Crash course in Large Deviation Theory

J. M. Swart
March 2, 2001[1]

/wisk/LDP/crash.tex

Contents

Preface

This crash course in Large Deviation Theory is intended for those who possess some basic knowledge about topology and measure theory. The reader is adviced to skip Sections 1.4 and 1.5 on first reading, and immediately start in Section 2, to see some non-trivial examples of the theory at work. One can then later page back to Section 1.5, and, in an even later stage, to Section 1.4.

1 General Theory

1.1 Measures on Polish spaces

By definition, a Polish space is a 'complete metrizable' separable topological space. With 'complete metrizable' we mean that it is possible to define a metric on $\mathcal{X}$ that generates the topology, in such a way, that $\mathcal{X}$ is complete in this metric.

Let $\mathcal{C}_b(\mathcal{X})$ be the Banach space of bounded continuous real functions on $\mathcal{X}$, equipped with the supremumnorm $\|\cdot\|_\infty$. Let $\mathcal{B}(\mathcal{X})$ be the Borel-σ-field on $\mathcal{X}$, i.e., the σ-field generated by the open subsets of $\mathcal{X}$. We write $\mathcal{M}(\mathcal{X})$ for the space of finite measures on

[1]With corrections from September 20, 2001

$\mathcal{X}$ and denote the space of probability measures on $\mathcal{X}$ by $\mathcal{M}_1(\mathcal{X})$. If f is a measurable real function on $\mathcal{X}$ and μ a measure on $\mathcal{X}$, we write

$$\langle \mu, f \rangle := \int f \, d\mu$$

for the integral of f with respect to μ, whenever it is defined. If f is nonnegative, then $\rho(dx) := f(x)\mu(dx)$ denotes the measure given by $\rho(A) := \int_A f \, d\mu$. By definition

$$\mathrm{supp}(\mu) := \bigcap \{A \text{ closed} : \mu(A^{\mathrm{c}}) = 0\},$$

is the support of μ. This is the unique smallest closed set on which μ is concentrated.

Every finite measure on $(\mathcal{X}, \mathcal{B})$ has the following property.

$$\forall \varepsilon > 0 \; \exists K \subset \mathcal{X} \text{ compact, such that } \mu(K^{\mathrm{c}}) \leq \varepsilon. \tag{1}$$

We are in particular interested in probability measures on $\mathcal{X}$.

Theorem 1.1 (Topology of weak convergence) *There exists a metric d on $\mathcal{M}_1(\mathcal{X})$ with the following properties.*[2]

1. *$(\mathcal{M}_1(\mathcal{X}), d)$ is a complete metric separable space.*
2. *$d(\mu_n, \mu) \to 0$ if and only if $\langle \mu_n, f \rangle \to \langle \mu, f \rangle$ for all $f \in \mathcal{C}_b(\mathcal{X})$.*

We call the topology generated by d the topology of weak convergence.

Theorem 1.2 (Prohorov) *A set $\mathcal{N} \subset \mathcal{M}_1(\mathcal{X})$ is compact if and only if $\mathcal{N}$ is closed and*

$$\forall \varepsilon > 0 \; \exists K \subset \mathcal{X} \text{ compact, such that } \mu(K^{\mathrm{c}}) \leq \varepsilon \; \forall \mu \in \mathcal{N}. \tag{2}$$

In particular, if $\mathcal{X}$ is compact then $\mathcal{M}_1(\mathcal{X})$ is compact. Note that (2) states that the compact set K in (1) can be chosen uniform in $\mathcal{N}$. A set or sequence of probability measures on $\mathcal{X}$ satisfying (2) is called tight. Prohorov's theorem states that tighness is equivalent to relative compactness. In particular, every tight sequence of probability measures (on a Polish space) has a convergent subsequence.

We write $\mu_n \Rightarrow \mu$ to denote weak convergence of probability measures. One can show that $\mu_n \Rightarrow \mu$ if and only if one of the following two equivalent conditions holds.

$$\begin{array}{rll} \text{(i)} & \limsup\limits_{n\to\infty} \mu_n(C) \leq \mu(C) & \forall C \text{ closed} \\ \text{(ii)} & \liminf\limits_{n\to\infty} \mu_n(O) \geq \mu(O) & \forall O \text{ open.} \end{array}$$

In fact, it suffices to check (ii) for all $O \in \mathcal{O}'$, where $\mathcal{O}'$ is a basis of the topology.

[2] One metric with these properties is for example the Prohorov metric. This metric is given by

$$d(\mu_1, \mu_2) = \inf\{\varepsilon \geq 0 : P[d(X_1, X_2) \geq \varepsilon] \leq \varepsilon, \; \mathcal{L}(X_1) = \mu_1, \; \mathcal{L}(X_2) = \mu_2\},$$

where the infimum is over all $\mathcal{X} \times \mathcal{X}$-valued random variables (X_1, X_2) such that the law of X_i is μ_i.

1.2 Lower (upper) semicontinuous functions

Let $\mathcal{X}$ be a metric space and let $\phi : \mathcal{X} \to (-\infty, \infty]$. We say that ϕ is lower semicontinuous if one of the following equivalent conditions holds

$$\begin{array}{ll} \text{(i)} & \{\phi \leq \alpha\} \text{ is closed for all } \alpha \in \mathbb{R} \\ \text{(ii)} & x_n \to x \text{ implies } \liminf_{n\to\infty} \phi(x_n) \geq \phi(x). \end{array}$$

A l.s.c. function assumes its minimum on (nonempty) compact sets. If ϕ_1, ϕ_2 are l.s.c. and $\lambda_1, \lambda_2 \geq 0$, then $\lambda_1\phi_1 + \lambda_2\phi_2$ is l.s.c. If ϕ_1, ϕ_2 are l.s.c. and nonnegative, then $\phi_1\phi_2$ is l.s.c. If (ϕ_i) is a sequence of l.s.c. functions and $\phi_i \uparrow \phi$, then ϕ is l.s.c. An indicator function 1_A is l.s.c. if and only if A is open.

A function $\phi : \mathcal{X} \to [-\infty, \infty)$ is called upper semicontinuous (u.s.c.) if $-\phi$ is l.s.c. The properties of u.s.c. functions are easily derived from those of l.s.c. functions.

1.3 The Large Deviation Principle

The Large Deviation Principle (LDP) gives a precise mathematical meaning to the statement that the probability that a certain random object takes values near a certain point decays with a certain exponential rate. Thus, we want to give a meaning to the statement that, roughly

$$P[X_n \in dx] \sim e^{-nI(x)} \qquad \text{as } n \to \infty, \tag{3}$$

where the X_n are stochastic variables. In order to get a sensible definition, it turns out that we must assume that the function I satisfies the following condition:

$$\{I \leq \alpha\} \text{ is compact for all } \alpha \in \mathbb{R}.$$

We describe this by saying that I has compact level sets. Clearly, such a function is lower semicontinuous. A function with compact level sets assumes its minimum on nonempty closed sets; in particular, since $\mathcal{X}$ is closed, such a function is bounded from below.

Definition 1.3 (Rate function) A good rate function is a function $I : \mathcal{X} \to (-\infty, \infty]$ with compact level sets. A rate function is a lower semicontinuous function $I : \mathcal{X} \to (-\infty, \infty]$ that is bounded from below. ◊

Definition 1.4 (Large Deviation Principle) A sequence (μ_n) of finite measures on a Polish space $\mathcal{X}$ satisfies the LDP with rate function I, if *both* of the following two conditions hold.

$$\begin{array}{llll} \text{(i)} & \displaystyle\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(C) \leq -\inf_{x\in C} I(x) & & \forall C \text{ closed} \\ \text{(ii)} & \displaystyle\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) \geq -\inf_{x\in O} I(x) & & \forall O \text{ open.} \end{array}$$

◊

Remark I If the μ_n are probability measures, then $\inf_{x\in\mathcal{X}} I(x) = 0$, as follows from the fact that $\mathcal{X}$ is both open and closed.

Remark II It is sufficient for (ii) if the condition holds for all $O \in \mathcal{O}'$, where $\mathcal{O}'$ is a basis of the topology.

Remark III If A is a measurable set such that $\inf_{x\in A^\circ} I(x) = \inf_{x\in\overline{A}} I(x)$, then (i) and (ii) imply that

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = -\inf_{x\in A} I(x).$$

Such a set is called an I-continuity set.

Lemma 1.5 (Varadhan-Bryc) *A sequence (μ_n) of finite measures on a Polish space $\mathcal{X}$ satisfies the LDP with the good rate function I, if and only if*

$$(*)\quad \lim_{n\to\infty} \frac{1}{n} \log\langle \mu_n, e^{nf}\rangle = \sup_{x\in\mathcal{X}} \Big\{ f(x) - I(x) \Big\}$$

for every $f : \mathcal{X} \to [-\infty, \infty)$ that is continuous and bounded from above.

At this point, the reader may wonder whether our definition really corresponds to the vague notion expressed in (3). We hope that this becomes clearer after the following lemma, which provides us with some examples of (μ_n) satisfying a LDP.

Lemma 1.6 (Sufficient conditions for LDP) *Let (σ_n) be a sequence of finite measures on a Polish space $\mathcal{X}$, converging weakly to a limit σ with compact support, and assume that* $\mathrm{supp}(\sigma_n) \subset \mathrm{supp}(\sigma)$ *for all n. Let $J_n, J : \mathcal{X} \to (-\infty, \infty]$ be continuous and bounded from above, and assume that*

$$\lim_{n\to\infty} \| e^{-J_n} - e^{-J} \|_\infty = 0$$

Then the measures (μ_n) given by $\mu_n(dx) := e^{nJ_n(x)}\sigma_n(dx)$ satisfy the LDP with the good rate function I given by $I(x) := J(x)$ if $x \in \mathrm{supp}(\sigma)$ and $I(x) := \infty$ otherwise.

Example We can take $\mathcal{X} := [0, 1]$ and $\sigma_n := \frac{1}{n}\sum_{i=1}^n \delta_{i/n}$ so that $\sigma_n \Rightarrow \sigma$, where σ is the Lebesgue measure. Now let I be any continuous function on $[0, 1]$. Then the measures $\mu_n(dx) := e^{-nI(x)}\sigma_n(dx)$ satisfy the LDP with good rate function I. In this example, every open set is an I-continuity set and hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) = -\inf_{x\in O} I(x) \quad \forall O \text{ open}.$$

To understand intuitively why this is so, rewrite the formula as

$$\mu_n(O) = e^{-n \inf_{x\in O} I(x) + o(n)} \qquad \text{as } n \to \infty..$$

When n is large, the main contribution to $\mu_n(O)$ comes from the part of O where the rate function I is lowest. The measure of other parts of O decays at a faster exponential rate and is therefore negligible in the limit.

1.4 Intermezzo: convergence of L^p-norms

In this section, which can be skipped at first reading, we sketch the proof of some of the lemmas in the last section. We show that the fact that a LDP holds for a sequence of finite measures (μ_n) can be translated into a statement about the convergence of certain L^p-norms associated with these measures.

As before, let $\mathcal{X}$ be a Polish space. Let $B(\mathcal{X})$ be the space of bounded measurable real functions on $\mathcal{X}$. If μ is a finite measure on $\mathcal{X}$, $p \geq 1$ and $\phi \in B(\mathcal{X})$, then we write

$$\|\phi\|_{p,\mu} := \langle \mu, |\phi|^p \rangle^{1/p}$$

for the L^p-norm associated with μ. By the continuity of measures, we have for nonnegative $\phi_i, \phi \in B(\mathcal{X})$:

$$\begin{array}{ll} \text{(i)} & \phi_i \downarrow \phi \text{ implies } \|\phi_i\|_{p,\mu} \downarrow \|\phi\|_{p,\mu} \\ \text{(ii)} & \phi_i \uparrow \phi \text{ implies } \|\phi_i\|_{p,\mu} \uparrow \|\phi\|_{p,\mu}. \end{array}$$

Now let I be a good rate function on $\mathcal{X}$. Define, for $\phi \in B(\mathcal{X})$:

$$\|\phi\|_{\infty,I} := \|e^{-I}\phi\|_\infty.$$

One can check that $\|\cdot\|_{\infty,I}$ is a norm on $B(\mathcal{X})$ and that

$$\begin{array}{lll} \text{(i)} & \phi_i \downarrow \phi \text{ implies } \|\phi_i\|_{\infty,I} \downarrow \|\phi\|_{\infty,I} & \text{if the } \phi_i, \phi \text{ are nonnegative, bounded, u.s.c.} \\ \text{(ii)} & \phi_i \uparrow \phi \text{ implies } \|\phi_i\|_{\infty,I} \uparrow \|\phi\|_{\infty,I} & \text{if the } \phi_i, \phi \text{ are nonnegative, measurable.} \end{array}$$

In the proof of (i), one needs the fact that I is a good rate function.

Now let (μ_n) be a sequence of finite measures on $\mathcal{X}$ and let I be a rate function. It is just a little calculation to check that the (μ_n) satisfy the LDP with rate function I if and only if both of the following conditions hold.

$$\begin{array}{lll} \text{(i)} & \limsup_{n\to\infty} \|1_C\|_{n,\mu_n} \leq \|1_C\|_{\infty,I} & \forall C \text{ closed} \\ \text{(ii)} & \liminf_{n\to\infty} \|1_O\|_{n,\mu_n} \geq \|1_O\|_{\infty,I} & \forall O \text{ open.} \end{array}$$

Lemma 1.5, translated into this language, says that (i) and (ii) are equivalent to

$$(*) \ \lim_{n\to\infty} \|\phi\|_{n,\mu_n} = \|\phi\|_{\infty,I} \qquad \forall \phi \in \mathcal{C}_b(\mathcal{X}).$$

(To see this, note that $\frac{1}{n}\log\langle \mu_n, e^{nf}\rangle = \log \|e^f\|_{n,\mu_n}$. If $f : \mathcal{X} \to [-\infty, \infty)$ is continuous and bounded from above, then $\phi := e^f$ is nonnegative, bounded and continuous.)

In the proof of these statements one needs the continuity of the norm $\|\cdot\|_{\infty,I}$ with respect to increasing and descreasing sequences. For the proof that (i) and (ii) imply $(*)$, one uses that l.s.c. (u.s.c.) functions can be approximated from below (above) by simple l.s.c. (u.s.c.) functions. One can then prove that (i) and (ii) imply

$$\begin{array}{lll} \text{(i}'\text{)} & \limsup_{n\to\infty} \|\phi\|_{n,\mu_n} \leq \|\phi\|_{\infty,I} & \forall \phi \text{ nonnegative bounded u.s.c.} \\ \text{(ii}'\text{)} & \liminf_{n\to\infty} \|\phi\|_{n,\mu_n} \geq \|\phi\|_{\infty,I} & \forall \phi \text{ nonnegative bounded l.s.c.,} \end{array}$$

which implies $(*)$. To show that $(*)$ implies (i) and (ii), one uses the fact that for every closed set C there exist continuous functions $\phi_i : \mathcal{X} \to [0,1]$ such that $\phi_i \downarrow 1_C$. Using the continuity of the norms, we then have

$$\begin{aligned}\limsup_{n\to\infty} \|1_C\|_{n,\mu_n} &= \limsup_{n\to\infty} \inf_i \|\phi_i\|_{n,\mu_n} \\ &\le \inf_i \lim_{n\to\infty} \|\phi_i\|_{n,\mu_n} = \inf_i \|\phi_i\|_{\infty,I} = \|1_C\|_{\infty,I}.\end{aligned}$$

The proof that $(*)$ implies (ii) is analogue.

1.5 Manipulations of LDP's

Once we know that a LDP holds for a sequence of measures (μ_n), this information can frequently be used to prove LDP's for other sequences of measures. The following is a simple consequence of Varadhan's lemma (Lemma 1.5).

Proposition 1.7 (Transformation through an exponential density) *Let (μ_n) be a sequence of probability measures on a Polish space $\mathcal{X}$, satisfying the LDP with rate function I. Assume that $f : \mathcal{X} \to [-\infty, \infty)$ is continuous and bounded from above. Let $Z_n := \langle \mu_n, e^{nf} \rangle$. Then the probability measures*

$$\tilde{\mu}_n(dx) := \frac{1}{Z_n} e^{nf(x)} \mu_n(dx)$$

satisfy the LDP with rate function $\tilde{I}$ given by

$$\tilde{I}(x) = I(x) - f(x) - \inf_{y\in\mathcal{X}} \Big\{ I(y) - f(y) \Big\}.$$

Remark I The function $\tilde{I}$ need in general not be a good rate function.

Remark II If the (μ_n) are finite measures and $\mu_n'(dx) = e^{nf(x)}\mu_n(dx)$, then the (μ_n') satisfy the LDP with rate function $I' := I - f$. Note that $\tilde{I}$ above is just $I - f$ 'normalized' so that $\inf_{x\in\mathcal{X}} \tilde{I}(x) = 0$.

Proposition 1.8 (Contraction Principle) *Let (μ_n) be a sequence of finite measures on a Polish space $\mathcal{X}$, satisfying the LDP with good rate function I. Let $\mathcal{Y}$ be a Polish space and let $\psi : \mathcal{X} \to \mathcal{Y}$ be a continuous map. Then the induced measures $\mu_n \circ \psi^{-1}$ on $\mathcal{Y}$ satisfy the LDP with the good rate function*

$$J(y) = \inf_{\substack{x\in\mathcal{X} \\ \psi(x)=y}} I(x).$$

Remark I Since ψ is continuous, $\psi^{-1}(\{y\})$ is closed. Now I is a good rate function, which assumes its minimum on closed sets, and hence

$$J(y) = \begin{cases} \min_{x\in\psi^{-1}(\{y\})} I(x) & \text{if } \psi^{-1}(\{y\}) \neq \emptyset \\ \infty & \text{otherwise.} \end{cases}$$

Remark II For those who read the intermezzo on L^p-norms: it is easy to see that $\|\phi\|_{n,\mu_n\circ\psi^{-1}} = \|\phi \circ \psi\|_{n,\mu_n}$ and $\|\phi\|_{\infty,J} = \|\phi \circ \psi\|_{\infty,I}$. Compare J to a measure induced by a measurable map.

2 LDP's for iid random variables

2.1 Relative entropy

Let μ, ν be probability measures on a measurable space $(\Omega, \mathcal{F})$ and assume that $\nu \ll \mu$. The relative entropy of ν with respect to μ is defined as the quantity

$$H(\nu|\mu) := \int d\nu \log \frac{d\nu}{d\mu},$$

where $d\nu/d\mu$ denotes the Radon-Nikodym derivative of ν with respect to μ. Note that $d\nu/d\mu$ is zero on a set of ν-measure zero, so that we do not have to bother how to define $\log 0$. It is often useful to write $H(\nu|\mu)$ as

$$H(\nu|\mu) = \int d\mu \, \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu}.$$

The function $x \mapsto \phi(x) := x \log x$ is convex, and hence by Jensen's inequality we have $\langle \mu, \phi(d\nu/d\mu) \rangle \geq \phi(\langle \mu, d\nu/d\mu \rangle) = \phi(1) = 0$. Thus $H(\nu|\mu) \geq 0$. In fact, since ϕ is strictly convex, it is not hard to show that $H(\nu|\mu) = 0$ if and only if $\nu = \mu$. If $\nu \not\ll \mu$ we set $H(\nu|\mu) := \infty$.

2.2 Sanov's Theorem for finite spaces

Let Σ be a finite space, and let $\mathcal{M}_1(\Sigma)$ be the space of probability measures on Σ. We may identify a measure $\pi \in \mathcal{M}_1(\Sigma)$ with the vector $(\pi(x))_{x \in \Sigma} \in \mathbb{R}^{|\Sigma|}$, where $\pi(x) := \pi(\{x\})$. Thus, we identify $\mathcal{M}_1(\Sigma)$ with the *simplex*

$$\mathcal{M}_1(\Sigma) = \Big\{ \pi \in \mathbb{R}^{|\Sigma|} : \pi(x) \geq 0 \ \ \forall x \in \Sigma, \ \sum_{x \in \Sigma} \pi(x) = 1 \Big\},$$

which is equipped with a topology in the obvious way.

Theorem 2.1 (Sanov, finite space case) *Let (X_i) be a sequence of iid Σ-valued random variables with law $\rho \in \mathcal{M}_1(\Sigma)$. Let $L_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ denote the emperical law of the first n random variables. Then the probability measures $\mu_n := \mathcal{L}(L_n)$ on $\mathcal{M}_1(\Sigma)$ satisfy the LDP with good rate function $H(\cdot|\rho)$.*

Before we turn to the proof, let us introduce some additional notiation. Let us call Σ an *alphabet*, and let us call a sequence $\vec{x} = (x_1, \ldots, x_n) \in \Sigma^n$ a *word* of length n. If $\vec{x}$ is a word of length n, then we call

$$\tau_n(\vec{x}) := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$$

the *type* of the word $\vec{x}$. Note that $L_n = \tau_n(X_1, \ldots, X_n)$. By definition, $\mathcal{L}_n \subset \mathcal{M}_1(\Sigma)$ is the space $\mathcal{L}_n := \tau_n(\Sigma^n)$ of all possible types of words of length n. Note that

$$\mathcal{L}_n = \Big\{ \pi \in (\tfrac{1}{n}\mathbb{Z})^{|\Sigma|} : \pi(x) \geq 0 \ \ \forall x \in \Sigma, \ \sum_{x \in \Sigma} \pi(x) = 1 \Big\},$$

where $\frac{1}{n}\mathbb{Z} = \{\frac{1}{n}k : k \in \mathbb{Z}\}$. For any $\pi \in \mathcal{L}_n$, we define the *type class* $T_n(\pi)$ as

$$T_n(\pi) := \tau_n^{-1}(\{\pi\}) = \{\vec{x} \in \Sigma^n : \tau_n(\vec{x}) = \pi\}.$$

For any $\rho \in \mathcal{M}_1(\Sigma)$ and $\pi \in \mathcal{L}_n$, let us write

$$p_n(\pi|\rho) := P[\tau_n(X_1, \ldots, X_n) = \pi],$$

where ρ is the law of the iid random variables (X_i).

Lemma 2.2 (Relative entropy) *For any $\rho \in \mathcal{M}_1(\Sigma)$ and $\pi \in \mathcal{L}_n$, we have*

$$p_n(\pi|\rho) = p_n(\pi|\pi)e^{-nH(\pi|\rho)}.$$

Proof For $\pi \not\ll \rho$ the statement is obvious, if we agree that $e^{-\infty} = 0$. For $\pi \ll \rho$, just write

$$\begin{aligned} p_n(\pi|\rho) &= \sum_{\vec{x} \in T_n(\pi)} P[(X_1, \ldots, X_n) = \vec{x}\,] = \sum_{\vec{x} \in T_n(\pi)} \prod_{i=1}^n \rho(x_i) = \sum_{\vec{x} \in T_n(\pi)} \prod_{x \in \Sigma} \rho(x)^{n\pi(x)} \\ &= |T_n(\pi)| \Big(\prod_{x \in \Sigma} \pi(x)^{n\pi(x)} \Big) \Big(\prod_{x \in \Sigma} (\tfrac{\rho(x)}{\pi(x)})^{n\pi(x)} \Big) = p_n(\pi|\pi) e^{-n \sum_{x \in \Sigma} \pi(x) \log(\frac{\pi(x)}{\rho(x)})}. \end{aligned}$$

∎

Proof of Theorem 2.1 We first treat the case that $\rho(x) > 0$ for all $x \in \Sigma$. In this case, $H(\cdot|\rho)$ is a continuous function on $\mathcal{M}_1(\Sigma)$. We are now going to apply Lemma 1.6. Thus, we are going to show that the measures (μ_n) (which are measures *on*, not *in* $\mathcal{M}_1(\Sigma)$) can be written as

$$\mu_n(d\pi) = e^{nJ_n(\pi)}\sigma_n(d\pi), \tag{4}$$

where the (σ_n) are finite measures converging weakly to some limit σ, and the functions J_n satisfy

$$\lim_{n \to \infty} \|e^{-J_n} - e^{-H(\cdot|\rho)}\|_\infty = 0. \tag{5}$$

Define measures σ_n on $\mathcal{M}_1(\Sigma)$ by

$$\sigma_n := \frac{1}{|\mathcal{L}_n|} \sum_{\pi \in \mathcal{L}_n} \delta_\pi.$$

It is not hard to see that $\sigma_n \Rightarrow \sigma$, where σ is the uniform probability measure on $\mathcal{M}_1(\Sigma)$. (It is illustrative to make a picture of the set $\mathcal{L}_n$ in the case that $|\Sigma| = 3$.)

Since the measures μ_n are concentrated on $\mathcal{L}_n$, it is clear that they have a density with respect to σ_n. In fact, Lemma 2.2 tells us that

$$\mu_n(d\pi) = p_n(\pi|\rho)|\mathcal{L}_n|\sigma_n(d\pi) = |\mathcal{L}_n|p_n(\pi|\pi)e^{-nH(\pi|\rho)}\sigma_n(\pi).$$

It is not hard to see that for any $\pi \in \tilde{\mathcal{L}}_n$:

$$|\tilde{\mathcal{L}}_n|^{-1} \leq p_n(\pi|\pi) \leq 1.$$

The upper bound is obvious since $p_n(\pi|\pi)$ is a probability. For the lower bound, use that $p_n(\tilde{\pi}|\pi) \leq p_n(\pi|\pi)$ for all $\tilde{\pi} \in \mathcal{L}_n$ (which is not hard to show) so that $1 = \sum_{\pi \in \mathcal{L}_n} p_n(\tilde{\pi}|\pi) \leq |\mathcal{L}_n| p_n(\pi|\pi)$. Combining our last two formulas, we get

$$e^{-nH(\pi|\rho)}\sigma_n(\pi) \leq \mu_n(d\pi) \leq |\mathcal{L}_n| e^{-nH(\pi|\rho)}\sigma_n(\pi).$$

It is not hard to see that

$$|\mathcal{L}_n| \leq (n+1)^{|\Sigma|}.$$

Thus, the functions J_n in (4) satisfy

$$H(\pi|\rho) \leq J_n(\pi) \leq H(\pi|\rho) + \frac{1}{n}|\Sigma| \log(n+1).$$

It is now easy to see that condition (5) is satisfied, so that the statement of the theorem follows from Lemma 1.6.

If some of the $\rho(x)$ are zero, define $\tilde{\mathcal{L}}_n := \{\pi \in \mathcal{L}_n : \pi \ll \rho\}$ and, in this case, define the measures σ_n by

$$\sigma_n(d\pi) := \frac{1}{|\tilde{\mathcal{L}}_n|} \sum_{\pi \in \tilde{\mathcal{L}}_n} \delta_\pi.$$

These measures converge weakly to a measure σ with $\mathrm{supp}(\sigma) = \{\pi \in \mathcal{M}_1(\Sigma) : \pi \ll \rho\}$. Now the argument proceeds more or less as before. ∎

2.3 Cramér's Theorem in the finite case

As in the last section, let Σ be a finite set and let (X_i) be a sequence of iid random variables with law $\rho \in \mathcal{M}_1(\Sigma)$. Let $f : \Sigma \to \mathbb{R}$ be some function. Often, we want to minimize the entropy $H(\nu|\rho)$ of a measure $\nu \in \mathcal{M}_1(\Sigma)$, subject to the restraint that $\langle \nu, f \rangle = z$. Here the following lemma is useful.

Lemma 2.3 (Entropy inequality) *Assume that Σ is a finite set, $f : \Sigma \to \mathbb{R}$ and $\rho \in \mathcal{M}_1(\Sigma)$. For $\lambda \in \mathbb{R}$, put*

$$\Lambda(\lambda) := \log\langle \rho, e^{\lambda f} \rangle.$$

Then, for every $\nu \in \mathcal{M}_1(\Sigma)$ and $\lambda \in \mathbb{R}$:

$$\lambda\langle \nu, f \rangle - H(\nu|\rho) \leq \Lambda(\lambda),$$

where equality holds if and only if

$$\nu(x) = e^{\lambda f(x) - \Lambda(\lambda)}\rho(x).$$

Proof Since $x \mapsto \log(x)$ is a concave function, Jensen's inequality gives

$$\lambda\langle\nu, f\rangle - H(\nu|\rho) = \sum_{x\in\Sigma} \nu(x) \log\Big(\frac{e^{\lambda f(x)}\rho(x)}{\nu(x)}\Big) \leq \log \sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x) = \Lambda(\lambda).$$

Using the fact that $x \mapsto \log(x)$ is strictly convex, it is not hard to see that equality holds iff the function $\frac{e^{\lambda f(x)}\rho(x)}{\nu(x)}$ is constant, which means that $\nu(x) = \frac{1}{Z}e^{\lambda f(x)}\rho(x)$. Here the normalizing constant is $Z = \sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x) = e^{\Lambda(\lambda)}$. ∎

We call Λ the *logarithmic moment generating function*, or *cumulant generating function* of the stochastic variable f. For every $\lambda \in \mathbb{R}$, let $\rho^{(\lambda)}$ be the probability measure on Σ given by

$$\rho^{(\lambda)}(x) := e^{\lambda f(x) - \Lambda(\lambda)}\rho(x).$$

Remark Note that by Lemma 2.3

$$\Lambda(\lambda) = \langle\rho^{(\lambda)}, f\rangle - H(\rho^{(\lambda)}|\rho).$$

In statistical mechanics, we meet the situation that f is the energy function, ρ normalized counting measure, $\rho^{(\lambda)}$ the Gibbs measure with inverse temperature λ, and $\langle\rho^{(\lambda)}, f\rangle$ the (expected) energy. In this context, $\Lambda(\lambda)$ is called the *free energy.*

Lemma 2.4 (Properties of the cumulant generating function) *Let Σ be a finite set, $f : \Sigma \to \mathbb{R}$ and $\rho \in \mathcal{M}_1(\Sigma)$. Define $\Lambda(\lambda)$ and $\rho^{(\lambda)}$ as above. Then the function Λ is infinitely differentiable. For every $\lambda \in \mathbb{R}$, we have*

$$\begin{aligned} \Lambda'(\lambda) &= E_{\rho^{(\lambda)}}(f) \\ \Lambda''(\lambda) &= \mathrm{Var}_{\rho^{(\lambda)}}(f). \end{aligned}$$

Proof The first equality follows by writing

$$\Lambda'(\lambda) = \tfrac{\partial}{\partial\lambda} \log \sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x) = \frac{\sum_{x\in\Sigma} f(x)e^{\lambda f(x)}\rho(x)}{\sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x)} = \sum_{x\in\Sigma} f(x)e^{\lambda f(x)-\Lambda(\lambda)}\rho(x),$$

while the second equality follows from

$$\begin{aligned} \Lambda''(\lambda) &= \tfrac{\partial}{\partial\lambda} \frac{\sum_{x\in\Sigma} f(x)e^{\lambda f(x)}\rho(x)}{\sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x)} \\ &= \frac{(\sum_{x\in\Sigma} f(x)^2e^{\lambda f(x)}\rho(x))(\sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x)) - (\sum_{x\in\Sigma} f(x)e^{\lambda f(x)}\rho(x))^2}{(\sum_{x\in\Sigma} e^{\lambda f(x)}\rho(x))^2}. \end{aligned}$$

∎

We see that Λ is a convex function, which assumes its minimum at that value (or, in exceptional cases, values) of λ for which $\langle\rho^{(\lambda)}, f\rangle = 0$ (if such a value exists). In particular, if $\langle\rho, f\rangle = 0$, then $\Lambda(0) = 0$ and Λ is nonnegative.

Theorem 2.5 (Cramér, finite case) *Let (X_i) be a sequence of iid random variables taking values in a finite set Σ and let $f : \Sigma \to \mathbb{R}$. Let*

$$S_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad \textit{and} \quad \mu_n := \mathcal{L}(S_n).$$

Then the (μ_n) satisfy the LDP with rate function

$$\Lambda^*(z) := \sup_{\lambda\in\mathbb{R}}\{\lambda z - \Lambda(\Lambda)\}.$$

Proof The fact that the (μ_n) satisfy an LDP follows from Sanov's theorem and the contraction principle. Define $\psi : \mathcal{M}_1(\Sigma) \to \mathbb{R}$ by $\psi(\pi) := \sum_{x\in\Sigma} f(x)\pi(x)$. Note that ψ is continuous and that $\psi(L_n) = S_n$, where $L_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ is the emperical law of $(X_1, \ldots, X_n)$. Note that

$$\psi^{-1}(\{z\}) = \{\nu \in \mathcal{M}_1(\Sigma) : \langle \nu, f\rangle = z\}.$$

Thus, the contraction principle and Sanov's theorem combine to give the following result.

Lemma 2.6 (Consequence of the contraction principle) *The (μ_n) satisfy the LDP with good rate function J given by*

$$J(z) = \begin{cases} \min_{\nu\in\psi^{-1}(\{z\})} H(\nu|\rho) & \textit{if } \psi^{-1}(\{z\}) \neq \emptyset, \\ \infty & \textit{otherwise.} \end{cases}$$

Proof of Theorem 2.5 (continued) It is left to verify that $J = \Lambda^*$. By Lemma 2.3,

$$H(\nu|\rho) \geq \lambda z - \Lambda(\lambda) \qquad \forall \lambda \in \mathbb{R},\ \nu \in \psi^{-1}(\{z\}),$$

and hence

$$\inf_{\nu\in\psi^{-1}(\{z\})} H(\nu|\rho) \geq \sup_{\lambda\in\mathbb{R}}\{\lambda z - \Lambda(\lambda)\}.$$

Now assume that $z \in (\min_{x\in\Sigma} f(x), \max_{x\in\Sigma} f(x))$. Then it is not hard to see that there exists a $\lambda \in \mathbb{R}$ such that $\langle \rho^{(\lambda)}, f\rangle = z$. By Lemma 2.3, for this λ

$$H(\nu|\rho) = \lambda z - \Lambda(\lambda),$$

and hence

$$\min_{\nu\in\psi^{-1}(\{z\})} H(\nu|\rho) = \max_{\lambda\in\mathbb{R}}\{\lambda z - \Lambda(\lambda)\}.$$

If z is either $\min_{x\in\Sigma} f(x)$ or $\max_{x\in\Sigma} f(x)$, then it is not hard to see that $H(\nu|\rho) = \lim_{\lambda\to-\infty} \lambda z - \Lambda(\lambda)$ or $H(\nu|\rho) = \lim_{\lambda\to\infty} \lambda z - \Lambda(\lambda)$, respectively. Finally, in the case that $z \notin [\min_{x\in\Sigma} f(x), \max_{x\in\Sigma} f(x)]$, one can check that $H(\nu|\rho) = \infty = \sup_{\lambda\in\mathbb{R}}\{\lambda z - \Lambda(\lambda)\}$.$\blacksquare$

2.4 The Gärtner-Ellis theorem

In this section we state, without proof, a version of the Gärtner-Ellis theorem, which is a generalization of Cramér's theorem. Let (X_i) be a sequence of real valued random variables, and define $S_n := \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu_n := \mathcal{L}(S_n)$. If the (X_i) are iid with law $\rho \in \mathcal{M}_1(\mathbb{R})$, then the cumulant generating function Λ_n of S_n is given by

$$\Lambda_n(\lambda) := \log E[e^{\lambda S_n}] = \log E\left[e^{\frac{1}{n}\lambda \sum_{i=1}^n X_i}\right] = \log \prod_{i=1}^n E\left[e^{\frac{1}{n}\lambda X_i}\right] = n(\Lambda(\tfrac{\lambda}{n}))^n,$$

where $\Lambda(\lambda) = E[e^{X_i}]$ is the cumulant generating function of ρ. In particular, we see that $\frac{1}{n}\Lambda_n(n\lambda) = \Lambda(\lambda)$.

Now consider a situation where the (X_i) are not iid, but assume that the limit $\lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda) =: \Lambda(\lambda)$ exists. In such situations, the Gärtner-Ellis theorem is frequently applicable.

Let $\Lambda : \mathbb{R} \to (-\infty, \infty]$ be a convex function. Put $\mathcal{D}_\Lambda := \{x \in \mathbb{R} : I(x) < \infty\}$ and let $\mathcal{D}_\Lambda^\circ$ denote its interior. Put

$$\begin{aligned} r_+ &:= \sup \mathcal{D}_\Lambda \\ r_- &:= \inf \mathcal{D}_\Lambda. \end{aligned}$$

We say that Λ is *essentialy smooth* if

(i) $\mathcal{D}_\Lambda^\circ$ is non-empty

(ii) I is differentiable on $\mathcal{D}_\Lambda^\circ$

(iii) $\lim_{x\uparrow r_+} I'(x) = \infty$ and $\lim_{x\downarrow r_-} I'(x) = -\infty$.

Theorem 2.7 (Gärtner-Ellis) *Let (S_n) be a sequence of real-valued random variables with cumulant generating function Λ_n and put $\mu_n := \mathcal{L}(S_n)$. Assume that*

$$\lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda) = \Lambda(\lambda),$$

for some essentially smooth convex function $\Lambda : \mathbb{R} \to (-\infty, \infty]$ such that $0 \in \mathcal{D}_\Lambda^\circ$. Then the measures (μ_n) satisfy the LDP with good rate function Λ^* given by

$$\Lambda^*(z) := \sup_{\lambda\in\mathbb{R}}\{\lambda z - \Lambda(\lambda)\}.$$

References

[DS89] J.-D. Deuschel and D.W. Stroock, *Large deviations.* Academic Press, Boston, 1989.

[DZ93] A. Dembo and O. Zeitouni, *Large deviations techniques and applications.* Jones and Bartlett Publishers, Boston, 1993.

[DZ98] A. Dembo and O. Zeitouni, *Large deviations techniques and applications 2nd edition*, Applications of Mathematics 38. Springer, New York, 1998.

[Hol00] F. den Hollander, *Large deviations*, Fields Institute Monographs 14. AMS, Providence, 2000.