AN INTRODUCTION TO LARGE DEVIATION THEORY

YU-LIANG WU

Table of Contents

1. Introduction

1.1. **Overview.** The aim of this lecture note is to give a very brief introduction to the theory of large deviations. In probability theory, a recurrent theme is the convergence of random variables, and the law of large numbers (LLN) is among the fundamental results concerning their "typical" behavior. But beyond this scope, mathematicians yearn for a more refined picture of the "atypical configurations" or "rare events" that deviate from the expected outcome either by a small or a large amount. In this pursuit, the former is addressed in the central limit theorem (CLT) while the latter gives rise to the large deviation principle (LDP). To demonstrate these, we would like to begin with the following example.

Example 1.1 (Bernoulli trial)**.** A Bernoulli trial is a series of identical and independent experiments with exactly two possible outcomes, say 0 and 1, which may be modeled as a sequence of i.i.d. random variables $(X_i)_{i=1}^{\infty}$ with the Bernoulli distribution:

$$\mathbb{P}(X_i = 1) = p \quad \text{and} \quad \mathbb{P}(X_i = 0) = 1 - p \quad (p \in (0,1)).$$

The core questions regarding these trials revolve around the number of $1'$s in an n -trial, which is mathematically phrased as a random variable $S_n = \sum_{i=1}^{n} X_i$. It is not hard to determine the distribution of S_n:

$$\mathbb{P}(S_n = \lfloor nx \rfloor) = \binom{n}{\lfloor nx \rfloor} p^{\lfloor nx \rfloor} (1-p)^{n - \lfloor nx \rfloor} \quad \text{if } x \in [0,1],$$

for which, here and throughout our study, we intend to examine the logarithm of the probability of the associated events. Specifically, by Stirling's approximation $(\log n! = n \log n - n + O(\log n))$,

$$\begin{aligned} & n^{-1} \log \mathbb{P}(S_n = \lfloor nx \rfloor) \\ = \; & \frac{\lfloor nx \rfloor}{n} \log \left(\frac{p}{\frac{\lfloor nx \rfloor}{n}} \right) + \left(1 - \frac{\lfloor nx \rfloor}{n} \right) \log \left(\frac{1-p}{1 - \frac{\lfloor nx \rfloor}{n}} \right) + O\left(\frac{\log n}{n} \right). \end{aligned}$$

Asymptotically, given that $x \mapsto -x \log x$ is uniformly continuous on $[0,1]$,

$$\lim_{n \to \infty} n^{-1} \log \mathbb{P}(S_n = \lfloor nx \rfloor) = -I(x),$$

where

$$I(x) = x \log \left(\frac{x}{p} \right) + (1-x) \log \left(\frac{1-x}{1-p} \right)$$

is plotted in Figure 1. In the language of large deviation theory, the sequence $n^{-1} S_n$ is said to satisfy the large deviation principle (which will be defined rigorously later) with rate function I.

Some comments are in order.

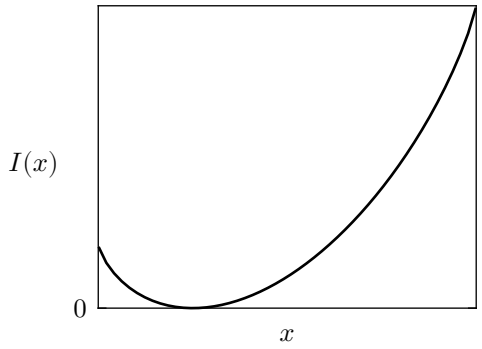

FIGURE 1. rate function $I(x)$

- $I(x)$ is a convex function with

$$I'(x) = \log\left(\frac{x}{1-x}\right) - \log\left(\frac{p}{1-p}\right) \quad \text{and} \quad I''(x) = \frac{1}{x} + \frac{1}{1-x}.$$

  In particular, its minimum is attained by a unique point $x = p = \mathbb{E}(X_1)$. This implies the law of large numbers by the Borel-Cantelli lemma.
- Observe that the second-order Taylor expansion

$$I(x) = \frac{(x-p)^2}{2p(1-p)} + O(|x-p|^2).$$

  When considering a small deviation $x = p + y/\sqrt{n}$, this is "roughly" consistent with the central limit theorem:

$$\mathbb{P}\left(\frac{\sqrt{n}(n^{-1}S_n - p)}{\sqrt{p(1-p)}} = \frac{y}{\sqrt{p(1-p)}}\right) \approx \frac{1}{\sqrt{2\pi p(1-p)}} e^{\frac{y^2}{2p(1-p)}}.$$

1.2. **General settings.** Throughout the note, we let $\{\mu_\varepsilon\}_\varepsilon$ be a collection of probability measures on a common measurable space $(\mathcal{X}, \mathcal{B})$. The family is indexed by a set of non-negative real numbers $\varepsilon > 0$ with an accumulation point at 0, as we are primarily interested in the limiting behavior of these measures as $\varepsilon \to 0$. Within this framework, we assume that $\mathcal{X}$ is a Hausdorff topological space and that $\mathcal{B}$ contains the Borel σ-algebra $\mathcal{B}_\mathcal{X}$.

Definition 1.2 (rate function)**.** Let $I : \mathcal{X} \to [0, \infty]$ be a function with *sublevel sets* $\Psi_I(\alpha) := \{x : I(x) \le \alpha\}$ and *effective domain* $\mathcal{D}_I := \{x : I(x) < \infty\}$.

- I is called a *rate function* if it is lower semicontinuous.
- A rate function I is said to be *good* if $\Psi_I(\alpha)$ is compact for all $\alpha < \infty$.

Definition 1.3 (large deviation principle)**.** The family $\{\mu_\varepsilon\}_\varepsilon$ is said to satisfy the *large deviation principle with rate* $I : \mathcal{X} \to [0, \infty]$ if for every $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \mathring{\Gamma}} I(x) \le \liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(\Gamma) \le \limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(\Gamma) \le -\inf_{x \in \overline{\Gamma}} I(x). \tag{1.1}$$

This definition of large deviation principle provides the most general framework for our discussion. In particular, one can further adapt the definition for any sequence of random variables as follows: The sequence $(X_n)_{n \in \mathbb{N}}$ is said to satisfy the large

deviation principle if the associated family of probability laws $\{\mu_{n^{-1}}\}_{n\in\mathbb{N}}$ satisfies the large deviation principle. This convention is applied to all statements concerning random variables.

The rationale behind our specific settings is tied directly to the following characterizations of the large deviation principle. By assuming $\mathcal{B}_{\mathcal{X}} \subseteq \mathcal{B}$, we immediately have the following equivalent statement.

Remark 1.4. *Suppose $\mathcal{B}_{\mathcal{X}} \subseteq \mathcal{B}$. Then, $\{\mu_\varepsilon\}_\varepsilon$ satisfies the large deviation principle with rate I if and only if the following hold.*

- *(upper bound) For any closed set $F \subset \mathcal{X}$,*
$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) \leq -\inf_{x\in F} I(x). \tag{1.2}$$
- *(lower bound) For any open set $G \subset \mathcal{X}$,*
$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(G) \geq -\inf_{x\in G} I(x). \tag{1.3}$$

As for the Hausdorff assumption, it ensures that compact sets are well-behaved, particularly concerning the notion of exponential tightness.

Definition 1.5. Suppose $\mathcal{B}$ contains all compact subsets of $\mathcal{X}$. The family $\{\mu_\varepsilon\}$ is *exponentially tight* if for every $\alpha < \infty$, there exists a compact set K such that
$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(K^c) < -\alpha.$$

As every compact set is closed in a Hausdorff space, we can relax the LDP conditions under the assumption of exponential tightness as follows:

Proposition 1.6. *Suppose $\mathcal{B}_{\mathcal{X}} \subset \mathcal{B}$. If $\{\mu_\varepsilon\}$ is exponentially tight, then the following hold.*

- *(upper bound) If (1.2) holds for every compact set, so does it for every closed set.*
- *(lower bound) If (1.3) holds for every open set, then I is good.*

Notably, any $\{\mu_\varepsilon\}$ admitting a rate function that satisfies the relaxed conditions is said to satisfy the weak LDP.

Proof. Suppose $\{\mu_\varepsilon\}$ is exponentially tight. If F is a closed set, then for any closed set F,
$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) \leq -\inf_{x\in F} I(x)$$
. For all $\infty > \alpha > 0$, there exists a compact set K_α such that
$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(K_\alpha^c) < -\alpha.$$
Hence,
$$\begin{aligned}\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) &\leq \max\left\{\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F\cap K_\alpha), \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(K_\alpha^c)\right\} \\ &= \lim_{a\to\infty} \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F\cap K_\alpha) \leq -\inf_{x\in F\cap K_\alpha} I(x),\end{aligned}$$

where the last inequality holds as $F \cap K_\alpha$ is compact due to the Hausdorff assumption.

With exponential tightness and the lower bound for open sets, we can find for every $\infty > \alpha > 0$ a compact set K_α such that

$$-\alpha > \liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(K_\alpha^c) \geq - \inf_{x \in K_\alpha^c} I(x).$$

This naturally implies

$$\Psi_I(\alpha) \subset K_\alpha,$$

which is a closed subset due to lower semicontinuity and therefore compact due to the Hausdorff assumption. □

2. Cramér's theorem in $\mathbb{R}$

In this section, we establish the Large Deviation Principle for the empirical mean of i.i.d. random variables. We begin by introducing the primary tools of our analysis: the logarithmic moment generating function and its Fenchel-Legendre transform. Let $(X_i)_{i=1}^{\infty}$ be i.i.d. random variables with law $\mu \in M_1(\mathbb{R})$, and let μ_n be the law of the empirical mean $\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We denote the expectation of X_1 by $\overline{x} := \mathbb{E}X_1$, whenever it is well-defined.

Definition 2.1. The *logarithmic moment generating function* (log MGF) associated with the law μ is defined as

$$\Lambda(\lambda) := \log \mathbb{E}\big(e^{\langle \lambda, X_1\rangle}\big). \tag{2.1}$$

Definition 2.2. The *Fenchel-Legendre transform* of $\Lambda(\lambda)$ is

$$\Lambda^*(x) := \sup_{\lambda\in\mathbb{R}}[\langle \lambda, x\rangle - \Lambda(\lambda)]. \tag{2.2}$$

Example 2.3. There is a geometric interpretation of the Fenchel-Legendre transform of $\Lambda(\lambda)$; that is, $\Lambda^*(x)$ is the supremum of the y-intercepts of all lines with slope x that lie below Λ.

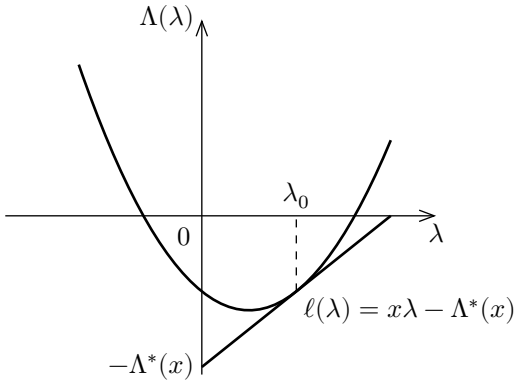

Figure 2. rate function $I(s)$

For the empirical means, the large deviation principle takes the following form:

Theorem 2.4 (Cramér)**.** *Let X_i, μ_n, Λ, and Λ^* be as defined on $\mathbb{R}$. Then, $\{\mu_n\}$ satisfies the LDP with the convex rate function Λ^*, namely,*

(1) For any closed set F, $\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(F) \le -\inf_{x\in F} \Lambda^*(x)$.

(2) For any open set G, $\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(G) \ge -\inf_{x\in G} \Lambda^*(x)$.

Furthermore,

(3) If $0 \in \mathcal{D}_\Lambda^\circ$, then Λ^* is good.

To motivate the proof, we first examine the role of the log MGF. For real-valued random variables, Markov's inequality provides the key estimate for the LDP upper bound ((1.2)). Specifically, for any $\lambda \ge 0$:

$$\mu_n[x,\infty) \le e^{-\lambda x}\mathbb{E}\Big(e^{\lambda \hat{S}_n}\Big) = e^{-\lambda x}\big(\mathbb{E}\big(e^{\lambda X_1}\big)\big)^n$$

$$\Rightarrow \limsup_{n\to\infty} \frac{1}{n}\log \mu_n[x,\infty) \le \inf_{\lambda\ge 0} -\lambda x + \Lambda(\lambda).$$

Similarly,

$$\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(-\infty,x] \le \inf_{\lambda\le 0} -\lambda x + \Lambda(\lambda).$$

These observations suggest that Λ^* is the natural candidate for the rate function. The following lemma summarizes the key properties of Λ and Λ^*.

Lemma 2.5. *Let Λ and Λ^* be as defined. Then,*

(1) Λ is a convex function and Λ^* is a convex rate function.

(2) Either of following holds.

- *If $\Lambda(\lambda) < \infty$ only when $\lambda = 0$, then Λ^* is identically zero.*
- *If $\Lambda(\lambda) < \infty$ for some $\lambda > 0$ (respectively, $\lambda < 0$), then $\overline{x} < \infty$ (respectively, $\overline{x} > -\infty$) is well-defined. Under the circumstances,*

$$\Lambda^*(x) = \begin{cases} \sup_{\lambda\ge 0}[\lambda x - \Lambda(\lambda)] \text{ if } x \ge \overline{x}, \\ \sup_{\lambda\le 0}[\lambda x - \Lambda(\lambda)] \text{ if } x \le \overline{x}, \end{cases} \tag{2.3}$$

  which satisfies

  - *Λ^* is decreasing on $(-\infty,\overline{x}]$ and increasing on $[\overline{x},\infty)$, and*
  - *$\inf_{x\in\mathbb{R}} \Lambda^*(x) = 0$ and if $\overline{x} \in \mathbb{R}$, then $\Lambda^*(\overline{x}) = 0$.*

(3) Λ is differentiable in $\mathring{\mathcal{D}}_\Lambda$ with $\Lambda'(\lambda) = \big(\mathbb{E}\big(e^{\lambda X_1}\big)\big)^{-1}\mathbb{E}\big(X_1 e^{\lambda X_1}\big)$ and

$$\Lambda'(\lambda) = x \Rightarrow \Lambda^*(x) = \lambda x - \Lambda(\lambda).$$

(4) If $0 \in \mathring{\mathcal{D}}_\Lambda$, then Λ^* is a good rate function. Moreover, if $\mathcal{D}_\Lambda = \mathbb{R}$, then $\lim_{|x\mapsto\infty} \frac{\Lambda^*(x)}{|x|} = \infty$.

Proof. (1) By Hölder's inequality, given any $\lambda, \lambda' \in \mathbb{R}$ and any $t, t' \in [0,1]$ satisfying $t + t' = 1$,

$$\log \mathbb{E}\Big(e^{\langle t\lambda + t'\lambda', X_1\rangle}\Big) \le t \log \mathbb{E}\big(e^{\langle \lambda, X_1\rangle}\big) + t' \log \mathbb{E}\Big(e^{\langle \lambda', X_1\rangle}\Big),$$

proving the convexity of Λ. The convexity of Λ^* follows from definition:

$$\sup_{\lambda\in\mathbb{R}}[\langle \lambda, tx + t'x\rangle - \Lambda(\lambda)] \le t \sup_{\lambda\in\mathbb{R}}[\langle \lambda, x\rangle - \Lambda(\lambda)] + t' \sup_{\lambda\in\mathbb{R}}[\langle \lambda, x'\rangle - \Lambda(\lambda)].$$

To prove the lower semicontinuity, observe that for any $\lambda \in \mathbb{R}$ and any $x \in \mathbb{R}$,

$$\liminf_{y\to x} \langle \lambda, y\rangle - \Lambda(\lambda) = \langle \lambda, x\rangle - \Lambda(\lambda),$$

implying $\liminf_{y\to x} \Lambda^*(y) \ge \Lambda^*(x)$. The non-negativity of Λ^* follows from the fact $\Lambda^*(x) \ge \langle 0, x\rangle - \Lambda(0) = 0$.

(2) The case $\mathcal{D}_\Lambda = \{0\}$ is automatic. If $\Lambda(\lambda) < \infty$ for some $\lambda > 0$, then, due to the fact that $e^x \ge x$ for all $x \ge 0$,

$$\mathbb{E}\Big[X_1 \mathbb{1}_{\{X_1\ge 0\}}\Big] \le \lambda^{-1}\mathbb{E}\Big[e^{\langle \lambda, X_1\rangle}\mathbb{1}_{\{X_1\ge 0\}}\Big] \le \lambda^{-1}\exp(\Lambda(\lambda)) < \infty,$$

meaning both $\mathbb{E}(\langle \lambda, X_1\rangle)$ and $\Lambda(\lambda)$ are well-defined. Hence, by Jensen's inequality,

$$\overline{x} = \lambda^{-1}\log \circ \exp(\mathbb{E}(\langle \lambda, X_1\rangle)) \le \lambda^{-1}\Lambda(\lambda) < \infty.$$

The argument proceeds similarly for $\Lambda(\lambda) > -\infty$ whenever $\lambda < 0$.

We then proceed to prove (2.3) and its related properties. Since now $\overline{x}$ is well-defined, by Jensen's inequality, for all $\eta < 0$ and $x \geq \overline{x}$ (similar for $\eta > 0$ and $x \leq \overline{x}$),

$$\langle \eta, x\rangle - \Lambda(\eta) = \log \mathbb{E}\big(e^{\langle \eta, x - X_1\rangle}\big) \leq \langle \eta, x - \overline{x}\rangle \leq 0 = \langle 0, x\rangle - \Lambda(0), \tag{2.4}$$

from which (2.3) follows. Moreover, (2.3) implies the monotonicity on $(\overline{x}, \infty)$ and $(\infty, \overline{x})$. Finally, if $\overline{x} \in \mathbb{R}$, then by (2.4), $\inf_{x\in\mathbb{R}} \Lambda^*(x) = \Lambda^*(\overline{x}) = 0$. If $\overline{x} = -\infty$ (similar for $\overline{x} = -\infty$), we deduce from Markov's inequality that for all $\lambda \geq 0$

$$\log \mu[x, \infty) \leq \log e^{-\langle \lambda, x\rangle} \mathbb{E}\big(e^{\langle \lambda, X_1\rangle}\big),$$

from which it follows that $0 \leq \inf_{x\in\mathbb{R}} \Lambda^*(x) \leq \lim_{x\to\overline{x}} \Lambda^*(x) \leq 0$, as desired.

(3) The differentiability follows from the dominated convergence theorem. Let $f_\varepsilon(x) = \frac{e^{(\lambda+\varepsilon)x} - e^{\lambda x}}{\varepsilon}$ so that $f_\varepsilon(x) \to xe^{\lambda x}$ as $\varepsilon \to 0$ and $|f_\varepsilon(x)| \leq \frac{e^{\lambda x}\left(e^{\delta|x|} - 1\right)}{\delta} =: h_\delta(x)$ whenever $|\varepsilon| < \delta$. Since $\mathbb{E}(h_\delta(X_1)) < \infty$ for all sufficiently small δ, we may apply the dominated convergence theorem to derive the derivative of Λ. Finally, the function $g(\eta) := \langle \eta, x\rangle - \Lambda(\eta)$ is concave and thus $g'(\lambda) = 0$ implies $g(\lambda) = \sup_{\eta\in\mathbb{R}} g(\eta)$, proving the proposed property.

(4) Suppose $[\lambda_-, \lambda_+] \subset \mathcal{D}_\Lambda$ is a non-degenerate interval containing 0. Then, for $\lambda \in [\lambda_-, \lambda_+]$

$$\liminf_{|x|\mapsto\infty} \frac{\Lambda^*(x)}{|x|} \geq \liminf_{|x|\mapsto\infty} \left[\lambda \operatorname{sign}(x) - \frac{\Lambda(\lambda)}{|x|}\right] \geq \min\{-\lambda_-, \lambda_+\} > 0,$$

implying $\Lambda^*(x) \to \infty$ as $|x| \to \infty$. Hence, the sublevel set $\Psi_{\Lambda^*}(\alpha)$ is closed and bounded for all $\alpha < \infty$, and Λ^* is good. When $\mathcal{D}_\Lambda = \mathbb{R}$, then the result follows by letting $\lambda_+ = -\lambda_- \to \infty$. □

Proof of Theorem 2.4. (1) For brevity, let $I_F = \inf_{x\in F} \langle \lambda, x\rangle - \Lambda(\lambda)$. If $I_F = 0$, then the inequality is trivial. Hence, we assume $I_F > 0$. Under the circumstances, the numbers $x_+ = \inf[\overline{x}, \infty) \cap F$ and $x_- = \sup(-\infty, \overline{x}] \cap F$ are different from $\overline{x}$. By Markov's inequality, for all $\lambda > 0$ and $\lambda' < 0$,

$$\mu_n(F) \leq \mu_n[x_+, \infty) + \mu_n(-\infty, x_-] \leq e^{-n(\langle \lambda, x_+\rangle - \Lambda(\lambda))} + e^{-n(\langle \lambda', x_-\rangle - \Lambda(\lambda'))}.$$

By Lemma 2.5,

$$\mu_n(F) \leq e^{-n\Lambda^*(x_+)} + e^{-n\Lambda^*(x_-)} \leq 2e^{I_F}.$$

Taking the normalized logarithm and letting $n \to \infty$ yields the desired inequality.

(2) It suffices to show that for all measures μ and all $\delta > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(-\delta, \delta) \geq \inf_{\lambda\in\mathbb{R}} \Lambda(\lambda) = -\Lambda^*(0), \tag{2.5}$$

for once this is proved, one may simply consider the translation $Y = X - x$ to deduce that the log MGF $\Lambda_Y(\lambda) = \Lambda(\lambda) - \langle \lambda, x\rangle$ and that $\Lambda_Y^*(z) = \Lambda^*(z + x)$, which in turn imply

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(x - \delta, x + \delta) \geq -\Lambda^*(x).$$

This proves the desired lower bound.

To prove (2.5), assume first that (a) $\mu(-\infty,0)>0$, (b) $\mu(0,\infty)>0$, and (c) μ is boundedly supported. Under the circumstances, (a) and (b) imply $\Lambda(\lambda)\to\infty$ as $|\lambda|\to\infty$, and (c) implies Λ is finite on $\mathbb{R}$. Consequently, by Lemma 2.5, there exists $\eta\in\mathbb{R}$ such that

$$\Lambda'(\lambda)=0 \quad\text{and}\quad \Lambda(\lambda)=\inf_{\eta\in\mathbb{R}}\Lambda(\eta).$$

Define a probability measure $\tilde{\mu}$ by

$$\frac{d\tilde{\mu}}{d\mu}(x)=e^{\langle\lambda,x\rangle-\Lambda(\lambda)}.$$

Observe that

$$\begin{aligned}\mu_n(-\varepsilon,\varepsilon)&=\int_{|\sum_{i=1}^n x_i|<n\varepsilon} d\mu(x_1)\cdots d\mu(x_n)\\&=\int_{|\sum_{i=1}^n x_i|<n\varepsilon} e^{\sum_{i=1}^n[-\langle\lambda,x_i\rangle+\Lambda(\lambda)]}d\tilde{\mu}(x_1)\cdots d\tilde{\mu}(x_n).\\&\geq e^{n\Lambda(\lambda)-n\varepsilon|\lambda|}\tilde{\mu}_n(-\varepsilon,\varepsilon).\end{aligned}$$

By Lemma 2.5,

$$\mathbb{E}_{\tilde{X}\sim\tilde{\mu}}(\tilde{X})=\int xe^{\langle\lambda,x\rangle-\Lambda(\lambda)}d\mu(x)=\big(\mathbb{E}(e^{\langle\lambda,x\rangle})\big)^{-1}\mathbb{E}\big(X_1e^{\langle\lambda,X_1\rangle}\big)=\Lambda'(\lambda)=0.$$

Hence, by the law of large numbers,

$$\lim_{n\to\infty}\tilde{\mu}_n(-\varepsilon,\varepsilon)=1.$$

Hence, for all $0<\varepsilon<\delta$,

$$\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(-\delta,\delta)\geq\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(-\varepsilon,\varepsilon)\geq\Lambda(\lambda)-\varepsilon|\lambda|,$$

proving (2.5) by letting $\varepsilon\to0$.

Now, if the support of μ is not bounded, fix $M>0$ such that (a) $\mu[-M,0)>0$, (b) $\mu(0,M]>0$. Hence, by letting ν be the normalized law of X_1 on $\{|X_1|\leq M\}$, we have that

$$\begin{aligned}\mu_n(-\delta,\delta)&\geq\int_{\frac{1}{n}\sum_{i=1}^n x_i\in(-\delta,\delta)}\prod_{i=1}^n\mathbb{1}_{[-M,M]}(x_i)d\mu(x_1)\cdots d\mu(x_n)\\&=\nu_n(-\delta,\delta)\cdot\mu[-M,M]^n\end{aligned}$$

and that the log MGF associated with ν is

$$\log\int_{-M}^{M}e^{\langle\lambda,x\rangle}d\mu(x)-\log\mu[-M,M]=:\Lambda_M(\lambda)-\log\mu[-M,M].$$

Hence, by the case of bounded support,

$$\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(-\varepsilon,\varepsilon)\geq\log\mu[-M,M]+\liminf_{n\to\infty}\frac{1}{n}\log\nu_n(-\varepsilon,\varepsilon)\geq\inf_{\lambda\in\mathbb{R}}\Lambda_M(\lambda),$$

and therefore, by writing $I_M=-\inf_{\lambda\in\mathbb{R}}\Lambda_M(\lambda)$ and $I^*=\limsup_{M\to\infty}I_M$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(-\varepsilon, \varepsilon) \geq -I^*. \tag{2.6}$$

Since Λ_M is increasing in M, so is $-I_M$. Therefore, $I^* >= -\infty$ and the level sets $\{\lambda : \Lambda_M(\lambda) \leq -I^*\}$ are decreasing compacted sets, admitting some λ_0 in their intersection. By monotone convergence theorem,

$$\Lambda(\lambda_0) = \lim_{M\to\infty} \Lambda_M(\lambda_0) \leq -I^*. \tag{2.7}$$

Combining (2.6) and (2.7) proves (2.5).

Finally, if $\mu(-\infty, 0) = 0$ or $\mu(0, \infty) = 0$, then Λ is monotone and $\inf_{\lambda\in\mathbb{R}} \Lambda(\lambda) = \log \mu(\{0\})$. The bound $\mu_n(-\delta, \delta) \geq \mu_{n(\{0\})} = \mu(\{0\})^n$ then yields (2.5).

(3) It follows from Lemma 2.5(4). □

3. Gärtner-Ellis theorem.

Let Λ_ε be the log MGF associated with d-dimensional real random variables Z_ε, which can be defined as

$$\Lambda_{\mu_\varepsilon}(\lambda) := \log \mathbb{E}\big[e^{\langle \lambda, Z_\varepsilon\rangle}\big].$$

The Gärtner-Ellis theorem states the following.

Definition 3.1. A convex function $\Lambda : \mathbb{R}^d \to (-\infty, \infty]$ is *essentially smooth* if:

(1) $\mathring{\mathcal{D}}_\Lambda$ is nonempty.
(2) Λ is differentiable throughout $\mathring{\mathcal{D}}_\Lambda$.
(3) Λ is steep, namely, $\lim_{n\to\infty} |\nabla \Lambda(\lambda_n)| = \infty$ whenever $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $\mathring{\mathcal{D}}_\Lambda$ converging to a boundary point of $\mathring{\mathcal{D}}_\Lambda$.

Theorem 3.2 (Gärtner-Ellis)**.** *Suppose that $\Lambda(\lambda) := \lim_{\varepsilon\to 0} \varepsilon \Lambda_\varepsilon(\varepsilon\lambda)$ exists for all $\lambda \in \mathbb{R}^d$ as extended real numbers and $0 \in \mathring{\mathcal{D}}_\Lambda$.*

(1) For any closed set F,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) \le -\inf_{x\in F} \Lambda^*(x).$$

(2) For any open set G,

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(G) \ge -\inf_{x\in G\cap \mathcal{F}} \Lambda^*(x),$$

where $\mathcal{F}$ is the set of exposed points of Λ^* admitting an exposing hyperplane belonging to $\mathring{\mathcal{D}}_\Lambda$.

(3) If Λ is an essentially smooth, lower semicontinuous function, then the LDP holds with the good rate function Λ^*.

The proof of the third statement relies on several results from convex analysis; for clarity, we state these results here and defer their proofs.

Lemma 3.3. *Under the same assumption as in Theorem 3.2, the following hold.*

(1) $\Lambda(\lambda)$ is a convex function, $\Lambda(\lambda) > -\infty$ everywhere.
(2) $\Lambda^*(x)$ is a good convex rate function.
(3) Suppose that $y = \nabla\Lambda(\eta)$ for some $\eta \in \mathring{\mathcal{D}}_\Lambda$. Then $\Lambda^*(y) = \langle \eta, y\rangle - \Lambda(\eta)$. Moreover $y \in \mathcal{F}$, with η being the exposing hyperplane for y.

Lemma 3.4 (Rockafellar)**.** *If $\Lambda : \mathbb{R}^d \to (-\infty, \infty]$ is an essentially smooth, lower semicontinuous, convex function, then* ri $\mathcal{D}_{\Lambda^*} \subseteq \mathcal{F}$.

With these tools, we can now proceed with the proof of the Gärtner-Ellis theorem. For convenience, we first introduce the following auxiliary function:

Definition 3.5. The δ-rate function associated with a rate function I is a function defined as

$$I^\delta(x) = \min\Big\{I(x) - \delta, \frac{1}{\delta}\Big\}. \tag{3.1}$$

Proof of Theorem 3.2. (1) It suffices to prove the inequality for all compact sets and that $\{\mu_\varepsilon\}$ is exponentially tight.

The upper bound for compact sets follows essentially from Markov's inequality. Choose for each $x \in \Gamma$ a $\lambda_x \in \mathbb{R}^d$ and an open neighborhood A_x of x such that

$$\langle \lambda_x, x\rangle - \Lambda(x) \geq I^\delta(x),$$
$$\inf_{y\in A_x}[\langle \lambda_x, y\rangle - \langle \lambda_x, x\rangle] \geq -\delta,$$

where I^δ is the δ-rate function associated with Λ^* as defined in (3.1). Applying Markov's inequality yields

$$\mu_\varepsilon(A_x) \leq \mathbb{E}\Big(e^{\langle \frac{\lambda_x}{\varepsilon}, Z_\varepsilon\rangle - \langle \frac{\lambda_x}{\varepsilon}, x\rangle}\Big) \exp\bigg(-\inf_{y\in A_x}\Big\langle \frac{\lambda_x}{\varepsilon}, y\Big\rangle - \Big\langle \frac{\lambda_x}{\varepsilon}, x\Big\rangle\bigg),$$

which in turn implies

$$\varepsilon \log \mu_\varepsilon(A_x) \leq \delta - \bigg[\langle \lambda_x, x\rangle - \varepsilon \Lambda_{\mu_\varepsilon}\Big(\frac{\lambda_x}{\varepsilon}\Big)\bigg].$$

Now that Γ is a compact set, one can find a finite cover $\{A_{x_i}\}_{i=1}^N$ with $x_i \in \Gamma$ such that

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(\Gamma) \leq \delta - \min_{1\leq i\leq N} I^\delta(x_i) \leq \delta - \inf_{x\in\Gamma} I^\delta(x),$$

proving the theorem by letting $\delta \to 0$.

It is left to demonstrate the exponential tightness of μ_ε, which is equivalent to that for all marginals μ^j of μ on j-th coordinate and follows essentially from Lemma 2.5 (4). Let $\beta > 0$ be sufficiently small so that $\overline{B}_\rho(0) \subset \mathring{\mathcal{D}}_{\Lambda^j}$. By Markov's inequality, we have that for all $\rho \in \mathbb{R}$,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu^j_\varepsilon[\rho, \infty) \leq \limsup_{\varepsilon\to 0} -\beta\rho + \varepsilon \Lambda_{\mu_\varepsilon}\big(\varepsilon^{-1}\beta \mathbf{e}_j\big) \leq -\beta\rho + \Lambda^*\big(\beta \mathbf{e}_j\big), \quad (3.2)$$

where $\mathbf{e}_j$ is the j-th vector in the standard basis and the right-hand side converges to $-\infty$ as $\rho \to \infty$. The same arguments applies to $\limsup_\infty \varepsilon \log \mu^j_\varepsilon(-\infty, \rho]$. The two estimates combined prove the exponential tightness.

(2) The case $\Lambda(\lambda) = -\infty$ for some $\lambda \in \mathcal{X}^*$ is trivial, for $\Lambda^*(\cdot) = \infty$ everywhere. Assuming $\Lambda(\lambda) < \infty$ for all $\lambda \in \mathcal{X}^*$, one may adopt the change of measure argument as before in the following manner. Let G be any fixed open set, $y \in G \cap \mathcal{F}$, $\delta > 0$, and $\eta \in \mathring{\mathcal{D}}_\Lambda$ be an exposing hyperplane for Λ^*. By continuity of η, we choose an open neighborhood $B_\delta \subset G$ of y such that

$$\sup_{z\in B_\delta} \langle \eta, z - y\rangle < \delta.$$

Now that $\Lambda_{\mu_\varepsilon}\big(\frac{\eta}{\varepsilon}\big)$ is well-defined for every sufficiently small $\varepsilon > 0$, we define a probability measure $\tilde{\mu}_\varepsilon$ equivalent to μ_ε:

$$\frac{d\tilde{\mu}_\varepsilon}{d\mu_\varepsilon}(z) = \exp\bigg[\Big\langle \frac{\eta}{\varepsilon}, z\Big\rangle - \Lambda_{\mu_\varepsilon}\Big(\frac{\eta}{\varepsilon}\Big)\bigg]$$

so that

$$\begin{aligned}\varepsilon \log \mu_\varepsilon(B_\delta) &= \varepsilon \log \int_{B_\delta} \exp\bigg[-\Big\langle \frac{\eta}{\varepsilon}, z\Big\rangle + \Lambda_{\mu_\varepsilon}\Big(\frac{\eta}{\varepsilon}\Big)\bigg] d\tilde{\mu}_\varepsilon(z)\\ &= \varepsilon \Lambda_{\mu_\varepsilon}\Big(\frac{\eta}{\varepsilon}\Big) - \langle \eta, y\rangle + \varepsilon \log \int_{B_\delta} \exp\bigg[\Big\langle \frac{\eta}{\varepsilon}, y - z\Big\rangle\bigg] d\tilde{\mu}_\varepsilon(z),\end{aligned}$$

yielding

$$\begin{aligned}\liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(B_\delta) &\geq \Lambda(\eta)-\langle\eta,y\rangle+\lim_{\delta\to 0}\liminf_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon(B_\delta),\\ &\geq -\Lambda^*(y)+\lim_{\delta\to 0}\liminf_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon(B_\delta).\end{aligned}$$

It remains to show that $\lim_{\delta\to 0}\liminf_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon(B_\delta)=0$. To this end, we claim that

$$\limsup_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon\big(B_\delta^c\cap[-\rho,\rho]^d\big)<0 \text{ for all } \delta>0,\rho>0, \tag{3.3}$$

$$\limsup_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon\Big(\big([-\rho,\rho]^d\big)^c\Big)<0 \text{ for all large } \rho>0, \tag{3.4}$$

which together implies the desired estimate. To prove (3.3), define the function

$$\tilde{\Lambda}(\cdot):=\limsup_{\varepsilon\to 0}\varepsilon\Lambda_{\tilde{\mu}_\varepsilon}\left(\frac{\cdot}{\varepsilon}\right)$$

and observe that

$$\begin{aligned}\tilde{\Lambda}(\cdot)&=\Lambda(\cdot+\eta)-\Lambda(\eta),\\ \tilde{\Lambda}^*(\cdot)&=\Lambda^*(\cdot)+\Lambda(\eta)-\langle\eta,\cdot\rangle.\end{aligned}$$

One deduces from the Markov's inequality that

$$\limsup_{\varepsilon\to 0}\varepsilon\log\tilde{\mu}_\varepsilon(B_\delta^c\cap K_\alpha)\leq-\inf_{z\in B_\delta^c\cap K_\alpha}\tilde{\Lambda}^*(z)<0,$$

since $\tilde{\Lambda}^*$ is lower semicontinuous and η is an exposing hyperplane. For (3.4), observe that $0\in\mathring{\mathcal{D}}_{\tilde{\Lambda}}$, so the inequality follows from (3.2).

(3) With Lemma 3.3 and Lemma 3.4, we have that for ever open set G,

$$\liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(G)\geq-\inf_{x\in G\cap\mathcal{F}}\Lambda^*(x)=-\inf_{x\in G}\Lambda^*(x)$$

□

We now prove the lemmas.

Proof of Lemma 3.3. (1) The convexity follows from Hölder's inequality as in Lemma 2.5. Precisely, given any $t,t'\in[0,1]$ satisfying $t+t'=1$,

$$\log\mathbb{E}\Big(e^{\langle t\lambda+t'\lambda',Z_\varepsilon\rangle}\Big)\leq t\log\mathbb{E}\big(e^{\langle\lambda,Z_\varepsilon\rangle}\big)+t'\log\mathbb{E}\Big(e^{\langle\lambda',Z_\varepsilon\rangle}\Big),$$

proving the convexity of $\Lambda_{\mu_\varepsilon}$ and hence $\overline{\Lambda}$.

If $\Lambda(\lambda)=-\infty$ for some $\lambda\in\mathbb{R}^d$, then by convexity $\Lambda(\alpha\lambda)=-\infty$ for all $\alpha\in(0,1]$. Since $\Lambda(0)=0$, it follows by convexity that $\Lambda(-\alpha\lambda)=\infty$ for all $\alpha\in(0,1]$, contradicting the assumption that $0\in\mathring{\mathcal{D}}$. Thus, $\Lambda>-\infty$ everywhere.

(2) Since $0\in\mathring{\mathcal{D}}_\Lambda$, it follows that $B_\delta(0)\subset\mathring{\mathcal{D}}_\Lambda$ for some $\delta>0$, and $c=\sup_{\lambda\in B_\delta(0)}\Lambda(\lambda)<\infty$ since the convex function Λ is continuous in $\mathring{\mathcal{D}}_\Lambda$. Therefore,

$$\Lambda^*(x)\geq\sup_{\lambda\in\overline{B}_\delta(0)}[\langle\lambda,x\rangle-\Lambda(\lambda)]\geq|\delta||x|-\sup_{\lambda\in\overline{B}_\delta(0)}\Lambda(x),$$

implying $\Psi_{\Lambda^*}(\alpha)$ is bounded for every $\alpha<\infty$. The function Λ^* is convex and lower semicontinuous by a routine check as conducted in Lemma 2.5. Combining these implies that Λ^* is a good convex rate function.

(3) Suppose now that for some $x\in\mathbb{R}^d$,

$$\Lambda(\eta) = \langle \eta, y\rangle - \Lambda^*(y) \le \langle \eta, x\rangle - \Lambda^*(x).$$

Then, for every $\theta \in \mathbb{R}^d$,

$$\langle \theta, x\rangle \le \Lambda(\eta+\theta) - \Lambda(\eta).$$

In particular,

$$\langle \theta, x\rangle \le \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}[\Lambda(\eta+\varepsilon\theta) - \Lambda(\eta)] = \langle \theta, \nabla\Lambda(\eta)\rangle.$$

Since the inequality holds for all $\theta \in \mathbb{R}^d$, $x = \nabla\Lambda(\eta) = y$. Hence, y is an exposed point of Λ^* with exposing hyperplane $\eta \in \mathcal{D}_\Lambda^\circ$. □

Before proving Rockafellar's lemma, let us recall the definition of relative interior of a convex set.

Definition 3.6. The *relative interior* of a nonempty convex set C is defined as

$$\text{ri } C = \{y \in C : \text{for all } x \in C, y - \varepsilon(x-y) \in C \text{ for some } \varepsilon > 0\}.$$

Proof of Lemma 3.4. Assume without loss of generality that $\mathcal{D}_{\Lambda^*} \neq \varnothing$ for the lemma is vacuous otherwise. Under the circumstances, fix henceforth $x \in \text{ri } \mathcal{D}_{\Lambda^*}$ and define a function

$$f(\lambda) := \Lambda(\lambda) - \langle \lambda, x\rangle + \Lambda^*(x).$$

Observe that $f : \mathbb{R}^d \to [0,\infty]$ is convex, lower semincontinuous, and $\inf_{\lambda\in\mathbb{R}^d} f(\lambda) = 0$. Moreover, $f^*(\cdot) = \Lambda^*(\cdot + x) - \Lambda^*(x)$. Therefore, from $x \in \text{ri } \mathcal{D}_{\Lambda^*}$ it follows that $0 \in \text{ri } \mathcal{D}_{f^*}$. By Lemma C.2, there exists $\eta \in \mathcal{D}_\Lambda$ such that $f(\eta) = 0$. Let $\tilde\Lambda(\cdot) = \Lambda(\cdot + \eta) - \Lambda(\eta)$, so that $\tilde\Lambda$ is an essentially smooth, convex function and $\tilde\Lambda(0) = 0$. Consequently, by Lemma C.3, $\tilde\Lambda$ is finite in a neighborhood of the origin and thus $\eta \in \mathcal{D}_\Lambda^\circ$, at which f is differentiable by the assumption. Hence, $f(\eta) = \inf_{\lambda\in\mathbb{R}^d} f(\lambda)$, implying that $\nabla f(\eta) = 0$, i.e., $x = \nabla\Lambda(\eta)$. It now follows from Lemma 3.3(2) that $x \in \mathcal{F}$. Since $x \in \text{ri } \mathcal{D}_{\Lambda^*}$ is arbitrary, the proof is complete. □

The following theorem is a generalization of the Gärtner-Ellis theorem (Theorem 3.2), which essentially follows from the same proof with marginal distributions replaced by projection measures in every direction. Its proof can be found in [1].

Theorem 3.7 (Baldi)**.** *Suppose $\{\mu_\varepsilon\}$ is an exponentially tight family of probability measures on $\mathcal{X}$.*

(1) For every closed set $F \subset \mathcal{X}$, $\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) \le -\inf_{x\in F} \overline{\Lambda}^*(x)$.

(2) Let $\mathcal{F}$ be the set of exposed points of $\overline{\Lambda}^*$ with an exposing hyperplane λ for which

$$\Lambda(\lambda) = \lim_{\varepsilon\to 0} \varepsilon \Lambda_{\mu_\varepsilon}\left(\frac{\lambda}{\varepsilon}\right) \text{ exists and } \overline{\Lambda}(\gamma\lambda) < \infty \text{ for some } \gamma > 1. \quad (3.5)$$

Then, for every open set $G \subset \mathcal{X}$,

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(G) \ge -\inf_{x\in G\cap\mathcal{F}} \overline{\Lambda}^*(x).$$

(3) If for every open set $G \subset \mathcal{X}$, $\inf_{x\in G\cap\mathcal{F}} \overline{\Lambda}^*(x) = \inf_{x\in G} \overline{\Lambda}^*(x)$, then $\{\mu_\varepsilon\}$ satisfies the LDP with the good rate function $\overline{\Lambda}^*$.

Proof. See [1, Theorem 4.5.20]. □

The theorem is accompanied by a generalized criterion of differentiability of Λ in the following sense.

Definition 3.8 (Gateaux differentiable)**.** Let $\mathcal{X}$ be a topological vector space. A function $f : \mathcal{X}^* \to \mathbb{R}$ is said to be *Gateaux differentiable* if for all $\lambda, \theta \in \mathcal{X}^*$, the function $t \in \mathbb{R} \mapsto f(\lambda + t\theta)$ is differentiable at $t = 0$.

Corollary 3.9. *Let $\{\mu_\varepsilon\}$ be exponentially tight probability measures on the Banach space $\mathcal{X}$. Suppose that $\Lambda(\cdot) = \lim_{\varepsilon \to 0} \Lambda_{\mu_\varepsilon}\left(\frac{\cdot}{\varepsilon}\right)$ is finite-valued, Gateaux differentiable, and lower semicontinuous in $\mathcal{X}^*$ with respect to the weak* topology. Then $\{\mu_\varepsilon\}$ satisfies the LDP with the good rate function Λ^*.*

Proof. See [1, Corollary 4.5.27]. □

4. Applications

4.1. **Cramér's theorem in $\mathbb{R}^d$.** Combining the Gärtner-Ellis theorem with our previous results, we can establish the LDP for d-dimensional i.i.d. random variables $(X_i)_i$. As before, let μ_n be the law of the empirical mean, and let Λ and Λ^* be defined accordingly.

Theorem 4.1 (Cramér)**.** *Suppose $\mathcal{D}_\Lambda = \mathbb{R}^d$, then the family $\{\mu_n\}$ satisfies the LDP with the convex good rate function Λ^*.*

While powerful, this theorem does not encompass all scenarios where an LDP holds; a refined version requires only that $0 \in \mathring{\mathcal{D}}_\Lambda$ for the result to remain valid.

Example 4.2. Let μ be a Borel probability measure on $\mathbb{R}$ with a density $f(x) = C_d \cdot e^{-\|x\|}/\big(1+\|x\|^{d+2}\big)$ and Λ be its log MGF, where C_d is the normalizing constant. We will show in the following that $0 \in \mathring{\mathcal{D}}_\Lambda = \big\{\lambda \in \mathbb{R}^d : \|\lambda\| \le 1\big\}$, so the large deviation principle holds with a good rate function; however, Λ is not steep and hence the Gärtner-Ellis theorem does not apply.

On one hand, with the aid of Cauchy-Schwarz inequality $|\langle \lambda, x\rangle| \le \|\lambda\|\|x\|$, we realize that the estimate

$$e^{\langle \lambda, x\rangle} f(x) \le \frac{C_d\, e^{\|x\|(\|\lambda\|-1)}}{1+\|x\|^{d+2}} \quad \big(\forall \lambda \in \mathbb{R}^d\big)$$

holds and therefore $\mathcal{D}_\Lambda \supseteq \big\{\lambda \in \mathbb{R}^d : \|\lambda\| \le 1\big\}$. On the other hand, given any $\lambda \neq 0$, we have that along direction λ,

$$e^{\langle \lambda, x\rangle} f(x) = \frac{C_d\, e^{t\,\|\lambda\|\,(\|\lambda\|-1)}}{1+\|x\|^{d+2}} \quad (x = t\lambda, \forall t \in \mathbb{R}),$$

is unbounded as $t \to \infty$ if $\|\lambda\| > 1$, implying $\mathcal{D}_\Lambda = \big\{\lambda \in \mathbb{R}^d : \|\lambda\| \le 1\big\}$. Finally, we have uniform bound

$$\frac{C_d e^{-2\|x\|}}{1+\|x\|^{d+2}} \le e^{\langle \lambda, x\rangle} f(x) \le \frac{C_d}{1+\|x\|^{d+2}} \quad (\lambda \in \mathcal{D}_\Lambda),$$

Therefore, within the interior of $\mathcal{D}_\Lambda$,

$$\sup_{\lambda \in \mathring{\mathcal{D}}_\Lambda} |\nabla \Lambda(\lambda)| \le \left(\int \frac{e^{-2\|x\|}}{1+\|x\|^{d+2}}\, dx\right)^{-1} \int \frac{\|x\|\, e^{-2\|x\|}}{1+\|x\|^{d+2}}\, dx < \infty.$$

This proves that Λ is not steep.

4.2. **Large deviations for finite state Markov chains.** In this section, we study the large deviations of Markov chains on a finite alphabet Σ. Let $\mathbf{\Pi} = \{\pi(i,j)\}_{i,j=1}^{|\Sigma|}$ be a stochastic matrix, and consider a Markov chain $(Y_k)_{k\ge 0}$ with transition probability $\mathbf{\Pi}$. We denote by $\mathbb{P}_\sigma^\pi$ the probability law of the chain starting at state $\sigma \in \Sigma$:

$$\mathbb{P}_\sigma^\pi(Y_1 = y_1, \cdots, Y_n = y_n) = \pi(\sigma, y_1) \prod_{i=1}^{n-1} \pi(y_i, y_{i+1}).$$

We are interested in the empirical mean

$$Z_n = \sum_{k=1}^{n} X_k,$$

where $X_k = f(Y_k)$ for a given function $f : \Sigma \to \mathbb{R}^d$. The log MGF can be analyzed using the following family of non-negative matrices:

$$\pi_\lambda(i,j) = \pi(i,j)e^{\langle \lambda, f(j)\rangle} \quad (i,j \in \Sigma).$$

Theorem 4.3 (Perron-Frobenius)**.** *Let $\mathbf{B} = \{B(i,j)\}_{i,j=1}^{|\Sigma|}$ be an irreducible matrix. Then B possesses an eigenvalue ρ (called the Perron–Frobenius eigenvalue) such that:*

(1) $\rho > 0$ is real.
(2) For any eigenvalue λ of B, $|\lambda| \le \rho$.
(3) There exist left and right eigenvectors corresponding to the eigenvalue ρ that have strictly positive coordinates.
(4) The left and right eigenvectors μ, θ corresponding to the eigenvalue ρ are unique up to a constant multiple.
(5) For every $i \in \Sigma$ and every $\varphi = \left(\varphi_1, \cdots, \varphi_{|\Sigma|}\right)$ such that $\varphi_j > 0$ for all j,

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(i,j)\varphi_j \right] = \lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} \varphi_j B^n(j,i) \right] = \log \rho.$$

Proof. See, for example, Wikipedia. □

Theorem 4.4. *Let $\{Y_k\}$ be a finite state Markov chain possessing an irreducible transition matrix $\mathbf{\Pi}$. For every $z \in \mathbb{R}^d$, define*

$$I(z) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, z\rangle - \log \rho(\mathbf{\Pi}_\lambda)\}$$

Then the empirical mean Z_n satisfies the LDP with the convex, good rate function I. Explicitly, for any set $\Gamma \subseteq \mathbb{R}^d$, and any initial state $\sigma \in \Sigma$,

$$-\inf_{z\in\mathring{\Gamma}} I(z) \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\sigma^\pi(Z_n \in \Gamma) \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\sigma^\pi(Z_n \in \Gamma) \le -\inf_{z\in\overline{\Gamma}} I(z).$$

Proof. Consider the logarithmic generating function

$$\Lambda_n(\lambda) := \log \mathbb{E}_\sigma^\pi \left[e^{\langle \lambda, Z_n\rangle} \right].$$

By Gärtner–Ellis theorem (Theorem 3.2), it is enough to check that the limit

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

exists, is finite and differentiable everywhere in $\mathbb{R}^d$, and satisfies $\Lambda(\lambda) = \log \rho(\mathbf{\Pi}_\lambda)$. To begin, note that

$$\Lambda_n(\lambda) = \log \sum_{y_1,\cdots,y_n} \mathbb{P}_\sigma^\pi(Y_1 = y_1, \cdots, Y_n = y_n) \prod_{i=1}^{n} e^{\langle \lambda, f(y_k)\rangle} = \log \sum_{y_n} (\mathbf{\Pi}_\lambda)^n(\sigma, y_n).$$

Since $\mathbf{\Pi}_\lambda$ is irreducible, we have

$$\Lambda(\lambda) = \log \rho(\mathbf{\Pi}_\lambda).$$

To show that it is differentiable, we apply the implicit function theorem. Explicitly, consider the functions $F_y : \mathbb{R} \times \mathbb{R}^{|\Sigma|} \times \mathbb{R} \to \mathbb{R} \times \mathbb{R}^{|\Sigma|}$, parametrized by $y \in \mathbb{R}^{|\Sigma|}$, defined by

$$F_y(\rho, x, \lambda) = (\langle y, x\rangle - 1, \mathbf{\Pi}_\lambda x - \rho x).$$

Clearly, F_y is continuously differentiable. Hence, it suffices to show that for every λ_0 with the associated Perron-Frobenius eigenvalue ρ_0 and left and right eigenvectors y_0 and x_0 of $\mathbf{\Pi}_{\lambda_0}$ satisfying $\langle y_0, x_0\rangle = 1$, we have (1) $F_{y_0}(\rho_0, x_0, \lambda_0) = 0$ and (2) the Jacobian matrix

$$\partial_{\rho,x} F_{y_0}(\rho_0, x_0, \lambda_0) = \begin{pmatrix} 0 & y_0^T \\ -x_0 & \mathbf{\Pi}_{\lambda_0} - \rho_0 I. \end{pmatrix}$$

is invertible. Indeed, (1) is clear, and (2) holds because if it did not, there exists $(u, v) \in \mathbb{R} \times \mathbb{R}^{|\Sigma|} \setminus \{(0, 0)\}$ such that

$$\begin{cases} \langle y_0, v\rangle = 0, \\ -ux_0 + \mathbf{\Pi}_{\lambda_0} v - \rho_0 v = 0, \end{cases}$$

which is impossible. This implies, by the implicit function theorem, that there exists a continuously differentiable function $\lambda \mapsto (\rho(\lambda), x(\lambda))$ on a neighborhood of λ_0 such that $F_{y_0}(\rho(\lambda), x(\lambda), \lambda) = 0$. In particular, $\lambda \mapsto \log \rho(\mathbf{\Pi}_\lambda) = \log \rho(\lambda)$ is continuously differentiable on a neighborhood of λ_0. The proof is finished by noting that λ_0 is arbitrary. □

As a corollary, we have the following.

Corollary 4.5. *The empirical averages $L_n^{\mathbf{Y}}(i) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_i(Y_k)$, with $i \in \Sigma$, satisfy the LDP with the good rate function*

$$I(q) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, q\rangle - \log \rho(\mathbf{\Pi}_\lambda)\} = \begin{cases} \sup_{u>0} \sum_{j\in\Sigma} q_j \log\left[\frac{u_j}{(\mathbf{u\Pi})_j}\right] & q \in M_1(\Sigma), \\ \infty & q \notin M_1(\Sigma) \end{cases}$$

where $\pi_\lambda(i, j) = \pi(i, j)e^{\lambda_j}$ and the inequality between vectors is compared entrywise.

Proof. The first equality follows immediately from Theorem 4.4 by taking

$$f = \left(\mathbb{1}_1, \mathbb{1}_2, \cdots, \mathbb{1}_{|\Sigma|}\right).$$

To prove the second inequality, we first note that one inequality is more obvious than the other: By taking $\mathbf{u}$ to be the left probability eigenvector of $\mathbf{\Pi}_\lambda$, we see the inequality "$\leq$". To prove the other inequality, assume $q \in M_1(\Sigma)$ and choose $\lambda_j = \log\left[u_j/(\mathbf{u\Pi})_j\right]$ so that $\mathbf{u\Pi}_\lambda = \mathbf{u}$ and that $\rho(\mathbf{\Pi}_\lambda) = 1$. Therefore, by definition,

$$I(q) \geq \sum_{j=1}^{|\Sigma|} q_j \log\left[\frac{u_j}{(\mathbf{u\Pi})_j}\right],$$

finishing the proof. □

We also consider a derived process $\{(Y_k Y_{k+1})\}_{k\geq 0}$ of consecutive pairs to obtain Sanov's theorem. Such a process has the transition matrix $\mathbf{\Pi}^{(2)}$ defined by

$$\pi^{(2)}((k \times \ell, i \times j)) = \mathbb{1}_\ell(i)\, \pi(i,j).$$

As is discussed in the following, one can determine the large deviations of the pair empirical measure

$$L^{\mathbf{Y}}_{n,2}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_y(Y_{i-1} Y_i)$$

For $q \in M_1(\Sigma^2)$, we write

$$q_1(i) = \sum_{j=1}^{|\Sigma|} q(i,j) \text{ and } q_2(i) = \sum_{j=1}^{|\Sigma|} q(j,i),$$

and call q shift-invariant if $q_1 = q_2$.

Theorem 4.6. *Assume that* $\mathbf{\Pi}$ *is irreducible. Then for every probability measure* $q \in M_1(\{(i,j) : \pi(i,j) > 0\})$,

$$I_2(q) = \begin{cases} \sum_i q_1(i) H(q(\cdot\,|i) \mid \pi(i,\cdot)) & \text{if } q \text{ is shift-invariant,} \\ \infty & \text{otherwise.} \end{cases}$$

Proof. By Corollary 4.5,

$$I_2(q) = \sup_{u>0} \sum_{i,j\in\Sigma} q(i,j) \log \left[\frac{u_{i,j}}{\left(\mathbf{u}\mathbf{\Pi}^{(2)}\right)_{i,j}} \right] = \sup_{u>0} \sum_{i,j\in\Sigma} q_{i,j} \log \left[\frac{u_{i,j}}{\sum_k u_{k,i}\pi_{i,j}} \right].$$

If q is not invariant, then $q_1(j_0) < q_2(j_0)$ for some j_0. For $\mathbf{u}$ such that $u(\cdot, j) = 1$ when $j \neq j_0$ and $u(\cdot, j_0) = e^\alpha$, we have $I_2(q) = \infty$ if we let $\alpha \to \infty$.

If q is invariant,

$$\sum_{i,j\in\Sigma} q(i,j) \log \left[\frac{\sum_k u_{k,i} q_2(j)}{\sum_k u_{k,j} q_1(i)} \right] = 0.$$

Hence,

$$I_2(q) - \sum_i q_1(i) H(q(\cdot\,|i) \mid \pi(i|\,\cdot)) = \sup_{u>0} \sum_{i,j\in\Sigma} q(i,j) \log \left[\frac{u_{i,j} q_1(i)}{\sum_k u_{k,i} q(i,j)} \right]$$

$$= \sup_{u>0} \left\{ -\sum_j q_2(j) H(q'(\cdot\,|j) \mid u'(\cdot\,|j)) \right\},$$

where $u'(\cdot\,|j) = \frac{u(\cdot,j)}{\sum_i u(i,j)}$ and $q'(\cdot\,|j) = \frac{q(\cdot,j)}{\sum_i q(i,j)}$. This implies

$$I_2(q) \leq \sum_{i\in\Sigma} q_1(i) H(q(\cdot\,|i) \mid \pi(i,\cdot)).$$

Taking $\mathbf{u} > 0$ approaching q proves the theorem. □

A Probability theory

A.1 Basic inequalities.

Proposition A.1 (Borel-Cantelli)**.** *Let $(S_n)_{n=1}^{\infty}$ be a sequence of events in a probability space $(\mathcal{X}, \mathcal{B}, \mu)$. Then,*

$$\mu\Big(\limsup_{n\to\infty} S_n\Big) \leq \lim_{N\to\infty} \sum_{n=N}^{\infty} \mu(S_n).$$

Proposition A.2 (Markov)**.** *Let X be a nonnegative random variable. Then, for every $\alpha > 0$,*

$$\mu(X \geq \alpha) \leq \alpha^{-1} \cdot \mathbb{E}[X].$$

Proposition A.3 (Hölder)**.** *Let X, Y be non-negative random variables. Then, for $p, q \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\mathbb{E}(X^p)^{1/p}\mathbb{E}(Y^q)^{1/q} \geq \mathbb{E}(XY),$$

where the equality holds if and only if $X = cY$ for some $c > 0$.

Proof. Without loss of generality, assume $\mathbb{E}(X^p) = \mathbb{E}(Y^q) = 1$ to paraphrase the proposition: $1 \geq \mathbb{E}(XY)$ with equality holding if and only if $X = cY$ for some $c > 0$. Under the circumstances, it is not hard to verify that Young's inequality

$$XY \leq \frac{X^p}{p} + \frac{Y^q}{q}$$

holds if and only if $X^p = Y^q$. The paraphrased proposition then follows by integrating both sides. □

Proposition A.4 (Jensen)**.** *Let X be a real-valued random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ is convex. If $\mathbb{E}(X)$ and $\mathbb{E}(\varphi(X))$ are defined, then $\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$ with the convention $\varphi(+\infty) = \lim_{x\to\infty} \varphi(x)$ and $\varphi(-\infty) = \lim_{x\to-\infty} \varphi(x)$.*

Proof. We make use of the property of convex functions that

$$\varphi(x) = \sup\{ax + b : \varphi(x) \geq ax + b \text{ for all } x \in \mathbb{R}\}. \tag{A.1}$$

The left-hand side is by definition larger than the other, and thus it remains to show the remaining. To this end, it suffices to show that for all (x_0, y_0) satisfying $\varphi(x_0) \geq y_0$, there exists a linear function $\psi(x) = a(x - x_0) + y_0$ such that $\varphi(x) \leq \psi(x)$ for all x. Essentially, this is achieved by choosing

$$\begin{aligned} a &\in \left[\sup_{x<x_0} \frac{\varphi(x) - \varphi(x_0)}{x - x_0}, \inf_{x>x_0} \frac{\varphi(x) - \varphi(x_0)}{x - x_0}\right] \\ &= \left[\liminf_{x\to x_0} \frac{\varphi(x) - \varphi(x_0)}{x - x_0}, \limsup_{x\to x_0} \frac{\varphi(x) - \varphi(x_0)}{x - x_0}\right]. \end{aligned}$$

Now that (A.1) coincides with

$$\varphi(x) = \sup\{ax + b : \varphi(x) \geq ax + b \text{ for all } x \in \mathbb{R}, a, b \in \mathbb{Q}\}, \tag{A.2}$$

one can enumerate the linear functions on the right-hand side by $(\psi_i)_{i=1}^{\infty}$ and obtain

$$\mathbb{E}(\varphi(X)) \geq \mathbb{E}\Big(\max_{1\leq i\leq n} \psi_i(X)\Big) \geq \max_{1\leq i\leq n} \psi_i(\mathbb{E}(X)).$$

If $\mathbb{E}(X) \in \mathbb{R}$, then the proposition holds by letting $n \to \infty$. If $\mathbb{E}(X) = \infty$ (similar for $\mathbb{E}(X) = -\infty$) and $\varphi(\infty) > -\infty$, then right-hand side of the above still converges to $\varphi(\infty)$, while the proposition is trivial when $\mathbb{E}(X) = \infty$ and $\varphi(\infty) = -\infty$. □

A.2 **Radon-Nikodym theorem.**

Definition A.5 (absolute continuity)**.** Let μ and ν be defined on a common measurable space $(\mathcal{X}, \mathcal{B})$. We say ν is *absolutely continuous* with respect to μ, denoted by $\nu \ll \mu$, if $\nu(A) = 0$ for every $A \in \mathcal{B}$ satisfying $\mu(A) = 0$.

Theorem A.6 (Radon-Nikodym)**.** *Suppose ν, μ are two σ-finite measures on a common measurable space $(\mathcal{X}, \mathcal{B})$ and $\nu \ll \mu$, then there exists a $\mathcal{B}$-measurable function $f : \mathcal{X} \to [0, \infty)$ such that for every $A \in \mathcal{B}$,*

$$\nu(A) = \int_A f d\mu.$$

A.3 **Laws of large numbers.**

Theorem A.7 (Weak Law of Large Numbers)**.** *Suppose $(X_i)_{i=1}^{\infty}$ are i.i.d. and $\mathbb{E}(X_1) = 0$. Then, $n^{-1}\sum_{i=1}^n X_i \to 0$ in probability.*

Proof. Equivalently, we can assume X_i is nonnegative and $\mathbb{E}(X_1) < \infty$ and prove $n^{-1}\sum_{i=1}^n X_i \to \mathbb{E}(X_1)$. Let $M > 0$ and $X_i = Y_{i,1} + Y_{i,2}$ with $Y_{i,1} = \max\{X_i, M\}$. Immediately,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_{i,1} > \varepsilon\right) \leq \frac{nM}{n^2\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

On the other hand,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_{i,2} > \varepsilon\right) \leq \frac{n\mathbb{E}\big(X_1 \mathbb{1}_{X_1 \geq M}\big)}{n\varepsilon} \to 0 \text{ as } M \to \infty.$$

The theorem is proved by combining the above. □

Lemma A.8 (Kronecker)**.** *Suppose $a_n > 0$ and $a_n \nearrow \infty$. Then $\sum_{i=1}^{\infty} a_n^{-1} x_n < \infty$ implies $a_n^{-1}\sum_{i=1}^n x_i \to 0$.*

Proof. Writing $b_n = (a_n^{-1} - a_{n-1}^{-1})$ with $a_0^{-1} := 0$, one may use summation by parts to deduce

$$a_n^{-1}\sum_{i=1}^n x_i = \sum_{i=1}^n a_i(a_i^{-1}x_i) = \sum_{i=1}^n a_i^{-1}x_i + \sum_{i=1}^{n-1} \frac{a_i - a_{i+1}}{a_n} \sum_{j=1}^i a_j^{-1}x_j.$$

The last expression converges to 0 as $n \to \infty$. □

Proposition A.9 (Kolmogorov's Criterion of SLLN)**.** *Suppose $(X_i)_{i=1}^{\infty}$ are independent such that $\mathbb{E}(X_n) = 0$ and $\sum_{i=1}^{\infty} i^{-2}\operatorname{Var}(X_i) < \infty$. Then, $n^{-1}\sum_{i=1}^n X_i \to 0$ a.s.*

Proof. By virtue of the independence, one observes that

$$\mathbb{E}\left(\left|\sum_{i=1}^{\infty}\frac{X_i}{i}\right|^2\right)\le\sum_{i=1}^{\infty}\frac{\mathrm{Var}(X_i)}{i^2}<\infty,$$

which, in particular, implies $\sum_{i=1}^{\infty}\frac{X_i}{i}$ is finite almost surely, which in turn implies the conclusion by Lemma A.8. □

Proposition A.10 (Strong Law of Large Numbers)**.** *Suppose* $(X_i)_{i=1}^{\infty}$ *are i.i.d. and* $\mathbb{E}(X_1)=0$*. Then,* $n^{-1}\sum_{i=1}^{n}X_i\to\mathbb{E}(X_1)$ *a.s.*

Proof. It suffices to prove the case $\mathbb{E}(|X_1|)<\infty$, for if otherwise, one may still apply the result to $Y_n=\min\{M,X_n\}$ for every $M>0$ if $\mathbb{E}(X_1^+)<\infty$ (resp., $Y_n=\max\{-M,X_n\}X_n$ if $\mathbb{E}(X_1^-)>-\infty$) so that $\mathbb{E}(|Y_n|)<\infty$ and that $Y_n\le X_n$ (resp., $Y_n\ge X_n$), leading to the conclusion by letting $M\to\infty$.

To begin, truncate X_n to define

$$Y_n=\mathbb{1}_{\{|X_n|\le n\}}X_n-\mathbb{E}\big(X_n\mathbb{1}_{\{|X_n|\le n\}}\big).$$

Then, verify the assumption of Proposition A.9

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\mathrm{Var}(Y_i)}{i^2}&\le\sum_{i=1}^{\infty}\frac{\mathbb{E}\big(|X_1|^2\,\mathbb{1}_{\{|X_1|\le i\}}\big)}{i^2}=\left(\mathbb{E}\left(|X_1|^2\sum_{i=\max\{1,|X_1|\}}^{n}\frac{1}{i^2}\right)\right)\\&\le\mathbb{P}(|X_1|<1)+\mathbb{E}\left(|X_1|^2\sum_{i=|X_1|}^{n}\frac{2}{i(i+1)}\right)\\&\le\mathbb{P}(|X_1|<1)+2\mathbb{E}(|X_1|).\end{aligned}$$

to deduce

$$\frac{1}{n}\sum_{i=1}^{\infty}Y_i\to 0. \tag{A.3}$$

On the other hand, by the Borel-Cantelli lemma,

$$\mathbb{P}(|X_n|>n\text{ infinitely often})\le\limsup_{n\to\infty}\sum_{i=n}^{\infty}\mathbb{P}(|X_i|\ge i)\le\limsup_{n\to\infty}\mathbb{E}(|X_1|). \tag{A.4}$$

Combining (A.3) and (A.4) gives that

$$\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}X_i=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big(Y_i+\mathbb{E}\big(|X_1|\,\mathbb{1}_{\{|X_1|\le i\}}\big)\Big)=\mathbb{E}(X_1).$$

□

B Functional analysis

Theorem B.1 (Hahn-Banach Extension Theorem)**.** *Let $\mathcal{X}$ be a real vector space. Suppose that*

- *$\mathcal{Y}$ is a linear subspace of $\mathcal{X}$,*
- *$p : \mathcal{X} \to \mathbb{R}$ is a convex functional,*
- *$f : \mathcal{Y} \to \mathbb{R}$ is linear and $f \le p|_{\mathcal{Y}}$.*

Then, there exists a linear functional $g : \mathcal{X} \to \mathbb{R}$ such that $g|_{\mathcal{Y}} = f$ and $g \le p$.

Proof. The theorem is proved by transfinite induction.

If $\mathcal{Y} \neq \mathcal{X}$, choose $y \in \mathcal{Y} \setminus \mathcal{X}$, $\mathcal{Y}' = \operatorname{span}(\{y\} \cup \mathcal{X})$ and extend f to $f' : \mathcal{Y}' \to \mathbb{R}$ by

$$f'(x + ty) = f(x) + t\alpha \text{ for all } x \in \mathcal{Y}, t \in \mathbb{R},$$

where one may choose, if possible,

$$\alpha (= f'(x')) \in \left[\sup_{t<0} \sup_{x \in \mathcal{Y}} \frac{p(x+tx') - f(x)}{t}, \inf_{t>0} \inf_{x \in \mathcal{Y}} \frac{p(x+tx') - f(x)}{t} \right]$$

so that $f'(x) \le p(x)$ on $\mathcal{Y}'$. To see the above interval is indeed nonempty, note that the convexity of p together with $p \ge f$ implies that for each $t > 0$,

$$\frac{p(z+tx') - f(z)}{t} \ge \frac{p(z'-tx') - f(z')}{-t} \text{ for all } z, z' \in \mathcal{Y}, t > 0,$$

and hence the non-emptiness follows from the finite intersection property.

To conclude the proof, consider the set $\mathcal{C}$ of all pairs $(\mathcal{Y}, f)$ such that $\mathcal{Y}$ is a subspace of $\mathcal{X}$ and that $f : \mathcal{Y} \to \mathbb{R}$ is linear on $\mathcal{Y}$ with $f \le p|_{\mathcal{Y}}$. Define further a partial order $\le$ by

$$(\mathcal{Y}, f) \le (\mathcal{Y}', f') \text{ if and only if } \mathcal{Y} \subseteq \mathcal{Y}' \text{ and } f'|_{\mathcal{Y}} = f.$$

Hence, for any chain $\{(\mathcal{Y}_i, f_i)\} \subset \mathcal{C}$, the space $\mathcal{Y} = \cup_i \mathcal{Y}_i$ is a vector space and the function $f(x) := f_i(x)$ whenever $x \in \mathcal{Y}_i$, which forms a maximal element $(\mathcal{Y}, f) \in \mathcal{C}$ of the chain. Hence, by Zorn's lemma, there exists a maximal element $(\mathcal{Y}, f) \in \mathcal{C}$, and $\mathcal{Y} = \mathcal{X}$ by the first part of the proof. □

In the context of topological vector spaces, an equivalent of Theorem B.1 is phrased in terms of separation, as follows. Recall that a *core point* x of a subset A of a vector space $\mathcal{X}$ is the point satisfying that for every $y \subset \mathcal{X}$ there exists $\varepsilon = \varepsilon(y)$ such that $y + \delta y \in A$ for all $|\delta| \le \varepsilon$, for which we denote $\operatorname{cor} A = \{x \in A : x \text{ is a core point}\}$.

Theorem B.2 (Hahn-Banach Separation Theorem)**.** *Let A and B be disjoint, nonempty, convex subsets of a real topological vector space $\mathcal{X}$. Assume that $\operatorname{cor} A \neq \varnothing$. Then there is a linear functional $f : \mathcal{X} \to \mathbb{R}$ satisfying $\sup_{a \in A} f(a) \le \inf_{b \in B} f(b)$ and $A \cup B \not\subset \{f = \alpha\}$ for all α. Moreover, if in addition $\operatorname{int} A \neq \varnothing$, then f can be chosen so that $f \in \mathcal{X}^*$ and $f(a) < f(b)$ for all $a \in A$ and $b \in B$.*

Proof of Theorem B.2 by Theorem B.1. We first prove the general case, after which we point out the necessary modifications for the case where $\operatorname{int} A \neq \varnothing$.

Let $C = A - B$ and $z \in \operatorname{cor}(A - B)$. Define $C = A - B - z$ so that $-z \notin C$ and its associated *gauge*

$$p_C(x) = \inf\{t > 0 : x \in tC\}.$$

Then, $p_C : \mathcal{X} \to \mathbb{R}$ is a sublinear functional. Indeed, since $0 \in \text{cor}C$, p_c is real-valued. Moreover, $p_C(tx) = tp_C(x)$ for all $t > 0$ by definition and $p_C(x+y) \leq p_C(x) + p_C(y)$ by convexity of C. Consider the subspace $\mathcal{Y} = \text{span}\{-z\}$ and define the linear functional $g(-tz) = t$. To apply the extension theorem, observe that $p_C(-z) \geq 1$ and thus $g \leq p|_y$, yielding a linear extension $f : \mathcal{X} \to \mathbb{R}$ of g satisfying $f \leq p_C$ by Theorem B.1. To see f separates A and B, observe that for any $a \in A$ and $b \in B$,

$$\begin{aligned} f(a) &= f(a-b-z) + f(z) + f(b) \\ &\leq p_C(a-b-z) + f(z) + f(b) \leq 1 - 1 + f(b) = f(b). \end{aligned} \tag{B.1}$$

The number α in question can be arbitrary in $[\sup_{a\in A} f(a), \inf_{b\in B} f(b)]$. To see that that $A \cup B \not\subset \{f = a\}$, note that $z = a - b$ for some $a \in A$ and $b \in B$ and $f(z) = -1 = f(a) - f(b)$ as desired.

Now if $\text{int}A \neq \varnothing$, pick $z \in \text{int}(A - B)$ and proceed as before. Under the circumstances, C is a neighborhood of 0. Hence, $|f(x)| \leq \max\{|p_C(x)|, |p_C(-x)|\} \leq 1$ for all x in the neighborhood $C \cap -C$ of 0, proving the continuity. Furthermore, $p_C(a - b - z) < 1$ due to the fact $z \in \text{int}(A - B)$, rendering (B.1) a strict inequality. □

Proof of Theorem B.1 by Theorem B.2. Let

$$A = \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} : p(x) < \alpha\} \text{ and } B = \{(y, \beta) \in \mathcal{Y} \times \mathbb{R} : f(y) \geq \beta\}.$$

Then, A and B are convex and every point in A is a core point. Thus, by Theorem B.2, there exists a linear functional $h : \mathcal{X} \to \mathbb{R}$ and numbers $\rho \in \mathbb{R}$ such that

$$\sup_{(x,\alpha)\in A} h(x) + \rho\alpha \leq \gamma := \inf_{(y,\beta)\in B} h(y) + \rho\beta \quad (\gamma \in \mathbb{R}). \tag{B.2}$$

Under the circumstances, we observe, by letting $\alpha \to \infty$, that $\rho \leq 0$. It is clear that $\rho \neq 0$, for if otherwise, the inequality (B.1) above holds if and only if $h(x) = 0$ for all $x \in \mathcal{X}$, contradicting Theorem B.2 as $h(A) = h(B) = \{0\}$. Now that $\rho < 0$, the inequality (B.2) implies that

$$\begin{aligned} \rho^{-1}h(y) + f(y) \leq \rho^{-1}\gamma \text{ for all } y \in \mathcal{Y} \quad &\Rightarrow \quad f = \rho^{-1}(\gamma - h|_y), \\ \rho^{-1}g(x) + p(x) \geq \rho^{-1}\gamma \text{ for all } x \in \mathcal{X} \quad &\Rightarrow \quad p \geq \rho^{-1}(\gamma - h). \end{aligned}$$

The proof is finished by choosing $g = \rho^{-1}(\gamma - h)$. □

Theorem B.3. *Suppose $\mathcal{X}$ is a locally convex real topological vector space. Suppose A and B are two disjoint, nonempty, convex sets in the locally convex topological vector space $\mathcal{X}$. If A is compact and B is closed, then there exists an $f \in \mathcal{X}^*$ such that $\sup_{a\in A} f(a) < \inf_{b\in B} f(b)$.*

Proof. Suppose V is a convex neighborhood of 0 such that $(A + V) \cap B = \varnothing$. Apply Theorem B.2 to $A + V$ and B to obtain $f \in \mathcal{X}^*$ such that $\sup_{a\in A} f(a) < \sup_{a\in A+V} f(a)$ since $f(A)$ is compact and $f(A + V)$ is open. □

C Basics in convex analysis

Lemma C.1 (duality)**.** *Let $\mathcal{X}$ be a locally convex topological vector space. If $f : \mathcal{X} \to (-\infty, \infty]$ is convex and lower semicontinuous, then $f^{**}(x) = f(x)$.*

Proof. We should assume, without loss of generality, that f is not identically ∞, for the lemma holds obviously otherwise. Define

$$\mathcal{E} = \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} : f(x) \le \alpha\},$$
$$\mathcal{E}^* = \{(\lambda, \alpha) \in \mathcal{X}^* \times \mathbb{R} : f^*(\lambda) \le \alpha\},$$

which are convex subsets in the locally convex topological vector spaces $\mathcal{X} \times \mathbb{R}$ and $\mathcal{X}^* \times \mathbb{R}$, respectively.

It is not hard to verify by definition that $f^{**} \le f$. Hence, it suffices to prove the other inequality. Equivalently, it asserts that if $f(x) > \alpha$, then there is $(\lambda, \beta) \in \mathcal{X}^*$ such that $\langle \lambda, x\rangle - \beta > \alpha$. Then, observe that $A = \{(x, \alpha)\}$ is compact and $B = \mathcal{E}$ is closed by lower semicontinuity of f and nonempty by the fact f is not identically ∞. Moreover, under the assumption $f(x) > \alpha$, $A \cap B = \varnothing$. These altogether allow us to apply the Hahn-Banach separation theorem (specifically, Theorem B.3) to yield some $\eta \in \mathcal{X}^*$ and $\rho \in \mathbb{R}$ satisfying

$$\gamma := \sup_{(y,\beta)\in\mathcal{E}} \langle \eta, y\rangle - \rho\beta < \langle \eta, x\rangle - \rho\alpha \tag{C.1}$$

where $\rho \ge 0$ must hold as there exists $y \in \mathcal{X}$ with $f(y) < \infty$.

If $\rho > 0$, then (C.1) implies that $\alpha < \langle \rho^{-1}\eta, x\rangle - \rho^{-1}\gamma$. Moreover, $(\rho^{-1}\eta, \rho^{-1}\gamma) \in \mathcal{E}^*$ as desired, for if otherwise, $f^*(\rho^{-1}\eta) > \rho^{-1}\gamma$ contradicting the definition of γ.

If $\rho = 0$, then for all $(\mu, \beta) \in \mathcal{E}^*$, define $(\mu_\delta, \beta_\delta) := \left(\frac{\eta}{\delta} + \mu, \frac{\gamma}{\delta} + \beta\right)$ for $\delta > 0$. In fact, $(\mu_\delta, \beta_\delta) \in \mathcal{E}^*$ since

$$f^*(\mu_\delta) - \beta_\delta \le \sup_{y\in\mathcal{X}}[\langle \mu_\delta, y\rangle - f(y) - \beta_\delta]$$
$$\le \sup_{y\in\mathcal{X}}\left[\frac{1}{\delta}(\langle \eta, y\rangle - \gamma) + (\langle \mu, y\rangle - f^*(\mu)) - f(y)\right] \le 0.$$

Moreover,

$$[\langle \eta_\delta, x\rangle - f^*(\eta_\delta)] \ge \lim_{\delta\to 0}\left[\frac{1}{\delta}(\langle \eta, x\rangle - \gamma) + (\langle \mu, x\rangle - f^*(\mu))\right] = \infty,$$

yielding the desired $\lambda = \eta_\delta$ when δ is sufficiently small. □

Lemma C.2. *Suppose $f : \mathbb{R}^d \to [0, \infty]$ is a convex, lower semicontinuous function with $\inf_{\lambda\in\mathbb{R}^d} f(\lambda) = 0$ and $0 \in \operatorname{ri} \mathcal{D}_{f^*}$, then $f(\eta) = 0$ for some $\eta \in \mathbb{R}^d$*

Proof. Note that $f^*(0) = 0$, and hence the lemma, by duality (Lemma C.1), is equivalent to the existence of $\eta \in \mathbb{R}^d$ such that $\langle \eta, x\rangle \le f^*(x)$ for all $x \in \mathbb{R}^d$. This latter claim, by convexity of f^*, is equivalent to

$$\langle \eta, x\rangle \le \lim_{\delta\to 0+} \frac{f^*(\delta x) - f^*(0)}{\delta} = \inf_{\delta>0} \frac{f^*(\delta x) - f^*(0)}{\delta} =: g(x) \text{ for all } x \in \mathbb{R}^d.$$

To prove the claim, define

$$A = \overline{\{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} : g(x) \le \alpha\}} \text{ and } B = \{(0, -1)\},$$

Obviously, B is convex, compact, and nonempty. On the other hand, A is nonempty, convex, and $A \cap B = \varnothing$. Obviously, $(0,0) \in A$. To prove convexity, note that g is a convex function and closure of a convex set is convex. Indeed,

$$g(tx + (1-t)y) = \lim_{\delta \to 0+} \frac{f^*(t\delta x + (1-t)\delta y)}{\delta}$$
$$\leq \lim_{\delta \to 0+} t \cdot \frac{f^*(\delta x)}{\delta} + (1-t) \cdot \frac{f^*(\delta y)}{\delta} = tg(x) + (1-t)g(y).$$

Finally, to prove $A \cap B = \varnothing$, we first show that $0 \in \text{ri } \mathcal{D}_g = \mathcal{D}_g$. Under the circumstances, we deduce that $g|_{\mathcal{D}_g}$ is continuous at $0 \in \text{ri } \mathcal{D}_g = \mathcal{D}_g$ and thus $A \cap B = \varnothing$ follows naturally. To this end, note that $0 \in \text{ri } \mathcal{D}_{f^*}$ implies $0 \in \text{ri } \mathcal{D}_g$. Indeed, $g(0) = 0$ and if $x \in \mathcal{D}_g$, then $f^*(x) < \infty$ and for all small $\varepsilon > 0$,

$$g(-\varepsilon x) = \inf_{\delta > 0} \frac{f^*(\delta(-\varepsilon)x)}{\delta} \leq f^*(-\varepsilon x) < \infty.$$

In particular, the above implies $0 \in \text{ri } \mathcal{D}_g$. Moreover, $0 \in \text{ri } \mathcal{D}_g$ implies also $\text{ri } \mathcal{D}_g = \mathcal{D}_g$ since $x \in \mathcal{D}_g$ if and only if $tx \in \mathcal{D}_g$ for all $t > 0$.

We now may apply Theorem B.3 to A and B to yield some $\mu \in \mathcal{X}$ such that

$$\langle \mu, 0 \rangle + \rho = \rho > \sup_{(x,\alpha) \in A} \langle \mu, x \rangle - \rho\alpha,$$

where $\rho > 0$ since $(0,0) \in A$. Hence, for every $x \in \mathcal{X}$ with $f^*(x) < \infty$, we have that

$$\langle \rho^{-1}\mu, \varepsilon x \rangle \leq g(\varepsilon x) + 1 \text{ for all } \varepsilon > 0 \Rightarrow \langle \rho^{-1}\mu, x \rangle \leq g(x).$$

The proof is concluded by choosing $\eta = \rho^{-1}\mu$. □

Lemma C.3. *Let f be an essentially smooth, convex function. If $f(0) = 0$ and $f^*(x) = 0$ for some $x \in \mathbb{R}^d$, then $0 \in \mathring{\mathcal{D}}_f$.*

Proof. Since $f(0) = 0$, it follows by convexity of f that

$$f(t\lambda) \leq tf(\lambda), \text{for all } t \in [0,1], \lambda \in \mathbb{R}^d.$$

Moreover, since $f^*(x) = 0$,

$$f(t\lambda) \geq \langle t\lambda, x \rangle - f^*(x) \geq -t|\lambda||x|.$$

Because f is essentially smooth, there exists a closed ball $\overline{B}_r(z) \subset \mathring{\mathcal{D}}_f$ in which f is differentiable. Hence,

$$M = \sup_{\lambda \in \overline{B}_r(z)} \{f(\lambda) \vee |\lambda||x|\} < \infty$$

Hence, for any $t \in (0,1]$ and any $\theta \in \overline{B}_{tr}(tz)$ different from tz, by the convexity of f,

$$f(\theta) - f(tz) \leq \frac{|\theta - tz|}{tr}(f(y) - f(tz)) \leq \frac{2tM}{tr}|\theta - tz|,$$

where $y = tz + \frac{tr}{|\theta - tz|}(\theta - tz) \in \overline{B}_{tr}(tz)$. Similarly, $f(\theta) - f(tz) \geq -\frac{2tM}{tr}|\theta - tz|$. Hence,

$$|f(\theta) - f(tz)| \leq \frac{2tM}{tr}|\theta - tz| \text{ for all } \theta \in \overline{B}_{tr}(tz).$$

Observe that $tz \in \mathring{\mathcal{D}}_f$ because of the convexity of $\mathcal{D}_f$. Hence, by assumption, $\nabla f(tz)$ exists, and by the preceding inequality, $|\nabla f(tz)| \leq 2\frac{M}{r}$. Since f is steep, by considering $t \to 0$, in which case $tz \to 0$, we conclude that $0 \in \mathring{\mathcal{D}}_f$. □

References

[1] Dembo, Amir, and Zeitouni, Ofer, *Large Deviations Techniques and Applications*, Springer Berlin Heidelberg, 2010.

Research Unit of Mathematical Sciences, University of Oulu, P.O. Box 3000, 90014 University of Oulu, Finland.
Email address: yu-liang.wu@oulu.fi
URL: s92077.github.io