\mathcal{P}(F)$ identifies with a canonical cylindrical σ-algebra (which is purely measure theoretical), and are independent of the specific Polish topology on F. There is no possible ambiguity when defining the measurable sets of $\mathcal{P}(F)$.

Exercise 2.6.3. Check that $(z_1, \dots, z_n) \mapsto \sum_{m=1}^{n} \delta_{z_m}$ is continuous hence measurable.

Other, stronger, topologies can be considered on $\mathcal{P}(F)$. First, the τ-topology:

Definition 2.6.4. *The τ-topology on $\mathcal{P}(F)$ is defined as the smallest topolgy making all maps $\mu \mapsto \int \phi d\mu$ continuous for all ϕ bounded measurable.*

Exercise 2.6.5. Check that if F is Polish, then the τ-topology is the smallest topology containing the union of all compatible (i.e. that does not creates new measurable sets) topologies of convergence in distribution. (Hint: use the fact that measurable = continuous for a choice of a Polish topology on F).

The Borel sets given by the total variation norm is much larger.

The total variation is a obviously a metric, and the τ-topology is a locally convex topology. Both are however very rich (non-separable), so that the associated Borel σ-algebras are rather nasty (and avoided). In particular, there are some open sets which are not cylindrically measurable.

This discussion explains why one should take care when stating Sanov theorem: random variables takes their value in a Polish space for 'security reasons' (proofs are a bit constructive), but the topology of the LDP is in fact valid, eventually, in much richer (fine) topology (here the τ-topology).

2.7 Topologies on vector spaces

We consider vector spaces on $\mathbb{R}$, and we denote by $A + B$ (and similarly $A + x$) the set obtained by the summing all the vectors contained in A or B. In the same way $c \times A$ is obtained by multiplying all vectors of A by c.

If E is a vector space, the vector space structure gives 'natural' constraints on possible topologies. For instance, one may assume the topology is generated by a subbase satisfying the following constraints:

i) The topology is generated by the sets $\{x + N,\, x \in E, N \in \mathcal{C}\}$ where $\mathcal{C}$ is a collection of subsets containg 0 (invariance by translation).

ii) The sets in $\mathcal{C}$ are convex ('local convexity').

iii) The intersection of any $C \in \mathcal{C}$ and any vector line $\{lx,\, l \in \mathbb{R}\}$ with $x \in E$ is of the form $\{lx,\, l \in]-l_0, l_0[\}$ with $l_0 \in]0, +\infty]$.

Such topologies are called *locally convex*, and are the most general usual topologies considered on vector spaces.

Exercise 2.7.1. Check that i) normed vector spaces, and ii) topologies generated by a given vector subspace of linear forms are locally convex. Those are the most usual ones. Give usual examples. Prove that addition and multiplication by a real coefficient are continuous functions. In particular the topology is invariant by mutliplication ($C \in \mathcal{C}$ implies lC is open and may be added in the base).

One needs to add a condition on the base for the topology to be able to separate points (Hausdorff condition):

- In axiom *iii*) above, for each given vector line, there is at least one $C \in \mathcal{C}$ such that the intersection is finite in the sense that $l_0 < +\infty$.

Exercise 2.7.2. Check out cases for which locally convex spaces that are not metrizable. (Nota Bene: weak-$*$ topologies are never metrizable, but their norm-ball are).

It turns out that locally convex topologies are the same as topologies generated by semi-norms.

Lemma 2.7.3. *A semi-norm is the generalization of a norm without the axiom* $\|x\| = 0 \Rightarrow x = 0$ *(Exercise:* $x \mapsto |\langle x, l\rangle|$ *where* l *is a linear form is a semi-norm). A topology on* E *is locally convex iff it has a subbase given by the open balls of (as infinitely many as necessary) semi-norms.*

Exercise 2.7.4. Describe bases of neighborhoods of locally convex space, for instance when the semi-norms are given by a space of linear forms.

Let E and L are two real vector spaces endowed with a duality, that is a bi-linear form

$$(x, l) \in E \times L \mapsto \langle x, l\rangle \in \mathbb{R}.$$

Exercise 2.7.5. Describe the natural duality when E a space of bounded measurable functions (e.g. continuous), and L is the space of finite measures.

It is possible to consider on E (or symmetrically on L) the topology generated by the half-spaces defined by elements of L, that is the coarsest topology making each linear form $l \in L$ continuous. Such topology is denoted $\sigma(E, L)$ and is by definition a locally convex topology (check it).

We will consider duality pairs that are non-degenerate: if $\langle x, l\rangle = 0$ for any $l \in L$ then $x = 0$ (which makes $\sigma(E, L)$ Hausdorff); and symetrically if $\langle x, l\rangle = 0$ for any $x \in E$ then $l = 0$ (which makes $\sigma(L, E)$ Hausdorff).

Exercise 2.7.6. Show that the convergence in distribution is the restriction on $\mathcal{P}(F)$ of the $\sigma\left(\mathcal{M}_{\mathbb{R}}(F), C_b(F)\right)$-topology.

If E has already a given topology (*e.g.* E is a normed space), one can consider the topological dual E^* defined as the space of *all continuous linear forms* on E.

Exercise 2.7.7. Show that if E is endowed with the weak $\sigma(E, L)$ topology, then $E^* = L$. If E is a normed space, what is called the 'weak' topology on E ? Detail for $\mathbb{L}^p$ spaces.

Unfortunately, it is well-known that neither the dual of continuous bounded functions, nor the dual of finite measures endowed with total variation give back measures or functions (such spaces are much more nasty): if F is non-compact

$$M_b(F) \subsetneq (\mathcal{M}_{\mathbb{R}}(F), \|\ \|_{\text{tv}})^*,$$

and

$$\mathcal{M}_{\mathbb{R}}(F) \subsetneq C_b(F)^*,$$

There are two important exceptions:

Theorem 2.7.8 (Riesz representation)**.** *If K is compact, then*

$$C(K)^* = \mathcal{M}_{\mathbb{R}}(K),$$

that is any continuous linear form can be represented by a finite real measure.

A similar remarkable fact holds for measures that have a density with respect to a reference measure

Theorem 2.7.9 (Riesz representation)**.** *Let μ be a σ-finite measure (it has a density w.r.t. a finite measure). Then*

$$L^1(\mu)^* = L^\infty(\mu),$$

that is any continuous linear form on L^1 can be represented by μ-everywhere bounded function.

If E is a Banach space, E is said to be *reflexive* if the inclusion of E in its bi-dual $E \subset E^{**}$ is an equality.

Exercise 2.7.10. What is a Hilbert space ? Which of the L^p spaces are reflexive ?

Obviously, neither $C_b(F)$, $L^1(\mu)$, $L^\infty(\mu)$, nor $(\mathcal{M}_{\mathbb{R}}(F), \|\ \|_{\text{tv}})$ are reflexive spaces. This can be easily seen since $C_b(K)$ and $L^1(\mu)$ are separable, whereas $(\mathcal{M}_{\mathbb{R}}(F), \|\ \|_{\text{tv}})$ and $L^\infty(\mu)$ are not. Indeed, reflexive pairs are either both separable, or either non-separable.

2.8 Some theorems on weak compactness

Banach-Alaoglu-Bourbaki Given a Banach space E, one can consider one strong topology on E^* (defined by the operator norm), and two types of weak topologies on E^*: the topology $\sigma(E^*, E)$ induced by the original E called the *weak-$*$* topology, and the less weak topology $\sigma(E^*, E^{**})$ induced by the bigger bi-dual. Those two differs if and only if E is not reflexive. The famous Banach-Alaoglu-Bourbaki theorem states that

Theorem 2.8.1 (Banach-Alaoglu-Bourbaki)**.** *In the dual E^* of a Banach space E, norm-bounded sets are pre-compact for the* weak-$*$ *topology.*

Exercise 2.8.2. Check using Riesz on continuous fonctions, that the space of probabilities on a compact set is compact for the topology of convergence in distribution.

Exercise 2.8.3. Using Riesz and Banach-Alaoglu on $l^1 \stackrel{\text{def}}{=} \left\{ x_n | \sum_{n\geq 1} |x_n| < +\infty \right\} = L^1(\mathbb{N}_*)$, recover Tychonov theorem for compactness of uniformly bounded sequences of real numbers.

Exercise 2.8.4 (Proof of Banach-Alaoglu)**.** Admit the general Tychonov theorem: any product of compact space is compact for the product topology. Consider the set of all real valued functions over E as a product space $E^{\mathbb{R}}$ with the product topology.

- Check that E^* is a subspace of $E^{\mathbb{R}}$.
- Check that the trace of the product topology on E^* is the the weak-$*$ topology.
- Check that the image in $E^{\mathbb{R}}$ of the unit (strong) ball of E^* is compact.

A result by Kakutani states that the latter compactness result is only valid for *weak-*$*$ topologies: if all strongly bounded sets are weakly pre-compact then the space is reflexive. This shows that for non-reflexive spaces, there are bounded sets that are not weakly pre-compact. So in short, proving pre-compactness in topologies that are not weak-star requires some extra work.

Prohorov Imagine you want to show that a certain set of probability distributions is pre-compact.

A seen above, we can apply Banach-Alaoglu-Bourbaki theorem only on probabilities on compact sets, but not on a general Polish space since the space of real measures is not the dual space of a simple space of test functions.

Fortunately, one can identify compact sets in $\mathcal{P}(F)$ when F is Polish.

A subset $A \subset \mathcal{P}(F)$ is said to be *tight*, when the tails of the probabilities in A are uniformly small outside compacts of F.

Definition 2.8.5. *$A \subset \mathcal{P}(F)$ is tight if for any $\varepsilon > 0$ there is a compact $K_\varepsilon \subset F$ such that*

$$\inf_{\mu \in A} \mu(K_\varepsilon) \geq 1 - \varepsilon.$$

Theorem 2.8.6 (Prohorov)**.** *Let F be Polish. The closure $\overline{A}$ of subset $A \subset \mathcal{P}(F)$ is compact if and only if it is tight.*

Exercise 2.8.7. Recall the characterization of compact sets in metric spaces in terms of sequences. Check that Prohorov theorem amounts to say: 'any sequence of probability distribution with uniform tails outside compacts converges in distribution up to extraction'.

A poof in a simple case will be given in the next paragraph.

Dunford-Pettis As shown above, $L^1(\mu)$ is not reflexive, so that bounded sets are not generally pre-compact. Here again, we can identify weakly compact sets using the concept of uniform integrability.

Definition 2.8.8. *Assume μ finite. $\mathcal{F} \subset L^1(\mu)$ is uniformly integrable if and only if*

$$\lim_{c \to +\infty} \sup_{f \in \mathcal{F}} \int |f| \mathbb{1}_{|f| \geq c} d\mu = 0.$$

We state:

Theorem 2.8.9. *If μ is finite, then weak pre-compacity in $L^1(\mu)$ is equivalent to uniform integrability.*

Exercise 2.8.10. Show that the set of function satisfying $\int |f| \ln |f| \, d\mu \leq c$ for some $c < +\infty$ is uniformly integrable, hence $L^1(\mu)$-weakly compact

In fact, a result by De la Vallé Poussin states that uniform integrability is equivalent with the existence of a function g increasing stricly faster than linearly such that

$$\sup_{f \in \mathcal{F}} \int g(|f|) \, d\mu < +\infty.$$

Exercise 2.8.11. Compare using De la Vallé Poussin, weak compactness in $\mathbb{L}^p$, $p > 1$ and in $\mathbb{L}^1$.

In the next exercise, we will prove show that in a discrete setting, Prohorov and Dunford-Pettis are the same, and we will make a simple proof using Banach-Alaoglu.

Exercise 2.8.12. We consider $E = \mathbb{L}^1(\mathbb{N}_*, \mu)$ where μ is a strictly positive probability on $\mathbb{N}_*$.

- Check that $\mathcal{P}(\mathbb{N}_*)$ is a subset of $\mathbb{L}^1(\mathbb{N}_*, \mu)$.
- Which topology on E is convergence in distribution on $\mathcal{P}(\mathbb{N}_*)$ the trace topology of ?
- Prove that for subsets of E, tightness is equivalent to uniform integrability of the density with respect to any given $\mu \in E$ with $\mu(x) > 0$ for all x.

We now prove a weaker version of Dunford-Pettis in a weighted l^1 space (we prove that unif. int. implies sequential compactness). Consider a sequence $f_n \in \mathbb{L}^1(\mathbb{N}_*, \mu)$ where μ is a strictly positive probability.

- Consider the function $\hat{f}_n$ with a an integer

  $$\hat{f}_n : (k, a) \mapsto f_n(k) \mathbb{1}_{|f_n(k)| \leq a}$$

  and check that $\hat{f}_n$ is uniformly bounded with respect to n in a well-chosen $\mathbb{L}^2$ space.
- Apply Banach-Alaoglu in $\mathbb{L}^2$ to extract a limit.
- Consider the limit as a $\mathbb{L}^1$ function for the variable k indexed by a, and prove that it is a Cauchy sequence when a becomes large. (Hint: uniform integrability is key here). We admit that norms are lower semi-continuous for the weak topology (it will be proven within the next two sections).
- Conclude by showing that up to extraction f_n converges weakly in L^1.

2.9 Lower semi-continuity

We consider in this section regular topologies: any pair of closed and compact sets can be separated by neighborhoods (Exe: write it formally)

Definition 2.9.1. *A function on a topological space taking value in* $[-\infty, +\infty]$ *is lower semi-continuous iff* $\{x | f(x) \leq a\}$ *is closed for all* $a \in [-\infty, +\infty]$.

Exercise 2.9.2. Check it suffices to verify closedness for $a \in \mathbb{R}$. Check that lower semi-continuous functions are Borel measurable.

Exercise 2.9.3. Prove that on metric space lower semi-continuity is equivalent to $\liminf f(x_n) \geq f(\lim x_n)$ for any converging sequence $\{x_n\}$.

Exercise 2.9.4. Check that if $\mathcal{F}$ is a collection of lower semi-continuous functions, then $g(x) = \sup_{f \in \mathcal{F}} (f(x))$ defines a lower semi-continuous function.

Exercise 2.9.5. Let $\mathcal{B}$ be a basis for a topology. Suppose we are given

$$\mathcal{L} : \mathcal{B} \to [-\infty, +\infty]$$

an decreasing functions with respect to the inclusion order.

- Prove that
$$f(x) \stackrel{\text{def}}{=} \sup_{O_x \in \mathcal{B} : x \in O_x} \mathcal{L}(O)$$
is a lower semi-continuous function.
- Assume that the topology is regular, that the function $\mathcal{L}$ is defined on the collection of all sets, and that it is decreasing with respect to inclusion. Prove that $f(x) = \sup_{O \in \mathcal{B} : x \in O} \mathcal{L}(\overline{O})$.
- Prove that if I is a function, then $I^{\text{lsc}}(x) = \sup_{O \in \mathcal{B} : x \in O} \inf_O I$ is the lower semi-continuous envelope of I: it is the largest lower semi-continuous functions which is lower than I (in a pointwise sense).

Lemma 2.9.6. *Let* $\mathcal{B}$ *be a basis of a regular topological space. Let* f *is lower semi-continuous if and only if either* $f(x) = \sup_{B_x \in \mathcal{B} : x \in B_x} \inf_{\overline{B}_x} f$ *or equivalently* $f(x) = \sup_{B_x \in \mathcal{B} : x \in B_x} \inf_{B_x} f$

2.10 Convexity and convex duality

Convex sets and convex functions can be defined on any real vector space. The most important tool to generalize geometrically intuitive results in the finite dimensional setting to the infinite dimensional one is The Hahn-Banach theorem.

Theorem 2.10.1 (Hahn-Banach)**.** *Let* E *be a locally convex topological vector space. Then any disjoint convex closed set* C *and convex compact set* K *can be* strictly *separated by a continuous hyper-plane: there is continuous linear form* $l : E \to \mathbb{R}$ *such that*

$$\limsup_{x \in C} \langle l, x \rangle < \liminf_{x \in K} \langle l, x \rangle$$

Exercise 2.10.2. Give a simple counter-example to strict separation when K is not compact.

Corollary 2.10.3. *Let E be a vector space with some locally convex topology. Then any convex closed set is also closed for the $\sigma(E, E^*)$ topology. In particular, norms are lower semi-continuous for the weak topology.*

Exercise 2.10.4. Prove the corollary (Hint: show that the complementary is open). Give a counter-example in the form of a non-convex set (Hint: check that a small open ball cannot be weakly open in infinite dimension).

In this section, we consider a dual pair $(E, L, \langle \quad \rangle)$, assumed to be non-degenerate in the sense that E and L are mutually separating: $\langle l, x\rangle = 0$ for all $l \in E$ implies $x = 0$, and symmetrically.

Definition 2.10.5 (Legendre-Fenchel)**.** *Let $(E, L, \langle \quad \rangle)$ a dual pair of vector spaces. If $f : E \to]-\infty, +\infty]$ is any function, the* Legendre-Fenchel transform *of f is the function on L defined by*

$$f^*(l) \stackrel{\text{def}}{=} \sup_{x \in E} \left(\langle l, x\rangle - f(x)\right).$$

Exercise 2.10.6. Compute the Legendre dual of usual convex functions on $\mathbb{R}$: power of $p \geq 1$, exponential, absolute value, functions taking value in $\{0, +\infty\}$.

Exercise 2.10.7. Show that a Legendre-Fenchel transform $f(x) = g^*(x)$ where g is a function on L is necessarily convex and lower semi-continuous for the $\sigma(E, L)$-topology. (Do it on $\mathbb{R}^d$ first).

Exercise 2.10.8. By interverting sup and inf, show that:

$$f^{**} \leq f.$$

Convex duality states that f^{**} is in fact the lower semi-continuous and convex envelope of f in the sense that is f lower semi-continuous and convex, then $f^{**} = f$.

Exercise 2.10.9. Assume $E = L = \mathbb{R}^d$ and $\langle\,\rangle$ is the usual scalar product. Define

$$A(f) \stackrel{\text{def}}{=} \{(l, \alpha) | f(x) \geq x \cdot l - \alpha \, \forall x\}.$$

i) Interpret graphically $A(f)$.

ii) Interpret graphically f^* by showing that $f^*(l)$ is the infimum of all α such that $(l, \alpha) \in A(f)$.

iii) Describe in the same way $f^{**}(x)$ as the supremum of all the $x \cdot l - \alpha$ such that $(l, \alpha) \in A(f)$. Interpret graphically.

iv) Consider the set:

$$C(f) \stackrel{\text{def}}{=} \left\{(x, a) \in \mathbb{R}^d \times \mathbb{R} | f(x) \leq a\right\}$$

Interpret graphically.

v) Assume that $C(f)$ is convex and closed. Pick some $x \in \mathbb{R}^d$ and some $\varepsilon > 0$ and strictly separates with an hyper-plane the set $C(f)$ and the point $(x, f(x) - \varepsilon)$. Deduce that $f^{**}(x) \geq f(x) - \varepsilon$, hence convex duality.

vi) Give an example of a convex function which is not lower semi-continuous.

Convex duality can be generalized to any locally convex space thanks to the Hahn-Banach theorem.

Theorem 2.10.10 (Legendre-Fenchel)**.** *Let $(E, L, \langle\,\rangle)$ be a non-degenerate (mutually separating) dual pair of vector spaces. Endow E with the locally convex $\sigma(E, L)$ (weak) topology, and assume that $f : E \to]-\infty, +\infty]$ is a convex lower semi-continuous function. Then convex duality holds:*

$$f^{**}(x) = f(x).$$

Exercise 2.10.11. Prove the theorem by revisiting the finite dimensional case.

3 Relative entropy and its variational formulation

Definition 3.0.1. *Let F be a Polish space and π a reference probability distribution on F. The* relative entropy *is defined by*

$$\mathrm{Ent}(\mu|\pi) \stackrel{\mathrm{def}}{=} \begin{cases} \int_F \ln \frac{d\mu}{d\pi} d\mu & \mu \ll \pi \\ +\infty & \textit{else.} \end{cases}$$

The relative entropy is extended to the space $\mathcal{M}_{\mathbb{R}}(F)$ of real valued measures on F by setting

$$\mathrm{Ent}(\mu|\pi) = +\infty, \qquad \mu \notin \mathcal{P}(F)$$

The goal of this section is to show that i) relative entropy is a good lower semi-continuous convex function for the topology of convergence in distribution (and for the stronger τ-topology), and ii) that its convex dual is th functional version of the (logarithmic) cumulant generating function.

Definition 3.0.2. *The (functional, logarithmic)* cumulant generating function *of a probability μ on F is the function defined on the space of bounded test functions by:*

$$\Lambda(\varphi) \stackrel{\mathrm{def}}{=} \ln \int_F \mathrm{e}^{\varphi} d\mu.$$

Exercise 3.0.3. Show that Ent and Λ are convex function. Why is Λ continuous for the uniform norm topology ?

Exercise 3.0.4. Assume that F is finite and let $\pi > 0$ be a reference measure.

i) Solve the Euler-Lagrange equation associated with the optimization problem defining the convex dual of $\mathrm{Ent}(\,|\pi)$. Describe the unique probability that solves the optimization problem.

ii) Do the same for the cumulant generating function. Nota Bene: the optimal function defining convex duality is unique up to an additive constant.

iii) Prove again i) and ii) by using Jensen inequality. Hint: Develop: $\mathrm{Ent}\left(\,.\,|\frac{\mathrm{e}^{\phi}\pi}{\int \mathrm{e}^{\phi} d\pi}\right)$

iv) Check that convex duality holds true.

We can now state the general theorem.

Theorem 3.0.5 (Convex duality for relative entropy)**.** *Let F be a Polish space, and π any probability measure on F. Relative entropy and the functional cumulant generating function are convex dual conjugates for the dual pair $(\mathcal{M}_{\mathbb{R}}(F), M_b(F))$ as well as $(\mathcal{M}_{\mathbb{R}}(F), C_b(F))$. In other words, for any finite measure μ:*

$$\mathrm{Ent}\left(\mu|\pi\right) = \sup_{\varphi \in C_b(F)} \left(\mu(\varphi) - \ln \pi\left(\mathrm{e}^{\varphi}\right)\right) = \sup_{\varphi \in M_b(F)} \left(\mu(\varphi) - \ln \pi\left(\mathrm{e}^{\varphi}\right)\right),$$

and for any bounded measurable φ:

$$\ln \pi\left(\mathrm{e}^{\varphi}\right) = \sup_{\mu \in \mathcal{M}_{\mathbb{R}}(F)} \left(\mu(\varphi) - \mathrm{Ent}\left(\mu|\pi\right)\right) = \sup_{\mu \in \mathcal{P}(F)} \left(\mu(\varphi) - \mathrm{Ent}\left(\mu|\pi\right)\right),$$

Moreover, the convex lower semi-continuous function $\mu \mapsto \mathrm{Ent}\left(\mu|\pi\right)$ is good (it has compact level sets) for the $\sigma(\mathcal{M}_{\mathbb{R}}(F), M_b(F))$-topology (and in particular also for the weaker $\sigma(\mathcal{M}_{\mathbb{R}}(F), C_b(F))$-topology).

Exercise 3.0.6 (Proof of the theorem)**.** We start by proving that $\mu \mapsto \mathrm{Ent}\,(\mu|\pi)$ is convex and lower semi-continuous in the $\sigma(\mathcal{M}_{\mathbb{R}}(F), M_b(F))$-topology.

- Prove convexity.
- Check that the result is equivalent to lower semi-continuity for the $L^1(\pi)$-weak topology.
- Prove strong lower semi-continuity using Fatou. Deduce weak lower semi-continuity.

Next, we prove that the cumulant generating function is the dual of entropy.

- Give the candidate solutions of the optimization problem.
- Prove the claim using Jensen inequality.

Next we prove convex duality in the $\sigma(\mathcal{M}_{\mathbb{R}}(F), C_b(F))$-topology.

- Prove that entropy and the cumulant generating function are convex conjugate using pointwise approximation by continuous functions.

Finally compacity of level sets is given by:

- Prove 'gooodness' using Dunford-Pettis.

4 Varadhan, Cramér, and Sanov

Remark on handling inequalities

Orientation of inequalities in LDPs may be quite confusing, especially because the sign convention

$$-\inf I$$

which is not very natural. A helpful method to deal with that is the following. Assume you want to prove an upperbound $\limsup_n \frac{1}{n} \ln \mathbb{P}[X_n \in F] \leq -\inf_F I$; then use the following scheme:

On the one hand, we have the upper bound

$$\limsup_n \frac{1}{n} \ln \mathbb{P}[X_n \in F] \leq \limsup_n A_n \leq B.$$

on the other hand, we have the upper bound

$$\inf_F I \leq C.$$

Then check that $B \leq -C$, typically up to a freely chosen $\epsilon > 0$ can be taken arbitrarily small.

4.1 Varadhan's lemmas

In this section, we prove Varadhan's upper bound and lower bound. The form is slightly more general than in Dembo and Zeitouni [1998], and the proof is vey similar to the one in Rassoul-Agha and Seppäläinen [2015]. We present it as a *generalization of the definition* of LDP upper and lower bounds.

As an introduction, let us recall the Laplace's principle (that requires no topology).

Lemma 4.1.1 (Laplace principle)**.** *Let μ be a probability on E, and $V : E \to [-\infty, +\infty]$ a measurable function then*

$$\lim_{n \to +\infty} \frac{1}{n} \ln \int_E \mathrm{e}^{-nV} d\mu = -\mu\text{-essinf } V.$$

Proof. Assume μ-essinf$V$ is finite. Check one can assume without lost of generality that $V \geq 0$ and μ-essinf$V = 0$. Then check $\mathbb{1}_{V-\varepsilon} \mathrm{e}^{-n\varepsilon} \leq \mathrm{e}^{-nV} \leq 1$ and conlude. Show the limit is $+\infty$ with the lower bound when $\mu\text{-essinf } V = -\infty$. □

Laplace principle enables to obtain Varadhan lemmas (and then – as we will see below – a LDP) for a sequence of the form $\mu_n \propto \mathrm{e}^{-nV} d\mu$ when V has some (semi-)continuity properties.

Lemma 4.1.2. *Let F be Polish. and μ a probability. We recall that* $\mathrm{supp}(\mu)$ *is the intersection of all closed sets with measure* 1*. For any $V : F \to [-\infty, +\infty]$ measurable*

$$\inf_{\mathrm{supp}(\mu)} V \leq \mu\text{-essinf} V,$$

and if V is upper semi-continuous then

$$\mu\text{-essinf}V \leq \inf_{\text{supp}(\mu)} V.$$

In particular if V is continuous the LDP holds true for $\frac{\mathrm{e}^{-nV}d\mu}{\int \mathrm{e}^{-nV}d\mu}$ with rate function

$$I^V = V + I^0 - \inf\left(V + I^0\right)$$

where

$$I^0 = \begin{cases} 0 & on\ \text{supp}(\mu) \\ +\infty & else \end{cases}$$

Exercise 4.1.3 (Proof and consequences of the lemma above)**.**

- Proof: check that $\left\{x | V(x) < \inf_{\text{supp}(\mu)} V\right\}$ is disjoint from $\text{supp}(\mu)$. Conclude.
- Proof: let $\varepsilon > 0$, check that $\left\{x | V(x) < \inf_{\text{supp}(\mu)} V + \varepsilon\right\}$ is open and contains a point of $\text{supp}(\mu)$. Conclude.
- Prove Varadhan upper and lower bounds.
- Assume V continuous and prove the LDP for $\propto \mathrm{e}^{-nV}d\mu$.

In Varadhan's lemmas, we consider $(\mu_n)_n$ be a sequence of probabilities on E, and $I : E \to [0, +\infty]$ a lower semi-continuous function.

Lemma 4.1.4 (Varadhan's lower bound)**.** *The lower bound of the LDP with rate function I is equivalent to the following: for any upper semi-continuous function $V : E \to]-\infty, \infty]$ one has*

$$\liminf_n \frac{1}{n} \ln \int_E \mathrm{e}^{-nV} d\mu_n \geq -\inf_E (V + I)$$

Exercise 4.1.5 (Proof)**.**

- Check that Varadhan $\Rightarrow$ LDP lower bound using a well chosen V.
- LDP $\Rightarrow$ Varadhan. First, check that $\mathrm{e}^{-nV} \geq \mathbb{1}_O \mathrm{e}^{-nv_0}$ for a well-chosen open set O defined by a v_0-level set of V. On the one hand, deduce a lower bound on $\liminf_n \frac{1}{n} \ln \int e^{-nV} d\mu_n$ from the LDP that depends on v_0. On the other hand check that for some x_ε, $\inf(I + V) \geq I(x_\varepsilon) + V(x_\varepsilon) - \varepsilon$ and conclude by choosing $v_0 = V(x_\varepsilon) + \varepsilon$.

Lemma 4.1.6 (Varadhan's upper bound)**.** *The upper bound of the LDP with rate function I is equivalent to: for any lower semi-continuous function $V : E \to]-\infty, \infty]$ with tail condition*

$$\lim_{v_0 \to -\infty} \limsup_n \frac{1}{n} \ln \int_E \mathrm{e}^{-nV} \mathbb{1}_{V \leq v_0} d\mu_n = -\infty. \tag{9}$$

one has

$$\limsup_n \frac{1}{n} \ln \int_E \mathrm{e}^{-nV} d\mu_n \leq -\inf_E (V + I)$$

Exercise 4.1.7 (Proof)**.**

- Check that Varadhan ⇒ LDP using a well chosen V.
- Assume V is bounded from below, and pick $v_{K+1} > \inf(I+V)$. Consider a finite partition of E defined by given level sets of V: $V^{-1}(]v_k, v_{k+1}])$, $1 \leq k \leq K$, and use it to upper bound the integral $\int_E \mathrm{e}^{-nV} d\mu_n$. Then apply the LDP and conclude.
- Assume V is general with the tail condition. Distinguish the two cases $V \leq v_0$ and $V < v_0$ and take the limsup. Finally take $v_0 \to +\infty$.

Corollary 4.1.8. *Prove that if μ_n satifies a LDP with rate function I and that V is continuous and satisifes the tail condition* (9)*, then*

$$\lim_n \frac{1}{n} \ln \int_E \mathrm{e}^{-nV} d\mu_n \leq -\inf_E (V + I)$$

and

$$\mu_n^V = \frac{1}{\int \mathrm{e}^{-nV} d\mu_n} \mathrm{e}^{-nV} d\mu_n$$

satisfies a LDP with rate function $I + V - \inf_E (I + V)$.

We end with a practical sufficient criteria for the tail condition:

Lemma 4.1.9. *Let $(\mu_n)_{n\geq 1}$ a sequence of probabilities and V a real valued measurable function. If there is a $\gamma > 1$ such that*

$$\limsup_n \frac{1}{n} \ln \int_E \mathrm{e}^{-n\gamma V} d\mu_n < +\infty$$

then the tail condition (9) *holds true.*

Proof. Use a Chernoff-like bound. □

4.2 Exponential tightness

The goal of this section is to prove that any exponentially tight sequence of probability distributions satisfies, up to extraction, a LDP with a good rate function.

In a first lemma, we will assume that E is endowed with a countably generated topology – we will assume that E is Polish for simplicity – and prove that a *weak version of a LDP* holds true up to extraction of a sub-sequence.

Lemma 4.2.1. *Let $(\mu_n = \mathrm{Law}(X_n))_{n\geq 1}$ denote a sequence of probabilities defined on the Borel sets of a countably generated (regular) topology of E. Then there is a sub-sequence and a lower semi-continuous $[0,+\infty]$-valued function I satisfying: i) an LDP lower bound, and ii) a LDP upper bound for compact sets:*

$$\forall K \subset E,\ K \text{ compact},\ \frac{1}{n}\ln \mathbb{P}(X_n \in K) \leq -\inf_K I.$$

i) and ii) may be called a weak *LDP.*

Proof. Step 1 (Extraction): Let $\mathcal{B}$ denote a countable base for the topology, and extract a sub-sequence such that $\mathcal{L}(O) \stackrel{\text{def}}{=} -\lim_n \frac{1}{n}\ln \mathbb{P}(X_n \in O)$ converges for each $O \in \mathcal{B}$ (why can we do this ?).

Step 2 (Construction of I): How would we recover I from $\mathcal{L}(O)$ if a LDP were holding true ? Propose a definition for I. Check it is indeed lower semi-continuous.

Step 3 (LDP lower bound) Let O be an open set and let $x \in O$ be given. Check using $\mathcal{L}$ that the infimum limit of $\frac{1}{n}\ln \mathbb{P}(X_n \in O)$ is greater that $-I(x)$, and conlcude.

Step 4 (LDP upper bound) Let $\varepsilon > 0$ and K a compact set be given. Associate to each point $x \in K$ an open set O_x^ε of the base such that $I(x) \leq \mathcal{L}(O_x^\varepsilon) + \varepsilon$. Consider a finite covering $K \subset \bigcup_{1\leq k\leq K}, O_{x_k}$ by the latter. On the one hand, give an upper bound $\frac{1}{n}\ln \mathbb{P}_n[X_n \in K]$ using the covering and compute the limit for large n (recall that $\lim_n \frac{1}{n}\ln(\text{sum}) = \max\left\{\lim_n \frac{1}{n}\ln\right\}$). On the other hand remark that $\inf_K I \leq \min(I(x_1),\ldots,I(x_K))$.

□

Exercise 4.2.2. Check the details of the above proof.

We can strengthen the above weak theorem to a usual LDP with good rate function if and only if exponential thightness holds true (Nota Bene: E is assumed to be Polish to avoid uninteresting technicalities – a typical assumption for countably generated topologies).

Exponential tightness is to LDPs what tightness is to convergence in distribution. Recall that for sequences, tightness means that

$$\inf_{K \text{ compact}} \limsup_n \mathbb{P}(X_n \notin K) = 0$$

(Exercise: check this).

Definition 4.2.3. *A sequence of probabilities on a Polish space is said to be* exponentially *tight iff*

$$\inf_{K\ compact} \limsup_n \frac{1}{n}\ln \mathbb{P}(X_n \notin K) = -\infty$$

In particular, the sequence is tight.

We then obtain:

Proposition 4.2.4. *If a sequence of probabilities on a Polish space is exponentially tight, then there is a subsequence satisfying a LDP with a good rate function I.*

Proof. Step 1: Extract a subsequence satisfying a weak LDP.

Step 2: Let F closed and K compact be given. Prove the LDP upper bound for F using $F \subset (F \cap K) \cup K^c$ and the weak upper bound.

Step 3: Prove that I is a good rate function by using the LDP lower bound for the complementary of compact sets.

□

Exercise 4.2.5. Check the details of the above proof.

Finally, let us remark that the converse is also true on a Polish state space:

Lemma 4.2.6. *Any sequence $(X_n)_{n\geq 1}$ satisfying a LDP with a good rate function on a Polish space is exponentially tight.*

Proof. Denote by

$$K^{\varepsilon} \stackrel{\text{def}}{=} \{x | d(x,K) < \varepsilon\}$$

the open ε-thickening of K. Step 1: Prove using tightness of a single probability on a Polish space that for each (small) $\varepsilon > 0$ and each (large) $a > 0$ there exists a compact set $K_{a,\varepsilon}$

$$\sup_{n\geq 1} \frac{1}{n} \ln \mathbb{P}\left[X_n \in \left(K_{a,\varepsilon}^{\varepsilon}\right)^c\right] \leq -a$$

Step 2: Prove that for any sequence $(K_p)_{p\geq 1}$ of compact sets then the intersection

$$\bigcap_{p\geq 1} K_p^{1/p}$$

is totally bounded (that is for any ε it can be covered by finitely many small balls of size ε) and thus pre-compact. Apply it to the set $K_{a+p,1/p}$ of Step 1. is compact. Step 3: Prove exponential tightness using the compact set constructed in Step 3. □

Exercise 4.2.7. Do a simple proof in the case $E = \mathbb{R}^d$ by working with balls.

4.3 Cramér's theorem

Cramér's theorem gives the LDP for empirical avergaes of i.i.d. variables.

Theorem 4.3.1 (Cramér). *Consider a sequence of empirical averages* $\left(X_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{m=1}^{n} Z_m \right)_{n \geq 1}$ *where* Z_m, $m \geq 1$ *are i.i.d. taking value in* $\mathbb{R}^d$. *Assume that the cumulant generating function given by*

$$\Lambda(l) \stackrel{\text{def}}{=} \ln \mathbb{E}\left[\mathrm{e}^{Z \cdot l} \right]$$

is finite on an open set of $\mathbb{R}^d$ *containing* $l = 0$.

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

$$I = \Lambda^*,$$

that is the convex dual of Λ *on* $\mathbb{R}^d$. *For* $d = 1$, *the result is true without condition on* Λ.

In the present note, we will only prove under the assumption that $\Lambda(l) < +\infty$ for any $l \in \mathbb{R}^d$. This excludes laws with exactly exponential tails, but not more. The minimal assumptions in $\mathbb{R}^d$ requires additional technical details.

Before considering i.i.d. sequences, we will consider a general lemma in the case where the space $E_0 \subset E$ is the convex subset of a topological vector space E. It may then be possible to identify the convex dual of the rate function I using convex duality.

Lemma 4.3.2. *Let* $(\mu_n = \mathrm{Law}(X_n))_{n \geq 1}$ *be a sequence of probabilities on a Polish space* E_0.

i) $E_0 \subset E$ *is a closed convex subset of a locally convex vector space* E.

ii) The Polish topology of E_0 *is given by the trace topology of* E.

iii) The sequence $(\mu_n)_{n \geq 1}$ *satisfies a LDP with rate function* I.

iv) The limit of the cumulant generating funtion

$$\Lambda(l) \stackrel{\text{def}}{=} \lim_{n \to +\infty} \frac{1}{n} \ln \mathbb{E}\left[\mathrm{e}^{n \langle X_n, l \rangle} \right] \in \mathbb{R}$$

exists (and in particular is finite) for all $l \in E^*$.

Then Λ *is the convex dual in* $(E, E^*, \langle\ \rangle)$ *of the rate functions* I *of any sub-sequence satisfying a LDP, that is:*

$$\Lambda(l) = I^*(l) = \sup_{x \in E_0} \left(\langle x, l \rangle_{E_0, L} - I(x) \right).$$

Proof. Apply Varadhan's lemma in E_0 to the cumulant generating function after having checked a moment condition. □

Exercise 4.3.3. Prove directly that Λ is convex and lower semi-continuous for the $\sigma(E^*, E)$-topology.

We can now start to focus on Cramér's theorem. It is clear from the lemma above that a key step will be to show that the rate funtcion I is convex which enables to itentify it with the dual of Λ.

Convexity of I will be a consequence of a simple convexity poperty of addition of independent random variables.

Lemma 4.3.4. *Let X, Y two random variables taking value in a real vector space E. Then for any measurable $A, B \subset E$ and any $\theta \in [0,1]$:*

$$\mathbb{P}[X \in A]\,\mathbb{P}[Y \in B] \leq \mathbb{P}[\theta X + (1-\theta)Y \in \theta A + (1-\theta)B],$$

where $\theta A + (1-\theta)B$ is the set of vectors of the form $\theta x + (1-\theta)y$ with $x \in A, y \in B$.

Exercise 4.3.5. Prove the lemma.

We can next prove convexity of rate functions.

Lemma 4.3.6. *Assume that the sequence of empirical averages $\left(X_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{m=1}^{n} Z_m\right)_{n\geq 1}$ where Z_m, $m \geq 1$ are i.i.d. takes value in $E_0 \subset E$ a convex subset of a locally convex vector space E, E_0 being Polish for the trace topology. Assume that X_n satisfies a LDP with rate function I in E_0. Then I is convex.*

Proof. Step 1: Check that, for the considered topology, there is a translation invariant base $\mathcal{B}$ of convex subsets. Denote $O_x = x + O_0$ elements of this base and simplify

$$\theta O_x + (1-\theta)O_y$$

Step 2: Pick $\theta \in]0,1[$, $x, y \in E_0$, and $O_x, O_y \in \mathcal{B}$ two convex open sets in the base containing respectively x and y. Prove using the LDP that

$$\inf_{O_{\theta x+(1-\theta)y}} I \leq \theta \inf_{O_x} I + (1-\theta)\inf_{O_y} I.$$

Step 2: Conclude by proving the convexity of I using its lower semi-continuity. □

We can now conclude by stating and proving Cramérs theorem in the general case.

Theorem 4.3.7. *Consider a sequence of empirical averages $\left(X_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{m=1}^{n} Z_m\right)_{n\geq 1}$ where Z_m, $m \geq 1$ are i.i.d. taking value in takes value in $E_0 \subset E$ a closed convex subset of a locally convex vector space E, E_0 being Polish for the trace topology.*

i) The sequence $(X_n)_{n\geq 1}$ is exponentially tight in E_0.

ii) The cumulant generating function given by

$$\Lambda(l) = \ln \mathbb{E}\left[\mathrm{e}^{\langle Z, l\rangle}\right] < +\infty$$

is assumed to be finite for all $l \in E^*$.

Then $(X_n)_{n\geq 1}$ satisfies a LDP in E_0 with good convex rate function:

$$I = \Lambda^*,$$

Λ^* being the convex dual of Λ for the duality pair $(E, E^*, \langle\ \rangle)$.

Proof. Step 1: extract a sub-sequence satisfying a LDP in E_0 with (good) and convex rate function I.

Step 2: Apply Lemma 4.3.2.

Step 3 Extend I to E by setting $I = +\infty$ outside E_0. Prove that $I : E \to [0, +\infty]$ is lower semi-continuous for the weak $\sigma(E, E^*)$-topology. Step 4 Conclude after having identified by convex duality the rate function I. □

Exercise 4.3.8. Prove the Cramér theorem in $\mathbb{R}^d$ with the assumption that $\mathbb{E}\left[\mathrm{e}^{\langle Z,l\rangle}\right] < +\infty$ for any $l \in \mathbb{R}^d$.

Exercise 4.3.9 (Reflexive separable Banach spaces, $*$)**.** Check that in the case where $E = E_0$ is reflexive, separable Banach then the condition

$$\mathbb{E}\left[\mathrm{e}^{\gamma\|Z\|_E}\right] < +\infty \qquad \forall \gamma > 0,$$

is sufficient to apply the general Cramérs theorem above. (Hint: use Banach-Alaoglu and a Chernoff bound to obtain exponential tightness. Then use the finiteness of operator norm of continuous linear forms).

Consider the case where $E = \mathbb{L}^p([0, T])$ and Z is a pure jump Markov process in $\mathbb{R}$ as well as its application for cash flow risks in insurance.

4.4 Sanov theorem

In this section, we will prove Sanov theorem using Theorem 4.3.7. We start by proving exponential tightness of empirical measures of i.i.d. variables, for the topology of convergence in distribution.

Proposition 4.4.1 (Exponential tightness of empirical distributions)**.** *Consider a sequence of empirical measures* $\left(\Pi_n = X_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{m=1}^{n} \delta_{Y_m}\right)_{m\geq 1}$ *where* Y_m, $m \geq 1$ *are i.i.d. taking value in a Polish space* F*. Then* $(\Pi_n)_{n\geq 1}$ *is exponentially tight in* $\mathcal{P}(F)$ *endowed with the topology of convergence in distribution.*

Proof. Step 1: In order to prove exponential tightness, we need to construct compact sets in $\mathcal{P}(F)$. Let $(\varepsilon_p)_{p\geq 1}$ be a given decreasing sequence with $\lim_p \varepsilon_p = 0$. Check you can construct a sequence (K_p) of compact sets in F such that for any p

$$\mathbb{P}(Y \in K_p^c) \leq \varepsilon_p.$$

Prove that

$$\mathcal{K}_{p_0} = \cap_{p\geq p_0} \left\{\mu \in \mathcal{P}(F) | \mu(K_p^c) \leq \frac{1}{p}\right\}$$

is compact for each $p_0 \geq 1$.

Step 2: Construct an upper bound of $\mathbb{P}(\Pi_n \in \mathcal{K}_{p_0}^c)$ using the probabilities $\mathbb{P}\left(\Pi_n(K_p^c) > \frac{1}{p}\right)$ for $p \geq p_0$. What do we need to conclude the proposition ?

Step 3: Check that if a function $f : \mathbb{N} \to \mathbb{R}$ grows at least as fast as linearly, i.e. $f(p) \geq ap + b$ with $a > 0$, then

$$\lim_{p_0 \to +\infty} \limsup_n \frac{1}{n} \ln \sum_{p\geq p_0} \mathrm{e}^{-nf(p)} = -\infty.$$

Step 4: Construct a Chernoff upper bound of the event $\Pi_n(K_p^c) > \frac{1}{p}$ which is a good candidate to be of order $\mathrm{e}^{-\mathrm{cte}\, pn}$ when p and n are large. Conclude.

□

Sanov theorem is then a simple application of the general Cramér theorem, Theorem 4.3.7.

Theorem 4.4.2 (Sanov)**.** *Let* $(Y_m)_{m\geq 1}$ *denote a sequence of of i.i.d. variables taking values in a in Polish* F*. Denote* $\mathrm{Law}(Y_m) = \pi_0$*. Then the sequence of empirical distribution defined for* $n \geq 1$ *by*

$$X_n = \Pi_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{m=1}^{n} \delta_{Y_m} \in E$$

satisfy a LDP on $\mathcal{P}(F)$ *endowed with convergence in distribution, with good rate funcion*

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

where Ent *denotes* relative entropy.

Proof. Step 1: Set $E = \mathcal{M}_{\mathbb{R}}(F)$, and endow E with the $\sigma(E, C_b(F))$-topology, so that the trace topology on $E_0 = \mathcal{P}(F)$ is the topology of convergence in distribution, which is Polish.

Step 2: Recall the result of Section 3 which enables identify the rate function I with relative entropy.

□

In, fact Sanov theorem is still true for the stonger τ-topology, that is the trace topology on $\mathcal{P}(F)$ of the $\sigma(\mathcal{M}_{\mathbb{R}}(F), M_b(F))$-topology.

Note that measurable sets in $\mathcal{P}(F)$ defining probabilistic events are still cylindrical; but one should be aware that not all open or closed sets of the τ-topology are cylindrically measurable.

Exercise 4.4.3 (Sanov theorem for the τ-topology)**.** Prove Sanov theorem for the τ-topology using basic results of Descriptive Set Theory: any Borel set (resp. any Borel measurable function) of a Polish space can be made open (resp. continuous) by refining the Polish topology.

5 List of main theorems

In this section, we provide a short list of general theorems that enables to establish a LDP.

5.1 General Principles

Proposition 5.1.1 (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

$$I_f(z) \stackrel{\text{def}}{=} \inf_{x\in E:\, f(x)=z} I(x).$$

Proposition 5.1.2 (Tensorization Principle)**.** *Assume $(X_n^1)_{n\geq 1}$ and $(X_n^2)_{n\geq 1}$ are independent and both satisfies a LDP in E_1 and E_2 and with good rate function I_1 and I_2. Then $(X_n^1, X_n^2)_{n\geq 1}$ satisfies a LDP with good rate function*

$$I(x^1, x^2) = I_1(x^1) + I_2(x^2).$$

Proposition 5.1.3 (Varadhan Principle)**.** *Assume $(\mu_n)_{n\geq 1}$ satisfies a LDP with rate function I and that V is continuous and satisfies the tail condition* (9)*, then*

$$\mu_n^V = \frac{1}{\int \mathrm{e}^{-nV} d\mu_n}\, \mathrm{e}^{-nV} d\mu_n$$

satisfies a LDP with rate function $I + V - \inf_E (I + V)$.

5.2 Specific Principles

Proposition 5.2.1 (Constant sequence)**.** *A constant sequence μ_0 satisfies a LDP with rate function*

$$I(x) = \begin{cases} 0 & x \in \operatorname{supp}(\mu_0) \\ +\infty & \text{else} \end{cases}$$

Proposition 5.2.2 (Cramér in $\mathbb{R}^d$)**.** *Empirical averages of i.i.d. variables with exponential moments (at least in a neighborhood of 0) satisfies a LDP with a good rate function given by the convex conjugate of the moment generating function.*

Proposition 5.2.3 (Sanov)**.** *Empirical distributions of i.i.d. variables satisfies a LDP (for the topology of convergence in distribution) with a good rate function given by relative entropy.*

6 Statistical Mechanics

6.1 Newton equation

Consider n physical particles in $\mathbb{R}^3$, with masses $m_m > 0$, $m = 1 \dots n$. Denoting the diagonal mass matrix in $\mathbb{R}^{3n \times 3n}$

$$M \stackrel{\text{def}}{=} \begin{pmatrix} m_1 \mathrm{Id}_{\mathbb{R}^3} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & m_n \mathrm{Id}_{\mathbb{R}^3}, \end{pmatrix}$$

as well as the position evolution of particles through (absolute) time

$$t \mapsto q(t) \in \mathbb{R}^{3N},$$

Newton equation for an isolated systems states that

$$M \frac{d^2}{dt^2} q(t) = -\nabla U(q(t))$$

where

$$U : \mathbb{R}^{3n} \to \mathbb{R}$$

is a potential energy, typically of the form (for two-body interactions)

$$U(q) = \sum_{m=1}^{n} U_{\text{free}}(q_m) + \sum_{1 \leq m < m' \leq n} U_{\text{int}}(q_m, q_{m'}).$$

Exercise 6.1.1. Give the precise formula for the n-body gravitational problem.

Exercise 6.1.2. Give conditions for well-podesness using Cauchy-Lipschitz theorem.

6.2 Hamilton and Lagrange formalism

If we definite the momentum of the system as

$$p \stackrel{\text{def}}{=} Mv \in \mathbb{R}^{3n},$$

where v is the velocity, as well as the kinetic energy

$$T(p) \stackrel{\text{def}}{=} \frac{1}{2} p^T M^{-1} p,$$

and the total energy

$$H(q, p) \stackrel{\text{def}}{=} T(p) + U(q),$$

It is easy to re-write Newton's equation as Hamilton's equation

$$\begin{cases} \frac{d}{dt} q(t) = \partial_p H(q(t), p(t)), \\ \frac{d}{dt} p(t) = -\partial_q H(q(t), p(t)), \end{cases} \tag{H}$$

where $\partial_q H \in \mathbb{R}^{3n}$ is the differential with respect to the $3n$ position coordinates.

Remark 6.2.1 (On manifolds)**.** This enables to generalize easily Hamilton's equation on a differentiable manifold $\mathcal{M}$, provided one is given a real valued smooth energy function on the co-tangent space

$$H : T^*\mathcal{M} \to \mathbb{R}$$

using the fact that the differential of a function on a manifold is an element of co-tangent space so that $\partial_p H(q,p) \in T_q\mathcal{M}$ and $\partial_q H(q,p) \in T_q^*\mathcal{M}$, and we have thus the canonical vector space identification

$$T_{(q,p)}\left(T^*\mathcal{M}\right) \equiv T_q\mathcal{M} \times T_q^*\mathcal{M}.$$

It can be checked that as soon as the kinetic energy is a (lower semi-continuous) convex function, Hamilton's equation are equivalent to the Euler-Lagrange equations in *path space* of a quantity called *action* an constructed from a function on tangent space $T\mathcal{M}$ called Lagrangian. Consider the Lagrangian function on the tangent space $T\mathcal{M} = \mathbb{R}^{3n} \times \mathbb{R}^{3n}$,

$$L(q,v) \stackrel{\text{def}}{=} T(v) - U(q),$$

where we have abused notation and denoted the kinetic energy

$$T(v) = \frac{1}{2} v^T M v \in \mathbb{R}$$

that takes values on velocities.

Then it is possible to check that the Hamilton's equations are equivalent to the Euler-Lagrange equations

$$\text{(H)} \Leftrightarrow \text{(EL)},$$

where $\frac{d}{dt}q(t) = v(t)$ and

$$\frac{d}{dt}\left[\partial_v L(q(t),v(t))\right] = -\partial_q L(q(t),v(t)), \tag{EL}$$

is associated satsified by critical points (vanishing first order variations) of the action

$$\mathcal{A}\left(q : [0,T] \to \mathbb{R}^{3n}\right) = \int_0^T L(q(t),v(t))dt$$

when the two endpoints $q(0)$ and $q(T)$ are fixed.

Remark that $v \mapsto L(q,v)$ is the convex dual of $p \mapsto H(q,p)$ for fixed q.

Exercise 6.2.2.

- Prove that critical points of $\mathcal{A}$ with fixed end configurations satisfies (EL).
- Check that (EL) is equivalent to (H).

  Show that the critical point of the action satisfies the Euler-Lagrange equations are the critical point

- Prove that the Hamiltonian-Lagrangian duality can be extended as follows. Check that the Hamiltonian equations are the Euler-Lagrange equations of the augmented action where endpoints positions are fixed.

$$\tilde{\mathcal{A}}\left((q,p):[0,T]\to T^*\mathcal{M}\right)=\int_0^T\left\langle p(t),\frac{d}{dt}q\right\rangle-H(q(t),p(t))dt.$$

  Relate the critical points of the augmented action with those of $\mathcal{A}$ in the case where $v\mapsto L(q,v)$ and $p\mapsto H(q,p)$ are convex conjugate.

6.3 Conservation of energy anf Liouville theorem

It turns out that Hamilton's equations (H) satisfies two fundamental conservation laws.

First, they obviously satisfy energy conservation.

Lemma 6.3.1 (Energy conservation)**.** *Solutions of* (H) *satisfy*

$$\frac{d}{dt}H(q(t),p(t))=0$$

Then they also conserve the canonical symplectic form of $T^*\mathcal{M}$. We won't give here the details of this property since we don't need it, and simply states a corollary: the conservation of phas-space volume (Liouville).

Definition 6.3.2. *The Lebesgue measure on the phase-space $\mathbb{R}^{3n}\times\mathbb{R}^{3n}=T^*\mathbb{R}^{3n}$ is invariant by a diffeomorphic change of coordinates of $\mathbb{R}^{3n}$, that is by the transformation*

$$\begin{cases}\tilde{q}=\Phi(q)\\ \tilde{p}=\left[D_q\Phi\right]^{-1}p.\end{cases}$$

This measure is called the phase-space (or Liouville) measure*.*

We obtain:

Lemma 6.3.3 (Phase-space conservation)**.** *Solutions of* (H) *conserve phase-space measure that is if*

$$t\mapsto q_{q,p}(t),p_{q,p}(t)$$

denotes a solution with initial $(q_{q,p}(0),p_{q,p}(0))=(q,p)$, then for all Borel set A and time t

$$\int\mathbb{1}_{(q,p)\in A}dqdp=\int\mathbb{1}_{q_{q,p}(t),p_{q,p}(t)\in A}dqdp$$

6.4 Classical statistical mechanics

In classical statistical mechanics, one usually consider some macroscopic parameters, or observables, for instance the total energy (if the system is isolated), the volume of a box, or some other averaged quantities. Then one assumes that this macroscopic variables are *either fixed, or very slow*, so that the remaining degrees of freedom are distributed according the probability defined by the *conditioned phase-space measure.*

Definition 6.4.1. *Let ξ denotes macroscopic variables taking values in $\mathbb{R}^p$. We call conditional phase-space measures denoted*

$$\delta_{\xi(q,p)=z}(dqdp)$$

measurable kernels satisfying

$$\delta_{\xi(q,p)=z}(dqdp)dz = dqdp.$$

where dz is the Lebesgue measure of $\mathbb{R}^p$.

Assumption (Statistical Mechanics)**.** *Unkown degrees of freedom are distributed according to a conditional phase-space measure, the conditioning being given by the considered macroscopic variables ξ.*

If

$$t \mapsto \xi(q_t, p_t)$$

is constant (stationary in time), or very slow, Liouville's theorem ensures that associated conditional probabilities are inviariant by the mechanical evolution equations.

In short, the above assumption simply states that the physical system is sufficiently "chaotic", so that, other things being equal, the system is distributed according to an invariant distribution of its dynamics that has been identified.

6.5 Mean-field model of interacting particles

In the present notes, we will consider two type of macroscopic variables

- Total energy, which is conserved by Hamilton's equations
- A generalized version of volume, parametrized by $l \in L$, for instance the length of a piston.

We will also restrict our study to interacting particles that have a simple structure called mean-field.

We will also consider the following simplifying setting/notations.

- The phase-space of particles is denoted by

$$F.$$

If the "volume" of the box is parametrized by an external device $l \in L$, and we have n particles, the state space will become

$$F^n \times L,$$

with L compact.

- in the same way, the full phase-space measure is denoted

$$\pi_0(dx_1) \otimes \ldots \otimes \pi_0(dx_n) \otimes \pi_1(dl)$$

and is assumed to be a probability (for simplicity, this can be quite easily generalized with additional details). We will also assume that

$$L = \mathrm{supp}(\pi_1)$$

A *mean-field* energy function for the particles is an energy function which can be written in the form:

$$H(x_1, \ldots, x_n, l) = \frac{1}{n} \sum_{m=1}^{n} H_{\mathrm{free}}(x_m, l) + \frac{1}{n^2} \sum_{1 \leq m < m' \leq n} H_{\mathrm{int}}(x_m, x_{m'}),$$

with the additional condition

$$H_{\mathrm{int}}(x, x) = \mathrm{cte}.$$

Mean-field means that it can be written in the form of function of the empirical distribution of particles

$$\frac{1}{n} \sum_{m=1}^{n} \delta_{x_m}.$$

We will thus abuse notation and also denote

$$H(\mu, l) \stackrel{\mathrm{def}}{=} \int_F H_{\mathrm{free}}(x, l)\mu(dx) + \frac{1}{2} \int_F \int_F H_{\mathrm{int}}(x, x')\mu(dx)\mu(dx')$$

for a probability μ so that

$$H(\frac{1}{n} \sum_{m=1}^{n} \delta_{x_m}, l) = H(x_1, \ldots, x_n, l).$$

Physically, this requires a preliminary adimensioning and a regime justifying the $\frac{1}{n^2}$ term in front of the interaction (in particular, usual solid/liquid/gas phases cannot be studied with such mean-field models). Note also that at the macro level, the total energy of the particle system is scaled so that it is of order 1 when n is large.

The exterior device is also associated to an energy

$$H_{\mathrm{dev}}(l, p)$$

wher p is a generalized version of pressure. One should think as l the length of piston with a mass on the top of it which is attracted by gravity. H_{dev} is then the total energy of the mass, which exerts a constant pressure on the gas inside the piston.

6.6 Ensembles

We can then define some usual ensembles of statistical mechanics with our setting.

Definition 6.6.1. *The* micro-canonical *ensemble with volume parameter l and energy interval $[e_1, e_2]$ is the probability distribution on F^n defined by the phase-space measure conditioned by total particle energy in the interval:*

$$H(x_1, \dots, x_n, l) \in [e_1, e_2].$$

Definition 6.6.2. *The* isobare micro-canonical *ensemble with pressure parameter p and energy interval $[e_1, e_2]$ is the probability distribution on $F^n \times L$ defined by the phase-space measure conditioned by total particle + device energy in the interval:*

$$H(x_1, \dots, x_n, l) + H_{\text{dev}}(l, p) \in [e_1, e_2].$$

Definition 6.6.3. *The* canonical *ensemble with volume parameter l and inverse temperature parameter β is the "Gibbs" probability distribution on $F^n \times L$ defined by the density (defined up to a normalizing constant)*

$$\exp\left(-\beta n H(x_1, \dots, x_n, l)\right).$$

The density is meant with respect to the phase-space measure $\pi_0^{\otimes n}$.

Definition 6.6.4. *The* isobare canonical *ensemble with fixed pressure parameter p and inverse temperature parameter β is the probability distribution on $F^n \times L$ defined by the density (defined up to a normalizing constant)*

$$\exp\left(-\beta n \left[H(x_1, \dots, x_n, l) + H_{\text{dev}}(l, p)\right]\right).$$

The density is meant with respect to the phase-space measure $\pi_0^{\otimes n} \otimes \pi_1$

The main principle of statistical mechanics is then the following

Principle of rigorous statistical mechanics Large deviation theory enables to obtain from the ensembles above entropic variational problems satisfied by the density of particles (and the device state l in the isobare case). Under some conditions that will be studied below, those variational problems are equivalent (equivalence of ensembles) and yields usual equilibrium thermodynamic relations.

Exercise 6.6.5. Write down formally the variational problems associated with the thermodynic limit $n \to +\infty$ of the different ensembles.

Remark 6.6.6. In the same way as micro-canonical ensembles are left invariant by deterministic Hamiltonian dynamics, one can construct Markovian perturbation that leave canonical ensembles invariant. Exemple: Langevin equation.

6.7 Thermodynamic limits

In this section, we will give simple assumptions under which one can study the many particle limit $n \to +\infty$. More precisely, one considers the random empirical distribution in $\mathcal{P}(F)$ defined by the above ensembles, and one does show they satisfy the Gibbs conditioning principle, in an appropriate sense.

This limits, called thermodynamical limits, will be quite straightforward to prove from Sanov theorem (with Gibbs conditioning principle) and Varadhan lemmas under the following assumption

Assumption (1)**.** *The functions*

$$(\mu, l) \mapsto H(\mu, l)$$

and for each p

$$l \mapsto H_{\mathrm{dev}}(l, p)$$

are continuous and bounded on $\mathcal{P}(F) \times L$ *or* L*.* $\mathcal{P}(F)$ *is endowed with convergence in distribution.*

To avoid degeneracy we may also ask that

Assumption (2)**.** *Define*

$$e_{\min}(l) \stackrel{\mathrm{def}}{=} \inf_{\mu \in \mathcal{P}(F),\, \mathrm{Ent}(\mu|\pi_0) < +\infty} H(\mu, l),$$

and

$$e_{\max}(l) \stackrel{\mathrm{def}}{=} \sup_{\mu \in \mathcal{P}(F),\, \mathrm{Ent}(\mu|\pi_0) < +\infty} H(\mu, l)$$

One has

$$e_{\min}(l) < H(\pi_0, l) \stackrel{\mathrm{def}}{=} e_0(l) < e_{\max}(l).$$

Exercise 6.7.1. Compute the solutions of the micro-canonical case for $e < e_{\min}(l)$, $e = e_0$, and $e < e_{\max}(l)$.

For that purpose, we consider n particles and the device position as random variables and denoted

$$Y_1, \ldots, Y_n, \Lambda \in F^N \times L.$$

We assume that they are distributed according to one of the ensembles. We can then consider the random variables

$$\left(\Pi_n \stackrel{\mathrm{def}}{=} \frac{1}{N} \sum_{m=1}^{n} \delta_{Y_m}, \Lambda \right) \in \mathcal{P}(F) \times L,$$

and study its limit when $n \to +\infty$. The latter will be concentrated in the following solution sets. For the micro-canonical case:

$$\mathcal{M}_{e,l} \stackrel{\mathrm{def}}{=} \operatorname*{arginf}_{\mu : H(\mu,l) = e} \mathrm{Ent}(\mu|\pi_0),$$

isobare micro-canonical:

$$\mathcal{M}_{\bar{e},p} \stackrel{\text{def}}{=} \operatorname*{arginf}_{\mu,l:H(\mu,l)+H_{\text{dev}}(l,p)=\bar{e}} \text{Ent}(\mu|\pi_0),$$

canonical:

$$\mathcal{M}_{\beta,l} \stackrel{\text{def}}{=} \operatorname*{arginf}_{\mu} \left[\text{Ent}(\mu|\pi_0) + \beta H(\mu,l)\right],$$

isobare canonical:

$$\mathcal{M}_{\beta,p} \stackrel{\text{def}}{=} \operatorname*{arginf}_{\mu,l} \left[\text{Ent}(\mu|\pi_0) + \beta H(\mu,l) + \beta H_{\text{dev}}(l,p)\right].$$

Lemma 6.7.2. *The sets of minimizers above are compact sets. The minimizers of the canonical ensembles are non void. The minimizers of the micro-canonical ensembles are non void in some energy interval.*

Exercise 6.7.3. Proof. Determined the sets where the solution sets are non void.

Proposition 6.7.4 (Micro-canonical)**.** *Let $e \in [e_{\min}(l), e_{\max}(l)]$. Assume that $(Y_1, \ldots, Y_n)$ are distributed according to the micro-canonical ensemble (which is well-defined) with energy constrained in $[e-\delta e, e+\delta e]$ with $\delta e > 0$ and fixed $l \in L$. Let O be any open set with*

$$\mathcal{M}_{e,l} \subset O \subset \mathcal{P}(F),$$

then

$$\lim_{\delta e \to 0} \lim_{n \to +\infty} \frac{1}{n} \ln \mathbb{P}\left[\Pi_n \notin O\right] < 0$$

Proof. Step 0: Apply the LDP lower bound to check that the micro-canonical conditioning indeed makes sense. Step 1: Prove that if I is good and lower semi-continuous, and $F = \cap_k O^k$ is a closed set that is also the intersection of a sequence of open sets O^k, then:

$$\inf_F I = \lim_{k \to +\infty} \inf_{\cap_{k' \leq k} O^{k'}} I.$$

Step 2: Apply Sanov theorem. Step 3: Use the goodness of the rate function to check that the infimum on O^c is strictly greater than the total infimum. □

in the same way, we next define

$$\bar{e}_{\min}(p) \stackrel{\text{def}}{=} \inf_{l \in L} e_{\min}(l) + H_{\text{dev}}(l,p).$$

and

$$\bar{e}_{\max}(p) \stackrel{\text{def}}{=} \sup_{l \in L} e_{\max}(l) + H_{\text{dev}}(l,p).$$

Proposition 6.7.5 (Isobare Micro-canonical)**.** *Let $e \in [\bar{e}_{\min}(p), \bar{e}_{\max}(p)]$. Assume that $(Y_1, \ldots, Y_n)$ are distributed according to the isobare micro-canonical ensemble (which is well-defined) with total energy constrained in $[\bar{e}-\delta e, \bar{e}+\delta e]$ and fixed $l \in L$. Let O be any open set with*

$$\mathcal{M}_{\bar{e},p} \subset O \subset \mathcal{P}(F) \times L,$$

then

$$\lim_{\delta e \to 0} \lim_{n \to +\infty} \frac{1}{n} \ln \mathbb{P}\left[(\Pi_n, \Lambda) \notin O\right] < 0$$

Proposition 6.7.6 (Canonical)**.** *Assume that* $(Y_1, \dots, Y_n)$ *are distributed according to the canonical ensemble with inverse temperature* $\beta \in \mathbb{R}$ *and fixed* $l \in L$*. Let* O *be any open set with*

$$\mathcal{M}_{\beta,l} \subset O \subset \mathcal{P}(F),$$

then

$$\lim_{n\to+\infty} \frac{1}{n} \ln \mathbb{P}\left[\Pi_n \notin O\right] < 0$$

Proposition 6.7.7 (Isobare Canonical)**.** *Assume that* $(Y_1, \dots, Y_n)$ *are distributed according to the isobare canonical ensemble with inverse temperature inverse temperature* $\beta \in \mathbb{R}$ *and fixed* $l \in L$*. Let* O *be any open set with*

$$\mathcal{M}_{\beta,p} \subset O \subset \mathcal{P}(F) \times L,$$

then

$$\lim_{n\to+\infty} \frac{1}{n} \ln \mathbb{P}\left[(\Pi_n, \Lambda) \notin O\right] < 0$$

Define the constant energy (negative) entropy as:

$$s(e,l) \stackrel{\text{def}}{=} \inf_{\mu: H(\mu,l)=e} \text{Ent}(\mu|\pi_0).$$

Exercise 6.7.8. Re-write the other minimization problems using the constant energy entropy. Comment the link with convex duality.

The exercice above justifies the introduction of the constant 'volume' free energy as:

$$f(\beta,l) \stackrel{\text{def}}{=} \inf_{e\in\mathbb{R}} e + \frac{1}{\beta} s(e,l)$$

as well as the Gibbs (isobare) free energy

$$g(\beta,p) \stackrel{\text{def}}{=} \inf_{e\in\mathbb{R}, l\in L} e + H_{\text{dev}}(p,l) + \frac{1}{\beta} s(e,l).$$

Exercise 6.7.9. Re-write the two free energies above using a minimizing solution of the two canonical ensembles.

We can then compare the solutions of the following ensembles as follows.

Proposition 6.7.10. *The set of minimizers of the isobare canonical problem is included in both the canonical and the isobare micro-canonical, which are both included in the micro-canonical.*

In the result below, we will make the simplifying assumption that

$$l = (v,m) \in L = [v_{\text{min,v}_{\text{max}}}] \times [-r,+r]$$

and $H(\mu,l)$ depends on l only through the volume v. We also assume that the device total energy is

$$H_{\text{dev}}(p,l) = pv + \frac{1}{2}m^2, \tag{10}$$

the potential energy being pressure times volume. We set

$$s(e,v) = +\infty \qquad v \notin [v_{\min}, v_{\max}].$$

Exercise 6.7.11. Interpret in terms of a piston model. Show that for the isobare canonical problem, the device momentum is 0, $m = 0$, and the minimization problem is simplified.

Proposition 6.7.12. *Assume* (10). *Then the two dimensional Legendre transform in* $\mathbb{R}^2$ *of micro-canonical entropy* $(e, v) \mapsto s(e, v)$ *is given by* $\tilde{g}(\beta, \gamma)$ *the Gibbs free energy under the change of variable:*

$$\tilde{g}(\beta, \gamma) = \beta g(\beta, \gamma/\beta).$$

First-order phase transition are points were $s^{**}(e, v) \neq s(e, v)$.

Remark 6.7.13. In the same way, the function

$$\beta \mapsto \beta f(\beta, v)$$

is the Legendre transform of constant energy entropy $e \mapsto s(e, l)$ for the conjugate variable pair (e, β).

The Gibbs free energy is then the Legendre transform of free energy for the conjugate variable pair (v, p).

6.8 Equivalence of ensembles and the convex case

The reader may be familiar with the usual thermodynamical formalism which enables to switch equivalently from representation of thermodynamical quantities (entropy, energy, free energy etc..) using energy e and volume v variable, to (inverse) temperature β and pressure p variables.

This is only possible when the thermodynamical minimization problems of the different ensembles are equivalent, which is not true in general (we lose information by taking the Legendre transform of $s(e, v)$.

In order to obtain this equivalence property, we will ask some convexity property of the energy function. First we assume that the device energy is of the form

$$H_{\text{dev}}(v, m, p) = pv + \frac{1}{2}m^2.$$

Then, we will ask

Assumption (3)**.**

$$(\mu, v) \mapsto H(\mu, v)$$

is a convex function, and for each μ *and* p*, there a unique minimizer of the function*

$$v \mapsto H(\mu, v) + pv.$$

We obtain

Proposition 6.8.1 (Equivalence of ensembles)**.** *Under Assumptions 1, 2, 3, there exists a unique a minimizer of each of the four minmization problems (the micro-canonical energy being taken in* $]e_{\min}, e_{\max}[$ *and* $]\bar{e}_{\min}(p), \bar{e}_{\max}(p)[$*, or if required on the boundary of such). Moreover, each solution one of the four minimization problem is solution of any of other four.*

Proof. Step 1: Study first the canonical minimization problem. Show that the unique solution μ_β is continuous with β, hence also the canonical energy $e(\beta) \stackrel{\text{def}}{=} H(\mu_\beta)$.

Step 2: Check that $e(\beta)$ and $s(\beta)$ are monotonous functions of β.

Step 3: Compute the canonical energy for $\beta = 0, +\infty, -\infty$.

Step 4: Show that for any $e = e(\beta)$, μ_e is solution of the micro-canonical problem with energy e. Conclude on equivalence of ensembles for fixed volume v.

Step 5: Redo the same work for the two isobare ensembles.

Step 6: Redo the same work for the two canonical ensembles, with p playing the role of β.

□

Note that we have proven that energy and entropy are monotone functions of β (either fixed volume or fixed pressure). We also have proved that volume and free energy are monotone functions of pressure (fixed temperature).

6.9 Two dimensional convexity of entropy

We will prova that $(e, v) \mapsto s(e, v)$ is a convex function. We can state the main proposition:

Proposition 6.9.1. *Assume that $v \mapsto H(\pi_0, v)$ is linear, as well as Assumptions 1, 2, 3. Then the micro-canonical (negative) entropy*

$$(e, v) \mapsto s(e, v)$$

is a lower semi-continuous convex function of the pair variable (e, l). Moreover it is strictly convex outside its one dimensional minimum at (π_0, v), $v \in \mathbb{R}$.

Proof. Step 1: Show that for $e \leq H(\pi_0, l)$

$$s(e, v) = \inf_{H(\mu, v) \leq e} \mathrm{Ent}(\mu | \pi_0)$$

Step 2: Prove the following lemma

Lemma 6.9.2. *Let I and H be convex. Then*

$$y \mapsto \inf_{x:\, H(x,y) \leq 0} I(x, y)$$

is convex.

Step 3:

Lemma 6.9.3. *Let I be lower semi-continuous and good, and let H be lower semi-continuous then the function*

$$y \mapsto \inf_{x:\, H(x,y) \leq 0} I(x, y)$$

is lower semi-continuous.

Proof. Check it for sequences by extracting a sub-sequence that converges in $E \times F$. □

Step 4: Prove strict convexity using the equivalence of ensembles. □

Exercise 6.9.4. Recover the monotony of the different functions using this convexity property (make a picture). Interpret the change of coordinates using Legendre transform.

Exercise 6.9.5. Recover the various formulas one can find in the literature about thermodynamics for instance:

$$\partial_e s(e, v) = -\beta$$

or

$$\partial_v f(\beta, v) = -p$$

And so on...

References

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2013.

V. Bogachev. *Measure Theory*. Number vol. 1 & 2. Springer Berlin Heidelberg, 2007.

H. Brézis. *Analyse fonctionnelle: théorie et applications*. Collection Mathématiques appliquées pour la maîtrise. Masson, 1987.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Applications of mathematics. Springer, 1998.

P. Dupuis and R. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics. Wiley, 1997.

J. Feng and T. Kurtz. *Large Deviations for Stochastic Processes*. Mathematical Surveys and Monographs. American Mathematical Society, 2015.

M. Freidlin, J. Szücs, and A. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012.

A. Kechris. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer New York, 2012.

A. S. Kechris. Topology and descriptive set theory. *Topology and its Applications*, 58(3):195–222, 1994.

F. Rassoul-Agha and T. Seppäläinen. *A Course on Large Deviations with an Introduction to Gibbs Measures*. Graduate Studies in Mathematics. American Mathematical Society, 2015.

H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1-3):1–69, 2009.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.