

# **MA4L3 - Large Deviation Theory**

## **Lecture Notes**

**Stefan Adams**

2023, update 29.03.2023

## Contents

<b>1</b>	<b>Introduction and Cramér's theorem</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Framework of large deviation theory . . . . .	4
1.3	General Cramér Theorem in $\mathbb{R}$ . . . . .	6
<b>2</b>	<b>Methods of types and Sanov's theorem</b>	<b>13</b>
2.1	The empirical measure LDP - Sanov's theorem . . . . .	15
2.2	The pair empirical measure . . . . .	17
2.3	Cramer's theorem for finite subsets in $\mathbb{R}$ . . . . .	20
<b>3</b>	<b>General Theory</b>	<b>22</b>
3.1	Basic theory . . . . .	22
3.2	Contraction principle . . . . .	25
3.3	Varadhan's Integral Lemma . . . . .	27
3.4	Bryc's Inverse Varadhan Lemma . . . . .	30
<b>4</b>	<b>The Gärtner-Ellis theorem</b>	<b>32</b>
4.1	Gärtner-Ellis for $\mathbb{R}^d$ . . . . .	32
4.2	A general upper bound - topological vector spaces . . . . .	37
4.3	Summary: general Cramér's theorem and general Sanov's theorem . . . . .	40
<b>5</b>	<b>Large deviations for Markov chains</b>	<b>41</b>
5.1	Discrete time finite state space Markov chains . . . . .	41
5.2	Pair empirical measures for Markov chains . . . . .	45
5.3	Markov process with continuous time and finite state space . . . . .	49
<b>6</b>	<b>The Gibbs Conditioning principle</b>	<b>49</b>
6.1	Conditional limit theorem for i.i.d. sequences . . . . .	50
6.2	Example: microcanonical ensemble for one-dimensional Ising model . . . . .	53
<b>7</b>	<b>Sample path large deviations</b>	<b>54</b>
7.1	Mogulskii's theorem . . . . .	54
7.2	Schilder's theorem . . . . .	60
7.3	Application: pinning reward for polymer chains and random interfaces . . . . .	63
<b>A</b>	<b>Preliminaries on Probability Theory</b>	<b>70</b>
A.1	Random variables . . . . .	70
A.2	Classical Inequalities . . . . .	73
A.3	$L^p$ -spaces . . . . .	74
<b>B</b>	<b>Modes of Convergence</b>	<b>76</b>
<b>C</b>	<b>Law of large numbers and the central limit theorem</b>	<b>78</b>
<b>D</b>	<b>Normal distribution</b>	<b>79</b>
<b>E</b>	<b>Gaussian integration formulae</b>	<b>80</b>

## 1 Introduction and Cramér's theorem

### 1.1 Introduction

**Example 1.1 (Coin-tossing)** Let  $(X_i)_{i \in \mathbb{N}}$  be an i.i.d. sequence with  $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = \frac{1}{2}$  and denote  $\widehat{S}_N := X_1 + \cdots + X_N$ . Then, for all  $x > \frac{1}{2}$ ,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(\widehat{S}_N \geq xN) = -I(x), \quad (1.1)$$

where

$$I(z) = \begin{cases} \log 2 + z \log z + (1 - z) \log(1 - z) & ; \text{ for } z \in [0, 1], \\ +\infty & ; \text{ for } z \notin [0, 1]. \end{cases} \quad (1.2)$$

We shall prove (1.1). For  $x > 1$  we have  $\{\widehat{S}_N \geq xN\} = \emptyset$ , and thus both sides are  $-\infty$ . For  $x \in (\frac{1}{2}, 1]$  we write

$$\mathbb{P}(\widehat{S}_N \geq xN) = 2^{-N} \sum_{k \geq xN} \binom{N}{k},$$

which yields the estimate

$$2^{-N} Q_N(x) \leq \mathbb{P}(\widehat{S}_N \geq xN) \leq (N + 1) 2^{N+1} Q_N(x),$$

where

$$Q_N(x) = \max_{k \geq xN} \binom{N}{k}.$$

The maximum is attained at  $k = \lceil xN \rceil$ , the smallest integer  $\geq xN$ . Stirling's formula  $N! = N^N e^{-N} \sqrt{2\pi N} (1 + O(1/N))$  now allows us to infer that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Q_N(x) = -x \log x + (1 - x) \log(1 - x).$$

Now our upper and lower bounds merge on an exponential scale as  $N \rightarrow \infty$ , and we arrive at our statement. Our results actually deals with large deviations in the upward direction because  $\mathbb{E}[X_1] = \frac{1}{2}$  and  $x > \frac{1}{2}$ . It is clear from the symmetry  $I(1 - z) = I(z)$  that the same holds for  $\mathbb{P}(\widehat{S}_N \leq xN)$  with  $x < \frac{1}{2}$ . The function  $z \mapsto I(z)$  is called the *rate function*. Note that the rate function is infinite outside of  $[0, 1]$ , finite and strictly convex inside  $[0, 1]$ , and has a unique zero at  $z = \frac{1}{2}$ . The zero corresponds to the Strong Law of Large Number (SLLN). Indeed, (1.1) implies that

$$\sum_{N \in \mathbb{N}} \mathbb{P}\left(\left|\frac{1}{N} \widehat{S}_N - \frac{1}{2}\right| < \delta\right) < \infty; \quad \text{for all } \delta > 0,$$

and so the SLLN follows via the Borel-Cantelli lemma. By computation we see that  $I'(\frac{1}{2}) = 0$ ,  $I''(\frac{1}{2}) = 4 = \frac{1}{\sigma^2}$ . ♣

**Theorem 1.2 (Cramér's theorem)** Let  $(X_i)_{i \in \mathbb{N}}$  be a sequence of i.i.d. real-valued random variables with law  $P \in \mathcal{M}_1(\mathbb{R})$  satisfying

$$M(\lambda) := \mathbb{E}[e^{\lambda X_1}] < \infty \quad \text{for all } \lambda \in \mathbb{R}, \quad (1.3)$$

and let  $\hat{S}_N$  be their partial sum and  $m \in \mathbb{R}$  their mean, and denote  $\Lambda(\lambda) = \log M(\lambda)$ ,  $\lambda \in \mathbb{R}$ , the logarithmic moment generating function. Then, for any  $x > m$ , we have

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(\hat{S}_N \geq Nx) = -I(x), \quad (1.4)$$

where

$$I(x) = \Lambda^*(x) := \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}. \quad (1.5)$$

**Proof.**

Upper bound: We use Chebyshev's inequality, but in an optimised form. Recall that for any non-negative, increasing function  $\Psi$  we have the following version of that inequality,

$$\mathbb{P}(\hat{S}_N \geq Nx) \leq \mathbb{P}(\Psi(\hat{S}_N) \geq \Psi(Nx)) \leq \frac{1}{\Psi(Nx)} \mathbb{E}[\Psi(\hat{S}_N)].$$

We choose  $\Psi(x) = e^{\lambda x}$  with  $\lambda \geq 0$  and optimise over  $\lambda \geq 0$  later. This yields, writing  $S_N := \frac{1}{N} \hat{S}_N$  for the *empirical mean*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \geq x) \leq -\lambda x + \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[\exp(\lambda \hat{S}_N)] \leq -\lambda x + \Lambda(\lambda).$$

We optimise over  $\lambda \geq 0$  to get the best upper bound,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \geq x) \leq -\sup_{\lambda \geq 0} \{\lambda x - \Lambda(\lambda)\}.$$

We show that we can optimise over all  $\lambda \in \mathbb{R}$  on the right hand side. For this we show that  $\lambda x - \Lambda(\lambda)$  is negative for  $\lambda < 0$ . For  $\lambda = 0$ , the expression in the curly brackets vanishes. For  $\lambda < 0$  and  $x > m$  it holds that

$$\lambda x - \Lambda(\lambda) \leq \lambda m - \Lambda(\lambda) \leq \Lambda^*(m) = 0, \quad (1.6)$$

which immediately implies our statement. To see (1.6), use Jensen's inequality A.11 to get  $\Lambda(\lambda) \geq \lambda m$  for all  $\lambda$ , and thus  $\lambda m - \Lambda(\lambda) \leq 0$  for all  $\lambda \in \mathbb{R}$ , and so we know that  $\Lambda^*(m) \leq 0$ . On the other hand we know that  $\Lambda^* \geq 0$  due to the fact the the expression in the curly brackets vanishes for  $\lambda = 0$ , and thus we get  $\Lambda^*(m) = 0$ . We conclude with

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \geq x) \leq -\sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}. \quad (1.7)$$

Lower bound: We employ a change of measure or tilting method. The idea is to change the law such that the event in question has probability approximately of one, that is, the event is a large number event under the new measure. Recall that  $P \in \mathcal{M}_1(\mathbb{R})$  is the law of  $X_1$ , and define a new law  $Q \in \mathcal{M}_1(\mathbb{R})$  via a Radon-Nikodym density, i.e.,

$$Q(dx) = e^{-\Lambda(\lambda) + \lambda x} P(dx). \quad (1.8)$$

Assume that for all  $\varepsilon > 0$  there exists a  $\lambda > 0$  such that

$$\mathbb{Q}(x + \varepsilon > S_N \geq x) \rightarrow 1 \quad \text{as } N \rightarrow \infty, \quad (1.9)$$

where  $\mathbf{Q} = Q^{\otimes N}$  is just the product measure. We justify our assumption (1.9) later. Under this assumption we obtain the lower bound as follows, using that

$$\mathbb{P}(x + \varepsilon < S_N \leq x) = \mathbb{E}_{\mathbf{Q}}[e^{N\Lambda(\lambda) - \lambda \hat{S}_N} \mathbb{1}\{x + \varepsilon > S_N \geq x\}],$$

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \geq x) &\geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(x + \varepsilon > S_N \geq x) \geq \Lambda(\lambda) - \lambda(x + \varepsilon) \\ &\quad + \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbf{Q}(x + \varepsilon > S_N \geq x) = \Lambda(\lambda) - \lambda x - \lambda \varepsilon \quad (1.10) \\ &\geq -\Lambda^*(x + \varepsilon). \end{aligned}$$

We conclude with the lower bound by using the lower semicontinuity of  $\Lambda^*$  and  $\varepsilon \downarrow 0$ .

To prove our assumption (1.9) above it suffices to show that  $\lambda > 0$  can be chosen such that

$$\Lambda'(\lambda) = e^{-\Lambda(\lambda)} \mathbb{E}[x e^{\lambda X}] = \mathbb{E}_{\mathbf{Q}}[X].$$

To obtain (1.9) we need to have that

$$\Lambda'(\lambda) = \mathbb{E}_{\mathbf{Q}}[X] = x + \frac{\varepsilon}{2}. \quad (1.11)$$

We know that  $\Lambda'(0) = m$  and  $\Lambda'(\infty) = \text{ess sup } X =: M$ , which follows with Exercise 1.3. Recall that  $\text{ess sup } X$  is the smallest number  $\alpha$  such that  $\mathbb{P}(X > \alpha) = 0$ . If  $m < x < M$ , by the Intermediate Value Theorem, we can find for all  $\varepsilon > 0$  a  $\lambda > 0$  with  $\Lambda'(\lambda) = x + \frac{\varepsilon}{2}$ . To complete our argument note that, in case  $M < \infty$ , for  $x > M$  both sides of the statement in the theorem are  $-\infty$  because  $\mathbb{P}(S_N \geq x > M) = 0$  and  $\mathbb{E}[e^{\lambda X}] = \mathbb{E}[e^{\lambda X} \mathbb{1}\{X \leq M\}]$  and  $\lambda(x - M) \rightarrow \infty$  as  $\lambda \rightarrow \infty$ . If  $x = M$  we have

$$\mathbb{P}(S_N \geq M) = \mathbb{P}(X = M),$$

and thus the right hand side is  $\log \mathbb{P}(X = M)$ , and for the left hand side we get the same by considering  $\mathbb{E}[e^{\lambda M} \mathbb{1}\{X = M\}]$ . □

**Exercise 1.3 (The essential supremum and  $\Lambda'$ )** Let  $X$  be a  $\mathbb{R}$ -valued random variable such that

$$M(\lambda) = \mathbb{E}[e^{\lambda X}] < \infty \quad \text{for all } \lambda \in \mathbb{R},$$

and  $\Lambda(\lambda) = \log M(\lambda)$ . Show that

$$\Lambda'(\lambda) \rightarrow \text{ess sup } X \quad \text{as } \lambda \rightarrow \infty.$$



**Solution.** TA class in Week 2.



## 1.2 Framework of large deviation theory

**Proposition 1.4 (Laplace principle)** Fix a sequence  $a_n \rightarrow \infty$  as  $n \rightarrow \infty$  and a finite number  $N$  of nonnegative sequences  $b_n^{(1)}, \dots, b_n^{(N)}$ . Then

$$\lim_{n \rightarrow \infty} \frac{1}{a_n} \log \left( \sum_{i=1}^N b_n^{(i)} \right) = \max_{1 \leq i \leq N} \left\{ \limsup_{n \rightarrow \infty} \frac{1}{a_n} \log b_n^{(i)} \right\}. \quad (1.12)$$

**Proof.** For every fixed  $N \in \mathbb{N}$  we have

$$0 \leq \log \sum_{i=1}^N b_n^{(i)} - \max_{1 \leq i \leq N} \log b_n^{(i)} \leq \log N.$$

Dividing by  $a_n$  and taking the  $\limsup_{n \rightarrow \infty}$  shows that

$$\lim_{n \rightarrow \infty} \frac{1}{a_n} \log \left( \sum_{i=1}^N b_n^{(i)} \right) = \limsup_{n \rightarrow \infty} \frac{1}{a_n} \max_{1 \leq i \leq N} \log b_n^{(i)} = \max_{1 \leq i \leq N} \left\{ \limsup_{n \rightarrow \infty} \frac{1}{a_n} \log b_n^{(i)} \right\}.$$

□

**Corollary 1.5** We write  $\overline{A}$  and  $\overset{\circ}{A}$  for the closure and the interior respectively, of a Borel  $A \subset \mathbb{R}$ . Under the assumptions of Theorem 1.2 above, for every Borel set  $A \subset \mathbb{R}$ , it holds that

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \in A) &\leq - \inf_{x \in \overline{A}} \{\Lambda^*(x)\}, \\ \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \in A) &\geq - \inf_{x \in \overset{\circ}{A}} \{\Lambda^*(x)\}. \end{aligned} \quad (1.13)$$

**Proof.** TA class Week 2.

□

The last results motivates the upcoming definition large deviation principle. The *large deviation principle* (LDP) characterises the limiting behaviour, as  $N \rightarrow \infty$ , of a sequence of probability measures  $(\mu_N)_{N \in \mathbb{N}}$  on  $(E, \mathcal{F})$  in terms of a *rate function*. Here, we assume that  $(E, d)$  is a Polish space, i.e., a complete metric space, and that  $\mathcal{F}$  a  $\sigma$ -algebra. Frequently we will choose the Borel  $\sigma$ -algebra on  $E$ , denoted  $\mathcal{B}_E$  by default and only keep the general  $\mathcal{F}$  for our definitions.

**Notation 1.6** In the following standard notation is used throughout the lecture; for any set  $A \subset E$ ,  $\overline{A}$  denotes the closure of  $A$ ,  $\overset{\circ}{A}$  the interior of  $A$ , and  $A^c$  the complement of  $A$ . The infimum of a function over an empty set is interpreted as  $\infty$ .

**Definition 1.7 (Rate function)** A *rate function*  $I$  is a lower semicontinuous mapping  $I: E \rightarrow [0, \infty]$ , that is, for all  $\alpha \in [0, \infty)$ , the level set  $\mathcal{L}_I(\alpha) = \{x \in E: I(x) \leq \alpha\}$  is closed. A *good rate function* is a rate function for which all the level sets  $\mathcal{L}_I(\alpha)$  are compact subsets of  $E$ . The (effective) *domain* of  $I$ , denoted  $\mathcal{D}_I = \{x \in E: I(x) < \infty\}$ , is the set of points in  $E$  of finite rate.

**Definition 1.8 (Large deviation principle (LDP))** Suppose  $(\mu_N)_{N \in \mathbb{N}}$  is a sequence of probability measures  $\mu_N \in \mathcal{M}_1(E, \mathcal{F})$ . The sequence  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the large deviation principle (LDP) with speed or rate  $N$  and rate function  $I$  if, for all  $A \in \mathcal{F}$ ,

$$-\inf_{x \in \overset{\circ}{A}} \{I(x)\} \leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(A) \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(A) \leq -\inf_{y \in \bar{A}} \{I(y)\}. \quad (1.14)$$

The right- and left-hand side of (1.14) are referred to as the upper and lower bounds, respectively. A set  $A \in \mathcal{F}$  is called *I continuity set* if

$$\inf_{x \in \overset{\circ}{A}} \{I(x)\} = \inf_{x \in \bar{A}} \{I(x)\} =: I_A. \quad (1.15)$$

**Remark 1.9** (a) Note that in (1.14) the  $\sigma$ -algebra  $\mathcal{F}$  need not necessarily be the Borel  $\sigma$ -algebra  $\mathcal{B}_E$ . In principle there can be a separation between the sets on which probability may be assigned and the values of the bounds. So (1.14) makes sense even if some open sets are not measurable. However, for the remaining lecture we shall always assume that  $\mathcal{B}_E \subset \mathcal{F}$ .

(b) Why the lower and upper bound differ in (1.14)? Suppose that we are dealing with non-atomic measures, i.e.,  $\mu_N(\{x\}) = 0$  for every  $x \in E$ . So if we want the lower bound in (1.14) to hold with the infimum over  $A$  instead of  $\overset{\circ}{A}$ , we would conclude that  $I(x) = \infty$  for every  $x \in E$  and thus  $I \equiv \infty$ , contradicting the upper bound in (1.14) because  $\mu_N(E) = 1$  for all  $N \in \mathbb{N}$ .

(c) For *I* continuity sets  $A$  it holds that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(A) = -I_A. \quad (1.16)$$

(d) Since  $\mu_N(E) = 1$  for all  $N \in \mathbb{N}$ , it is necessary that  $\inf_{x \in E} \{I(x)\} = 0$  for the upper bound to hold. When  $I$  is a good rate function there exists at least one point  $x \in E$  for which  $I(x) = 0$ .

(e) Suppose that  $I$  is a rate function. Then (1.14) is equivalent to the following two bounds:

(i) **Upper bound:** For every  $\alpha \in (0, \infty)$  and every measurable set  $M$  with  $\bar{M} \subset \mathcal{L}_I(\alpha)^c$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(M) \leq -\alpha. \quad (1.17)$$

(ii) **Lower bound:** For any  $x \in \mathcal{D}(I)$  and any measurable  $M$  with  $x \in \overset{\circ}{M}$ ,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(M) \geq -I(x). \quad (1.18)$$

◇

### 1.3 General Cramér Theorem in $\mathbb{R}$

In this section we explore some generalisation of Theorem 1.2, and in particular our assumption 1.3.

**Definition 1.10** (a) The *logarithmic moment generating function* for  $\mu \in \mathcal{M}_1(\mathbb{R})$  is the mapping

$$\Lambda_\mu: \mathbb{R} \rightarrow (-\infty, \infty], \quad \lambda \mapsto \Lambda_\mu(\lambda) = \log \left( \int_{\mathbb{R}} e^{\lambda x} \mu(dx) \right).$$

We write  $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}]$  when  $\mu$  known and clear from the context being the law of  $X$ . The domain is  $\mathcal{D}_\Lambda = \{\lambda \in \mathbb{R}: \Lambda(\lambda) < \infty\}$ .

(b) The *Legendre-Fenchel transform* of  $\Lambda_\mu$  is denoted  $\Lambda_\mu^*$  and is defined as

$$\Lambda_\mu^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda_\mu(\lambda)\}, \quad x \in \mathbb{R}. \quad (1.19)$$

We drop the index  $\mu$  when the underlying probability measure is clear from the context. The domain is  $\mathcal{D}_{\Lambda^*} = \{x \in \mathbb{R}: \Lambda^*(x) < \infty\}$ .

We consider now that following setting. Let  $(X_i)_{i \in \mathbb{N}}$  be i.i.d.  $\mathbb{R}$ -valued random variables with law  $\mu \in \mathcal{M}_1(\mathbb{R})$ . We write  $\mathcal{M}_1(\mathbb{R})$  for  $\mathcal{M}_1(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ . The following lemma states all the properties of  $\Lambda^*$  and  $\Lambda$  that are needed to prove Theorem 1.12 which is our general version of Cramér's theorem in  $\mathbb{R}$ .

**Lemma 1.11 (Properties of  $\Lambda$  and  $\Lambda^*$ )** (a)  $\Lambda$  is a convex function and  $\Lambda^*$  is a convex rate function.

(b) (i) If  $\mathcal{D}_\Lambda = \{0\}$ , then  $\Lambda^* \equiv 0$ .

(ii) If  $\Lambda(\lambda) < \infty$  for some  $\lambda > 0$ , then  $m = \mathbb{E}[X_1] = \int_{\mathbb{R}} x \mu(dx) < \infty$ , and for all  $x \geq m$ ,

$$\Lambda^*(x) = \sup_{\lambda \geq 0} \{\lambda x - \Lambda(\lambda)\} \quad (1.20)$$

is, for  $x > m$ , a nondecreasing function.

(iii) If  $\Lambda(\lambda) < \infty$  for some  $\lambda < 0$ , then  $m > -\infty$ , and for all  $x \leq m$ ,

$$\Lambda^*(x) = \sup_{\lambda \leq 0} \{\lambda x - \Lambda(\lambda)\} \quad (1.21)$$

is, for  $x < m$ , a nonincreasing function.

(iv) When  $m \in \mathbb{R}$ ,  $\Lambda^*(m) = 0$ , and always,

$$\inf_{x \in \mathbb{R}} \{\Lambda^*(x)\} = 0. \quad (1.22)$$

(c)  $\Lambda$  is differentiable in  $\mathring{\mathcal{D}}_\Lambda$  with

$$\Lambda'(\lambda) = \frac{1}{M(\lambda)} \mathbb{E}[X_1 e^{\lambda X_1}] \quad (1.23)$$

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

**Proof.**

(a) By Hölder's inequality, for any  $\alpha \in [0, 1]$ ,

$$\begin{aligned}\Lambda(\alpha\lambda_1 + (1 - \alpha)\lambda_2) &= \log \mathbb{E}[(e^{\lambda_1 X_1})^\alpha (e^{\lambda_2 X_1})^{1-\alpha}] \leq \log \left( \mathbb{E}[e^{\lambda_1 X_1}]^\alpha \mathbb{E}[e^{\lambda_2 X_1}]^{1-\alpha} \right) \\ &= \alpha\Lambda(\lambda_1) + (1 - \alpha)\Lambda(\lambda_2),\end{aligned}$$

implying convexity for  $\Lambda$ .

$$\begin{aligned}\alpha\Lambda^*(x_1) + (1 - \alpha)\Lambda^*(x_2) &= \sup_{\lambda \in \mathbb{R}} \{\alpha\lambda x_1 - \alpha\Lambda(\lambda)\} + \sup_{\lambda \in \mathbb{R}} \{(1 - \alpha)\lambda x_2 - (1 - \alpha)\Lambda(\lambda)\} \\ &\geq \sup_{\lambda \in \mathbb{R}} \{(\alpha x_1 + (1 - \alpha)x_2)\lambda - \alpha\Lambda(\lambda)\} = \Lambda^*(\alpha x_1 + (1 - \alpha)x_2).\end{aligned}$$

Furthermore,  $\Lambda(0) = 0$ , and so  $\Lambda^*(x) \geq 0x - \Lambda(0) = 0$ . Suppose that  $x_N \rightarrow x$  as  $N \rightarrow \infty$ . Then, lower semicontinuity of  $\Lambda^*$  follows since

$$\liminf_{N \rightarrow \infty} \Lambda^*(x_N) \geq \liminf_{N \rightarrow \infty} (\lambda x_N - \Lambda(\lambda)) = \lambda x - \Lambda(\lambda).$$

Hence,  $\Lambda^*$  is a convex rate function.

(b) Clearly,  $\mathcal{D}(\Lambda) = \{0\}$  implies  $\Lambda^*(x) = \Lambda(0) = 0$  for all  $x \in \mathbb{R}$ . For all  $\lambda \in \mathbb{R}$ , by Jensen's inequality,

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X_1}] \geq \mathbb{E}[\log e^{\lambda X_1}] = \lambda m,$$

and thus if  $\Lambda(\lambda) < \infty$  we get that  $m < \infty$ . If  $m = -\infty$ , then  $\Lambda(\lambda) = \infty$  for  $\lambda$  negative, and (1.20) trivially holds. In case  $m \in \mathbb{R}$ , we obtain with the previous estimate that  $\lambda m - \Lambda(\lambda) \leq 0$  for all  $\lambda \in \mathbb{R}$ , and thus  $\Lambda^*(m) = 0$ . We also have that for  $x \geq m$  and  $\lambda < 0$ ,

$$\lambda x - \Lambda(\lambda) \leq \lambda m - \Lambda(\lambda) \leq \Lambda^*(m) = 0,$$

and therefore (1.20) follows. The monotonicity of  $\Lambda^*$  on  $[m, \infty)$  (nondecreasing) follows from (1.20), since for every  $\lambda \geq 0$ , the function  $\lambda x - \Lambda(\lambda)$  is nondecreasing as a function of  $x$ . The complementary case that  $\Lambda(\lambda) < \infty$  for some negative  $\lambda < 0$  follows by considering the logarithmic moment generating function of  $-X_1$ . We are finally left to show that  $\inf_{x \in \mathbb{R}} \Lambda^*(x) = 0$ . This is immediate from our reasoning above, as for  $\mathcal{D}(\Lambda) = \{0\}$  we have  $\Lambda^* \equiv 0$  and for  $m \in \mathbb{R}$  we have  $\Lambda^*(m) = 0$ . We shall now consider the case  $m = -\infty$  while  $\Lambda(\lambda) < \infty$  for some positive  $\lambda > 0$ . Then, by Chebychev's inequality and (1.20),

$$\log P(X_1 \geq x) = \log \mu([x, \infty)) \leq \inf_{\lambda \geq 0} \log \mathbb{E}[e^{\lambda(X_1 - x)}] = -\sup_{\lambda \geq 0} \{\lambda x - \Lambda(\lambda)\} = -\Lambda^*(x).$$

Hence,

$$\lim_{x \rightarrow -\infty} \Lambda^*(x) \leq \lim_{x \rightarrow -\infty} (-\log \mu([x, \infty))) = 0,$$

and  $\inf_{x \in \mathbb{R}} \Lambda^*(x) = 0$  follows. The only case left to discuss is that of  $m = \infty$  while  $\Lambda(\lambda) < \infty$  for some negative  $\lambda < 0$ . This is again settled by considering the logarithmic moment generating functions of  $-X_1$ .

(c) The identity (1.23) follows by interchanging the order of differentiation and integration which we justify by the dominated convergence theorem as follows:

$$f_\varepsilon(x) = (e^{(\eta+\varepsilon)x} - e^{\eta x})/\varepsilon$$

converges pointwise to  $xe^{\eta x}$  as  $\varepsilon \rightarrow 0$ , and, for  $\delta > 0$  small enough,

$$|f_\varepsilon(x)| \leq e^{\eta x}(e^{\delta|\eta|} - 1)/\delta =: h(x), \quad \varepsilon \in (-\delta, \delta),$$

and  $\mathbb{E}[h(X_1)] < \infty$ . Let  $\Lambda'(\eta) = y$  and define  $g(\lambda) := \lambda y - \Lambda(\lambda)$ . Note that  $g$  is concave and  $g'(\eta) = 0$ , and thus it follows that  $g(\eta) = \sup_{\lambda \in \mathbb{R}} g(\lambda) = \Lambda^*(y)$ .  $\square$

**Theorem 1.12 (Cramér's theorem in  $\mathbb{R}$  - general version)** *Let  $(X_i)_{i \in \mathbb{N}}$  be i.i.d.  $\mathbb{R}$ -valued with law  $\mu \in \mathcal{M}_1(\mathbb{R})$  and denote  $\mu_N = \mu^{\otimes N} \circ S_N^{-1}$ , where  $S_N$  is the empirical mean. Then the sequence  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with the convex rate function  $\Lambda^*$ , that is,*

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(A) &\leq - \inf_{x \in A} \{\Lambda^*(x)\}, \text{ for any closed set } A \subset \mathbb{R}; \\ \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu(G) &\geq - \inf_{x \in G} \{\Lambda^*(x)\}, \text{ for any open set } G \subset \mathbb{R}. \end{aligned} \quad (1.24)$$

**Proof of Theorem 1.12.** Proof of the upper bound in (1.24): Let  $\emptyset \neq F \subset \mathbb{R}$  closed. The upper bound certainly trivially holds when  $I_F := \inf_{x \in F} \Lambda^*(x) = 0$ . Thus assume that  $I_F > 0$ . By part (b) of Lemma 1.11 it follows that  $m$  exists (possibly as extended real number). For all  $x$  and  $\lambda \geq 0$ , an application of the (exponential with function  $e^{\lambda x}$ ,  $\lambda \geq 0$ ) Chebychev inequality yields

$$\mu_N([x, \infty)) = P(S_N \geq x) \leq \mathbb{E}[e^{N(S_N - x)}] = e^{-N\lambda x} \prod_{i=1}^m \mathbb{E}[e^{\lambda X_i}] = e^{-N(\lambda x - \Lambda(\lambda))}.$$

Now, if the mean  $m < \infty$ , then by (1.20) in Lemma 1.11, for every  $x > m$ , we obtain an upper by optimising over all  $\lambda \in \mathbb{R}$ , i.e.,

$$\mu_N([x, \infty)) \leq e^{-N\Lambda^*(x)} \quad \text{for every } x > m. \quad (1.25)$$

This follows from the proof of (1.20). Equivalently, if  $m > -\infty$  and  $x < m$ , we can use an estimate via the exponential Chebychev inequality for  $\lambda > 0$ ,

$$P(-S_N \geq -x) \leq \mathbb{E}[\exp(-N(\lambda(-S_N) - \tilde{\Lambda}(\lambda)))] ,$$

where  $\tilde{\Lambda}$  is the logarithmic moment generating function for  $-X_1$ . Note that  $\tilde{\Lambda}(-\lambda) = \Lambda(\lambda)$ . Hence,

$$P(-S_N \geq -x) \leq \exp(-N \sup_{\lambda \leq 0} \{\lambda x - \Lambda(\lambda)\}) = \exp(-N\Lambda^*(x)),$$

as for  $\lambda > 0$ , due to  $x < m$  we have

$$\lambda x - \Lambda(\lambda) \leq \lambda m - \Lambda(\lambda) \leq \Lambda^*(m) = 0,$$

and thus optimising for positive  $\lambda$  is not changing the supremum over  $\lambda \leq 0$  as long as  $x < m$ . Therefore,

$$\mu_N((-\infty, x]) \leq e^{-N\Lambda^*(x)}, \quad \text{for every } x < m. \quad (1.26)$$

After this preparation, we handle the three cases (i)  $m \in \mathbb{R}$ , (ii)  $m = -\infty$  and (iii)  $m = +\infty$  separately.

(i) Suppose  $m \in \mathbb{R}$ . Then, as seen in Lemma 1.11,  $\Lambda^*(m) = 0$ , and as  $I_F > 0$ , the mean  $m$  must be contained in the open set  $F^c$ . Denote  $(x_-, x_+)$  the union of all open intervals in  $F^c$  containing  $m$ . Clearly,  $x_- < x_+$  and either  $x_- \in \mathbb{R}$  or  $x_+ \in \mathbb{R}$  since  $F$  is nonempty. If  $x_- \in \mathbb{R}$ , then  $x_- \in F$ , and consequently  $\Lambda^*(x_-) \geq I_F$ . Likewise,  $\Lambda^*(x_+) \geq I_F$  whenever  $x_+ \in \mathbb{R}$ . Now we apply (1.25) for  $x = x_+$  and (1.26) for  $x = x_-$  such that the union of events bounds ensures that

$$\mu_N(F) \leq \mu_N((-\infty, x_-]) + \mu_N([x_+, \infty)) \leq 2e^{-NI_F},$$

and the upper bound in (1.24) follows as  $N \rightarrow \infty$ .

(ii) Suppose now  $m = -\infty$ . As  $\Lambda^*$  is nondecreasing, it follows from  $\inf_{x \in \mathbb{R}} \Lambda^*(x) = 0$  that  $\lim_{x \rightarrow -\infty} \Lambda^*(x) = 0$ , and hence  $x_* = \inf\{x \in \mathbb{R} : x \in F\}$  is finite for otherwise  $I_F = 0$ . As  $F$  is closed,  $x_* \in F$ , and thus  $\Lambda^*(x_*) \geq I_F$ . Noting that  $F \subset [x_*, \infty)$  and using (1.25) for  $x = x_*$ , we obtain the large deviations upper bound in (1.24) of Theorem 1.12. The third case (iii)  $m = +\infty$  follows analogously to the second case.

Proof of the lower bound in (1.24) of Theorem 1.12: The key idea is to prove that for every  $\delta > 0$  and every probability measure  $\mu \in \mathcal{M}_1(\mathbb{R})$ ,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N((-\delta, \delta)) \geq \inf_{\lambda \in \mathbb{R}} \{\Lambda(\lambda)\} = -\Lambda^*(0), \quad (1.27)$$

where  $\mu_N$  is the law of  $S_N$  under  $\mu^{\otimes N}$ . The proof of (1.27) will keep us busy below, it is actually the major part of the work. Suppose now that (1.27) holds. We can then quickly see that the lower bound in (1.24) holds. First recall that we write  $\Lambda$  for the logarithmic moment generating function for a real-valued random variable  $X$ , if we consider the random variable  $Y = X - x$ ,  $x \in \mathbb{R}$ , we write  $\Lambda_Y$  for the logarithmic moment generating function. It is easy to see that then  $\Lambda_Y(\lambda) = \Lambda(\lambda) - \lambda x$ , and hence with  $\Lambda_Y^*(y) = \Lambda^*(y + x)$  for all  $y \in \mathbb{R}$ , it follows from (1.27) that for every  $x \in \mathbb{R}$  and every  $\delta > 0$ ,

$$\liminf_{N \rightarrow \infty} \mu_N((x - \delta, x + \delta)) \geq -\Lambda^*(x). \quad (1.28)$$

For any open set  $G \subset \mathbb{R}$ , any element  $x \in G$ , and any  $\delta > 0$  small enough one has  $(x - \delta, x + \delta) \subset G$ . Thus we obtain

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(G) \geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N((x - \delta, x + \delta)) \geq -\Lambda^*(x),$$

and we can optimise the right hand side of (1.28) over all  $x' \in G$  to obtain the large deviation lower bound in (??).

**Proof of (1.27):** We split the proof according to the support of the measure  $\mu \in \mathbb{R}$ .

1.) Suppose  $\mu((-\infty, 0)) > 0, \mu(0, \infty) > 0$ , and that  $\text{supp}(\mu) \subset \mathbb{R}$  is a bounded subset. These assumptions ensure that  $\Lambda(\lambda) \rightarrow \infty$  when  $|\lambda| \rightarrow \infty$  and that  $\Lambda$  is finite everywhere, i.e.,  $\mathcal{D}(\Lambda) = \mathbb{R}$ . Then, according to part (c) of Lemma 1.11,  $\Lambda$  is a continuous, differentiable function, and hence there exists  $\eta \in \mathbb{R}$  such that

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

We define now a new probability measure  $\tilde{\mu} \in \mathcal{M}_1(\mathbb{R})$  by tilting the measure  $\mu$ , that is, we define the Radon-Nikodym density to be

$$\frac{d\tilde{\mu}}{d\mu}(x) = e^{\eta x - \Lambda(\eta)}, \quad (1.29)$$

and quickly check that this indeed defines a probability measure by computing writing

$$M(\eta) := e^{\Lambda(\eta)} = \mathbb{E}[e^{\eta X_1}],$$

$$\int_{\mathbb{R}} \tilde{\mu}(dx) = \frac{1}{M(\eta)} \int_{\mathbb{R}} e^{\eta x} d\mu = 1.$$

We now denote  $\tilde{\mu}_N$  the law of  $S_N$  under  $\tilde{\mu}^{\otimes N}$ , and we observe that for every  $\varepsilon > 0$  we obtain the estimate

$$\begin{aligned} \mu_N((-\varepsilon, \varepsilon)) &= \int_{\{x \in \mathbb{R}^N : |\sum_{i=1}^N x_i| < N\varepsilon\}} \mu(dx_1) \cdots \mu(dx_N) \\ &\geq e^{-N\varepsilon|\eta|} \int_{\{x \in \mathbb{R}^N : |\sum_{i=1}^N x_i| < N\varepsilon\}} \exp\left(\eta \sum_{i=1}^N x_i\right) \mu(dx_1) \cdots \mu(dx_N) \\ &= e^{-N\varepsilon|\eta|} e^{N\Lambda(\eta)} \tilde{\mu}_N((-\varepsilon, \varepsilon)). \end{aligned}$$

By (1.23) and our choice of  $\eta$ ,

$$\mathbb{E}_{\tilde{\mu}}[X_1] = \frac{1}{M(\eta)} \int_{\mathbb{R}} x e^{\eta x} d\mu = \Lambda'(\eta) = 0.$$

Thus the expectation is zero under the new measure  $\tilde{\mu}$ , and hence, by the law of large numbers,

$$\lim_{N \rightarrow \infty} \tilde{\mu}_N((-\varepsilon, \varepsilon)) = 1. \quad (1.30)$$

Our estimate above now gives, for every  $0 < \varepsilon < \delta$ ,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N((-\delta, \delta)) \geq \liminf_{n \rightarrow \infty} \frac{1}{N} \log \mu_N((-\varepsilon, \varepsilon)) \geq \Lambda(\eta) - \varepsilon|\eta|,$$

and (1.27) follows by taking the limit  $\varepsilon \rightarrow 0$  and using

$$\Lambda(\eta) \geq -\sup_{\lambda \in \mathbb{R}} \{-\Lambda(\lambda)\} = -\Lambda^*(0).$$

2.) Suppose that  $\text{supp}(\mu)$  is unbounded, while both  $\mu((-\infty, 0)) > 0$  and  $\mu((0, \infty)) > 0$ . Fix a cutoff parameter  $M > 0$  large enough so that  $\mu([-M, 0)) > 0$  as well as  $\mu((0, M]) > 0$ , and define

$$\Lambda_M(\lambda) := \log \int_{-M}^M e^{\lambda x} \mu(dx).$$

Denote  $\nu$  the law of  $X_1$  conditioned on the event  $\{|X_1| \leq M\}$ , and let  $\nu_N$  the law of  $S_N$  conditioned on  $\{|X_i| \leq M; i = 1, \dots, N\}$ . Then for every  $\delta > 0$  and for all  $N \in \mathbb{N}$ ,

$$\mu_N((-\delta, \delta)) \geq \nu((-\delta, \delta))\mu([-M, M])^N.$$

It is easy to see that (1.27) holds for  $\nu_N$ . The logarithmic moment generating function for  $\nu$  is

$$\Lambda_\nu(\lambda) = \log \left( \frac{\mathbb{E}[e^{\lambda X_1} \mathbb{I}\{|X_1| \leq M\}]}{\mu([-M, M])} \right) = \Lambda_M(\lambda) - \log \mu([-M, M]),$$

Thus

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N((-\delta, \delta)) \leq \log \mu([-M, M]) + \liminf_{N \rightarrow \infty} \frac{1}{N} \log \nu_N((-\delta, \delta)) \geq \inf_{\lambda \in \mathbb{R}} \{\Lambda_M(\lambda)\}.$$

Let  $I_M := -\inf_{\lambda \in \mathbb{R}} \{\Lambda_M(\lambda)\}$  and  $I^* = \limsup_{M \rightarrow \infty} I_M$ . Then

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N((-\delta, \delta)) \geq -I^*, \quad (1.31)$$

and we shall show that  $\inf_{\lambda \in \mathbb{R}} \{\Lambda(\lambda)\} \leq -I^*$  to conclude with (1.27). Note that  $\Lambda_M$  and thus  $-I_M$  is denote decreasing in  $M$ , and

$$-I_M \leq \Lambda_M(0) \leq \Lambda(0),$$

which shows that  $-I^* \leq 0$ . We see now that  $-I^* > -\infty$  as  $-I_M$  is finite for sufficiently large  $M$ . Thus the level sets  $\mathcal{L}_{\Lambda_M}(-I^*)$  are non-empty, compact sets and are nested with respect to  $M$ , and henceforth there is a point  $\lambda_0$  in their intersection. By Lebesgue's monotone convergence theorem,

$$\Lambda(\lambda_0) = \lim_{M \rightarrow \infty} \Lambda_M(\lambda_0) \leq -I^*,$$

and thus our bound (1.31) yields (1.27).

3.) Suppose now that either  $\mu((-\infty, 0)) = 0$  or  $\mu((0, \infty)) = 0$ , then  $\Lambda$  is a monotone function with  $\inf_{\lambda \in \mathbb{R}} \{\Lambda(\lambda)\} = \log \mu(\{0\})$ . Hence, in this case, (1.27) follows from

$$\mu_N((-\delta, \delta)) \geq \mu_N(\{0\}) = \mu(\{0\})^N.$$

□

**Remark 1.13** (a) The pivotal step in proving the large deviation upper bound is to optimise over exponential Chebychev inequalities for  $\lambda \geq 0$  considering the function  $e^{\lambda x}$ . Then consideration of the mean  $m$  and the argument  $x$  of  $\Lambda^*$  one extend the optimisation over all  $\lambda \in \mathbb{R}$  to obtain the Legendre-Fenchel transform.

(b) The crucial step in the proof of the lower bound was an exponential change of measure, sometimes also called tilting of the measure.

◇

We can strengthen our results concerning the goodness of the rate function.

**Lemma 1.14** *In the setting of Theorem 1.12 we have the following results. If  $0 \in \mathring{\mathcal{D}}_\Lambda$  then  $\Lambda^*$  is a good rate function. Moreover, if  $\mathcal{D}_\Lambda = \mathbb{R}$ , then*

$$\lim_{|x| \rightarrow \infty} \frac{\Lambda^*(x)}{|x|} = \infty. \quad (1.32)$$

**Proof.** There are  $\lambda_- < 0$  and  $\lambda_+ > 0$ ,  $\lambda_-, \lambda_+ \in \mathcal{D}_\Lambda$  since  $0 \in \mathring{\mathcal{D}}_\Lambda$ . Since for any  $\lambda \in \mathbb{R}$ ,

$$\frac{\Lambda^*(x)}{|x|} \geq \lambda \text{sign}(x) - \frac{\Lambda(\lambda)}{|x|},$$

it follows that

$$\liminf_{|x| \rightarrow \infty} \frac{\Lambda^*(x)}{|x|} \geq \min\{\lambda_+, -\lambda_-\} > 0.$$

We get  $\Lambda^*(x) \rightarrow \infty$  as  $|x| \rightarrow \infty$ , and its levels sets are closed and bounded, hence compact. Thus  $\Lambda^*$  is a good rate function. Note that (1.32) follows for  $\mathcal{D}_\Lambda = \mathbb{R}$  by considering  $-\lambda_- = \lambda_+ \rightarrow \infty$ . □

**Exercise 1.15** Prove by an application of Fatou's lemma that  $\Lambda$  is lower semicontinuous.



**Exercise 1.16** Compute  $\Lambda^*$  for the following distributions:

- (a)  $X \sim \text{Poi}(\lambda)$ , Poisson distribution with parameter  $\lambda > 0$ .
- (b)  $X \sim \text{Ber}(p)$ ,  $p \in [0, 1]$ , Bernoulli distributed with success probability  $p$ .
- (c)  $X \sim \text{Exp}(\lambda)$ , exponentially distributed with parameter  $\lambda > 0$ .
- (d)  $X \sim \text{N}(\mu, \sigma^2)$ .



**Exercise 1.17** Prove that  $\Lambda$  is  $\mathcal{C}^\infty$  in the interior  $\mathring{\mathcal{D}}_\Lambda$  and that  $\Lambda^*$  is strictly convex, and  $\mathcal{C}^\infty$  in the interior of the set  $F := \{\Lambda'(\lambda) : \lambda \in \mathring{\mathcal{D}}_\Lambda\}$



We now want to obtain the Cramér Theorem in  $\mathbb{R}^d$ . Some of the techniques for the  $\mathbb{R}$  - version are not available in  $\mathbb{R}^d$ . Suppose that  $(X_i)_{i \in \mathbb{N}}$  is a sequence of independent, identically distributed random vectors in  $\mathbb{R}^d$  with law  $\mu \in \mathcal{M}_1(\mathbb{R}^d)$ . We extend the definition of the *Legendre-Fenchel transform* in Definition 1.10 to the vector valued case in  $\mathbb{R}^d$ ,

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{ \langle x, \lambda \rangle - \Lambda(\lambda) \}, \quad x \in \mathbb{R}^d, \quad (1.33)$$

with  $\langle \cdot, \cdot \rangle$  the Euclidean inner product.

**Theorem 1.18 (Cramér Theorem in  $\mathbb{R}^d$ )** *Let  $(X_i)_{i \in \mathbb{N}}$  be a sequence of independent, identically distributed  $\mathbb{R}^d$ -valued random variables with law  $\mu \in \mathcal{M}_1(\mathbb{R}^d)$  and denote  $\mu_N$  the law of the empirical mean  $S_N$  under  $\mu^{\otimes N}$ . Assume that  $\mathcal{D}(\Lambda) = \mathbb{R}^d$ . Then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP on  $\mathbb{R}^d$  with rate  $N$  and good rate function  $\Lambda^*$ .*

## 2 Methods of types and Sanov's theorem

In this section we consider only a finite sample space  $E$  and write  $|E|$  for the number of elements of  $E$ . Before we prove the first large deviation principle we briefly discuss the role of the entropy as a measure of uncertainty. As is well-known, it was Ludwig Boltzmann who first gave a probabilistic interpretation of the *thermodynamic entropy*. He coined the formula  $S = k_B \log W$  which is engraved on his tombstone in Vienna: the entropy  $S$  of an observed state is nothing else than the logarithmic probability for its occurrence, up to some scalar factor  $k_B$  (the Boltzmann constant  $k_B = 1.3806 \times 10^{-23} \text{m}^2 \text{kgs}^{-2} \text{K}^{-1}$ ) which is physically significant but can be ignored from a mathematical point of view. The set  $E$  represents in Boltzmann's picture the possible energy levels for a system of particles, and  $\mu \in \mathcal{M}_1(E)$  corresponds to a specific histogram of energies describing some macro state of the system. Assume for a moment that each  $\mu(x), x \in E$ , is a multiple of  $\frac{1}{N}$ , i.e.,  $\mu$  is a histogram for  $N$  trials or, equivalently, a *macro state* for a system of  $N$  particles. On the microscopic level, the system is then described by a sequence  $\omega \in E^N$ , the *micro state*, associating to each particle its energy level. Boltzmann's idea is now the following:

*The entropy of a macro state  $\mu$  corresponds to the degree of uncertainty about the actual micro state  $\omega$  when only  $\mu$  is known, and can thus be measured by*

$$\log |\mathcal{T}_N(\mu)|,$$

*the logarithmic number of micro states leading to  $\mu$ .*

Recall, for a given micro state  $\omega \in E^N$ , that

$$L_N^\omega := \frac{1}{N} \sum_{i=1}^N \delta_{\omega_i}$$

is the associated macro state describing how the particles are distributed over the energy levels, and

$$\mathsf{T}_N(\nu) := \{\omega \in E^N : L_N^\omega = \nu\} \quad (2.1)$$

is the set of all  $\omega \in E^N$  of *type*  $\mu$ .

**Definition 2.1** Denote  $\mathcal{L}_N$  the set of all possible types of sequences of length in  $E$ , i.e.,

$$\mathcal{L}_N := \{\nu \in \mathcal{M}_1(E) : \nu = L_N^\omega \text{ for some } \omega \in E^N\}.$$

The *type class*  $\mathsf{T}_N(\nu)$  of  $\nu \in \mathcal{M}_1(E) \cap \mathcal{L}_N$  is the set  $\mathsf{T}_N(\nu) := \{\omega \in E^N : L_N^\omega = \nu\}$ .

Note that a type class consists of all permutations of a given vector in this set. We are using throughout the following convention,

$$0 \log 0 \triangleq 0 \quad \text{and} \quad 0 \log(0/0) \triangleq 0.$$

**Proposition 2.2 (Entropy as degree of ignorance)** Let  $\mu_N, \mu \in \mathcal{M}_1(E)$  be probability measures such that  $\mu_N \rightarrow \mu$  as  $N \rightarrow \infty$  and  $N\mu(x) \in \mathbb{N}_0$  for all  $x \in E$ . Then,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log |\mathsf{T}_N(\mu_N)| = - \sum_{x \in E} \mu(x) \log \mu(x). \quad (2.2)$$

**Proof.** This can be achieved easily with Stirling's formula and the weak convergence of the sequence of probability measures. Detailed error analysis and proof in [CK81].  $\square$

**Definition 2.3 (Shannon Entropy)** Suppose  $E$  is finite and  $\mu \in \mathcal{M}_1(E)$ . The (*Shannon*) *entropy* of  $\mu$  is defined as

$$H(\mu) := - \sum_{x \in E} \mu(x) \log \mu(x).$$

**Definition 2.4 (Relative entropy)** Suppose  $E$  is finite and  $\mu, \nu \in \mathcal{M}_1(E)$ . For  $\mu \in \mathcal{M}_1(E)$  denote

$$E_\mu := \{x \in E : \mu(x) > 0\}$$

its support. The *relative entropy* of  $\nu$  with respect to  $\mu$  is

$$H(\nu|\mu) := \begin{cases} \sum_{x \in E} \nu(x) \log \frac{\nu(x)}{\mu(x)} & \text{if } E_\nu \subset E_\mu, \\ +\infty & \text{otherwise.} \end{cases} \quad (2.3)$$

**Exercise 2.5 (Properties of relative entropy)** Show that  $H(\cdot|\mu)$  is (i) nonnegative and convex, (ii)  $H(\cdot|\mu)$  is finite on  $\{\nu \in \mathcal{M}_1(E) : E_\nu \subset E_\mu\}$ , (iii)  $H(\cdot|\mu)$  is a good rate function.



Suppose  $(X_i)_{i \in \mathbb{N}}$  is an  $E$ -valued sequence, then the *empirical measure* is the random variable

$$L_N = \frac{1}{n} \sum_{i=1}^N \delta_{X_i}$$

taking values in  $\mathcal{M}_1(E)$ . We sometimes write  $L_N^X$  for the vector  $X = (X_1, \dots, X_N)$ . As  $E$  is finite, we endow  $\mathcal{M}_1(E)$  with the metric inherited from the embedding into  $\mathbb{R}^{|E|}$  given by the mapping  $\mu \mapsto (\mu(x))_{x \in E}$ . The probability simplex

$$\text{Sim}_E := \{\nu = (\nu(x))_{x \in E} \in [0, 1]^{|E|} : \sum_{x \in E} \nu(x) = 1\} \subset \mathbb{R}^{|E|}$$

can be identified with  $\mathcal{M}_1(E)$ . We endow the simplex with the *total variation distance*

$$d(\mu, \nu) := \frac{1}{2} \sum_{x \in E} |\mu(x) - \nu(x)|, \quad (2.4)$$

which turns  $(\mathcal{M}_1(E), d)$  into a Polish space.

**Exercise 2.6** Show that, according to the SLLN, Theorem C.4,

$$d(L_N, \mu) \xrightarrow[N \rightarrow \infty]{} 0 \text{ a.s.}$$



## 2.1 The empirical measure LDP - Sanov's theorem

In this section we illustrate how combinatorial, or counting, arguments can help providing large deviation principles. For this section assume that  $E$  is a finite sample space with  $\#E = |E|$  elements. We endow  $E$  with the power set as  $\sigma$ -algebra.

**Theorem 2.7 (Sanov's theorem for finite spaces)** *Let  $(X_i)_{i \in \mathbb{N}}$  be an independent, identically distributed sequence of  $E$ -valued random variables with law  $\mu \in \mathcal{M}_1(E)$ . Denote  $\mu_n$  the distribution of  $L_n$  under  $\mu^{\otimes n}$ . Then  $(\mu_n)_{n \in \mathbb{N}}$  satisfies the LDP on  $\mathcal{M}_1(E)$  with rate  $n$  and rate function*

$$I_\mu(\nu) = H(\nu|\mu).$$

For the proof we shall need the following two lemmas.

**Lemma 2.8** *If  $x \in \mathbb{T}_N(\nu)$ ,  $\nu \in \mathcal{L}_N$ , then*

$$P((X_1, \dots, X_N) = x) = \exp(-N(H(\nu) + H(\nu|\mu))). \quad (2.5)$$

**Proof.**

$$H(\nu) + H(\nu|\mu) = - \sum_{x \in E} \nu(x) \log \mu(x).$$

Then, using independence, for  $x = (x_1, \dots, x_N) \in \mathbb{T}_N(\nu) \subset E^N$ ,

$$P((X_1, \dots, X_N) = x) = \prod_{i=1}^N \mu(x_i) = \prod_{y \in E} \mu(y)^{N\nu(y)} = \exp\left(N \sum_{y \in E} \nu(y) \log \mu(y)\right).$$

□

**Lemma 2.9** (a)  $|\mathcal{L}_N| \leq (N+1)^{|E|}$ .

(b) There exist polynomials  $p_1, p_2$  with positive coefficients such that for every  $\nu \in \mathcal{L}_N$ ,

$$\frac{1}{p_1(N)} e^{NH(\nu)} \leq |\mathbb{T}_N(\nu)| \leq p_2(N) e^{NH(\nu)}.$$

**Proof.** (a) For any  $y \in E$ , the number  $L_n^\omega(y)$  belongs to the set  $\{0, \frac{1}{N}, \dots, \frac{N-1}{N}, 1\}$  (frequency of  $y$  in  $\omega \in E^N$ ), whose cardinality is  $(N+1)$ .

(b)  $\mathbb{T}_N(\nu)$  is in bijection to the number of ways one can arrange the objects from a collection containing the object  $x \in E$  exactly  $N\nu(x)$  times. Hence  $|\mathbb{T}_N(\nu)|$  is multinomial,

$$|\mathbb{T}_N(\nu)| = \frac{N!}{\prod_{x \in E} (N\nu(x))!}.$$

Stirling's formula tell us that for suitable constants  $c_1, c_2 > 0$  we have for all  $N \in \mathbb{N}$ ,

$$N \log \frac{N}{e} \leq \log N! \leq N \log \frac{N}{e} + c_1 \log N + c_2.$$

Now,

$$\begin{aligned} \log |\mathbb{T}_N(\nu)| &\leq \log N! - \sum_{x \in E} \log (N\nu(x))! \leq N \log \frac{N}{e} - \sum_{x \in E} N\nu(x) \log \frac{N\nu(x)}{e} + c_1 \log N + c_2 \\ &= nH(\nu) + c_1 \log N + c_2, \end{aligned}$$

which yields the desired upper bound with  $p_2(N) = c_2 N^{c_1}$ . The proof of the lower bound is analogous. □

**Proof of Theorem 2.7.** Pick a Borel set  $A \subset \mathcal{M}_1(E)$ . Then, using the upper bound in Lemma 2.9,

$$\begin{aligned} P(L_N \in A) &= \sum_{\nu \in \mathcal{L}_N \cap A} P(L_N = \nu) = \sum_{\nu \in \mathcal{L}_N \cap A} \sum_{x \in \mathbb{T}_N(\nu)} P(X = (X_1, \dots, X_N) = x) \\ &\leq \sum_{\nu \in \mathcal{L}_N \cap A} p_2(N) e^{NH(\nu)} e^{-N(H(\nu) + H(\nu|\mu))} \\ &\leq (N+1)^{|E|} p_2(N) e^{-N \inf_{\nu \in A \cap \mathcal{L}_N} H(\nu|\mu)}. \end{aligned}$$

The lower bound reads

$$\begin{aligned} P(L_N \in A) &= \sum_{\nu \in \mathcal{L}_N \cap A} P(L_N = \nu) \geq \sum_{\nu \in \mathcal{L}_N \cap A} \frac{1}{p_1(N)} e^{NH(\nu|\mu)} \\ &\geq \frac{1}{p_1(N)} e^{-N \inf_{\nu \in A \cap \mathcal{L}_N} H(\nu|\mu)}. \end{aligned}$$

Since

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log(N+1)^{|E|} = \lim_{n \rightarrow \infty} \frac{1}{N} \log p_2(N) = \lim_{N \rightarrow \infty} \frac{1}{N} \log \frac{1}{p_1(N)} = 0,$$

we obtain

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log P(L_N \in A) &= - \liminf_{N \rightarrow \infty} \left\{ \inf_{\nu \in A \cap \mathcal{L}_N} H(\nu|\mu) \right\} \\ \liminf_{N \rightarrow \infty} \frac{1}{N} \log P(L_N \in A) &= - \limsup_{N \rightarrow \infty} \left\{ \inf_{\nu \in A \cap \mathcal{L}_N} H(\nu|\mu) \right\}. \end{aligned}$$

The desired upper bound of the large deviation principle in Theorem 2.7 follows, since  $A \cap \mathcal{L}_N \subset A$  for all  $N$ .

For the large deviation lower bound we pick  $\nu \in \overset{\circ}{A}$  from the interior of  $A$  such that  $E_\nu \subset E_\mu$ . We then find  $\delta > 0$  small enough such that the ball

$$\{\nu' \in \mathcal{M}_1(E) : d(\nu', \nu) < \delta\}$$

is contained in  $A$ . Observe that  $\mathcal{L}_N$  contains all probability measures taking values in  $\{0, \frac{1}{N}, \dots, 1\}$ . Thus, for each  $\nu \in \mathcal{M}_1(E)$  there is a  $\nu' \in \mathcal{L}_N$  such that for all  $x \in E$ :  $|\nu(x) - \nu'(x)| \leq C/N$  for some  $C > 0$ . Thus there exist a sequence  $\nu_N \in A \cap \mathcal{L}_N$  such that  $\nu_N \rightarrow \nu$  as  $N \rightarrow \infty$ . Moreover, without loss of generality, we may assume that  $E_{\nu_N} \subset E_\mu$ , and hence

$$- \limsup_{N \rightarrow \infty} \left\{ \inf_{\nu' \in A \cap \mathcal{L}_N} H(\nu'|\mu) \right\} \geq - \lim_{n \rightarrow \infty} H(\nu_N|\mu) = -H(\nu|\mu).$$

Recall that  $H(\nu|\mu) = \infty$  whenever, for some  $x \in E$ ,  $\nu(x) > 0$  while  $\mu(x) = 0$ . Therefore, by the preceding inequality, optimising over  $\nu \in \overset{\circ}{A}$ ,

$$- \limsup_{N \rightarrow \infty} \left\{ \inf_{\nu' \in A \cap \mathcal{L}_N} H(\nu'|\mu) \right\} \geq - \inf_{\nu \in \overset{\circ}{A}} H(\nu|\mu).$$

□

**Exercise 2.10** Prove that for every open set  $A \subset \mathcal{M}_1(E)$ ,

$$- \lim_{N \rightarrow \infty} \left\{ \inf_{\nu \in A \cap \mathcal{L}_N} H(\nu|\mu) \right\} = \lim_{N \rightarrow \infty} \frac{1}{N} \log P(L_N \in A) = - \inf_{\nu \in A} H(\nu|\mu).$$

☕

## 2.2 The pair empirical measure

We now study a generalisation of the empirical measure. For this we are recording two successive values at each instant of time of outcomes  $X_1, X_2, \dots$ . We assume that our random variables  $X_i$  are  $E$ -valued with  $E$  being a finite sample space. This generalisation will be useful when we will drop the assumption that the sequence  $(X_i)_{i \in \mathbb{N}}$  is i.i.d. and consider instead Markov sequences.

**Definition 2.11** Suppose  $(X_i)_{i \in \mathbb{N}}$  i.i.d. with  $X_i \in E$  and write  $X = (X_1, \dots, X_N)$ ,  $N \in \mathbb{N}$ . The *pair empirical measure* is the random probability measure on  $E \times E$ , defined as

$$L_N^{2,X} \equiv L_N^2 = \frac{1}{N} \sum_{i=1}^N \delta_{(X_i, X_{i+1})} \quad (2.6)$$

with the convention that  $X_{N+1} = X_1$  (periodic boundary conditions). We write  $\nu = (\nu_{x,y})_{x,y \in E}$  for  $\nu \in \mathcal{M}_1(E \times E)$ . Denote

$$\widetilde{\mathcal{M}}_1(E \times E) := \left\{ \nu \in \mathcal{M}_1(E \times E) : \nu^{(1)}(\cdot) = \sum_{y \in E} \nu_{\cdot,y} = \sum_{y \in E} \nu_{y,\cdot} = \nu^{(2)}(\cdot) \right\} \quad (2.7)$$

the set of probability measures on  $E \times E$  with equals marginals.

We turn  $\widetilde{\mathcal{M}}_1(E \times E)$  into a Polish space with the total variation metric/distance

$$d(\mu, \nu) = \frac{1}{2} \sum_{x,y \in E} |\mu_{x,y} - \nu_{x,y}|, \quad \mu, \nu \in \mathcal{M}_1(E \times E).$$

It follows from Birkhoff's Ergodic Theorem that

$$d(L_N^2, \mu \otimes \mu) \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \mathbb{P} - \text{a.s.},$$

where  $\mu \otimes \mu$  is the product measure of  $\mu \in \mathcal{M}_1(E)$ , the law of the i.i.d. sequence  $(X_i)_{i \in \mathbb{N}}$ .

**Theorem 2.12 (LDP: Pair empirical measure)** Let  $(X_i)_{i \in \mathbb{N}}$  be a sequence i.i.d.  $E$ -valued random variables with law  $\mu \in \mathcal{M}_1(E)$ . Under periodic boundary conditions,  $X_{N+1} = X_1$ ,  $N \in \mathbb{N}$ , for the pair empirical measures the following holds for all  $\varepsilon > 0$ ,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(L_N^2 \in B_\varepsilon^c(\mu \otimes \mu)) = - \inf_{\nu \in B_\varepsilon^c(\mu \otimes \mu)} \{I_\mu^2(\nu)\}, \quad (2.8)$$

where

$$B_\varepsilon^c(\mu \otimes \mu) = \{\nu \in \widetilde{\mathcal{M}}_1(E \times E) : d(\nu, \mu \otimes \mu) \leq \varepsilon\}$$

is the closed ball around  $\mu \otimes \mu$  with radius  $\varepsilon$  and

$$I_\mu^2(\nu) = \sum_{x,y \in E} \nu_{x,y} \log \frac{\nu_{x,y}}{\nu_x^{(1)} \mu_y}, \quad (2.9)$$

where  $\nu^{(1)} \in \mathcal{M}_1(E)$  is the first marginal of  $\nu \in \mathcal{M}_1(E \times E)$ .

**Remark 2.13** (a) Extend the definition of relative in Definition 2.4 to the sample space  $E \times E$  check that

$$I_\mu^2(\nu) = H(\nu | \nu^{(1)} \otimes \mu), \quad \nu \in \widetilde{\mathcal{M}}_1(E \times E). \quad (2.10)$$

(b) Comparing  $I_\mu^2$  with the rate function  $I_\mu, u$  in Sanov's theorem, Theorem 2.7, we realise that  $\nu_x^{(1)} \mu_y$  appears in the denominator instead of  $\mu_x \mu_y$  as we would expect from a direct extension of Sanov's theorem to  $E \times E$ . This discrepancy comes from the fact that in Theorem 2.12 we are recording pairs  $(X_1, X_1), (X_2, X_3), (X_3, X_4), \dots$  rather than the pairs  $(X_1, X_2), (X_3, X_4), (X_5, X_6), \dots$ . We see that the pairs in Theorem 2.12 are interlocked.

(c) Define for any  $x \in E$  the (conditional) probability measure  $\bar{\nu}_x$  by

$$\bar{\nu}_x(y) := \frac{\nu_{x,y}}{\nu_x^{(1)}}, y \in E. \quad (2.11)$$

Then

$$I_\mu^2(\nu) = I_\mu(\nu^{(1)}) + H(\nu | \nu^{(1)} \otimes \nu^{(1)}) = \sum_{x \in E} \nu_x^{(1)} H(\bar{\nu}_x | \mu), \quad \nu \in \widetilde{\mathcal{M}}_1(E \times E). \quad (2.12)$$

◇

**Proof of Theorem 2.12.** The proof is very similar the one of Theorem 2.7 though the combinatorics is more involved here. Denote

$$F_N := \{\mathbf{f} = (f_{x,y})_{x,y \in E} \in \mathbb{N}_0^{E \times E} : \sum_{x,y \in E} f_{x,y} = N, \sum_{y \in E} f_{x,y} = \sum_{y \in E} f_{y,x} = \bar{f}_x, x \in E\} \quad (2.13)$$

the set of possible frequencies of pairs from  $N$  samples. Clearly,

$$\frac{1}{N} F_N = \left\{ \frac{1}{N} \mathbf{f} : \mathbf{f} \in F_N \right\} \subset \widetilde{\mathcal{M}}_1(E \times E).$$

We write

$$\bar{f}_x = \sum_{y \in E} f_{x,y}, \quad x \in E.$$

For a given  $\mathbf{f} \in F_N$  we know the probability

$$\mathbb{P}\left(L_N^2(x, y) = \frac{f_{x,y}}{N} \forall x, y \in E\right) = \text{Comb} \prod_{x \in E} \mu_{\bar{f}_x}^{\bar{f}_x},$$

where Comb is a combinatorial factor accounting for all possible arrangements of the sample  $X_1, \dots, X_N$  that give rise to the frequency matrix  $\mathbf{f}$ . We mark each occurrence of a pair  $(x, y)$  of states  $x, y \in Y$  in the sample  $X_1, \dots, X_N$  by drawing an arrow from  $x$  to  $y$ . We obtain an oriented graph  $G(\mathbf{f})$  with vertex set  $E$  and the arrows as its set of oriented edges. We impose periodic boundary conditions for our sample,  $X_1 = X_{N+1}$ , and thus for every  $x \in E$  we have that

$$\#\{\text{ingoing arrows to } x\} = \#\{\text{outgoing arrows from } x\}.$$

The total number of arrows is exactly  $N$ . Thus

$$\text{Comb} = \#(\text{SG}(\mathbf{f})) \frac{\mathcal{E}(G(\mathbf{f}))}{\prod_{x,y \in E} f_{x,y}!},$$

where  $\mathcal{E}(G(f))$  denotes the number of Euler circuits on  $G(f)$ , that is, the number of looped paths respecting the arrows and using each arrow of the graph precisely once. The fact compensates for distinguishing between different arrows from  $x$  to  $y$ , and  $\#(SG(f))$  counts the number of cyclic shifts of the sample  $X_1, \dots, X_N$  that are distinct. We immediately see that  $1 \leq \#(SG(f)) \leq N$ . For an estimate of the number of Euler circuits, see Lemma 2.14 below. With Lemma 2.14 we thus get

$$\mathbb{P}\left(L_N^2(x, y) = \frac{f_{x,y}}{N} \forall x, y \in E\right) = e^{O(\log N)} \frac{\prod_{x \in E} \bar{f}_x!}{\prod_{x,y \in E} f_{x,y}!} \prod_{x \in E} \mu_x^{\bar{f}_x}$$

uniformly for  $f \in F_N$ . Then, as before,

$$Q_N(\varepsilon) \leq \mathbb{P}\left(L_N^2 \in \mathbb{B}_\varepsilon^c(\mu \otimes \mu)\right) \leq |F_N| Q_N(\varepsilon)$$

with

$$Q_N(\varepsilon) = \max_{f \in F_N: \frac{1}{N}f \in B_\varepsilon^c(\mu \otimes \mu)} \left\{ \mathbb{P}\left(L_N^2(x, y) = \frac{f_{x,y}}{N} \forall x, y \in E\right) \right\}.$$

We observe that  $|F_N| = O(N^{|E|-1}) = e^{O(\log N)}$ , and with Stirling's formula,

$$\frac{1}{N} \log \mathbb{P}\left(L_N^2 \in \mathbb{B}_\varepsilon^c(\mu \otimes \mu)\right) = O\left(\frac{\log N}{N}\right) - \min_{f \in F_N: \frac{1}{N}f \in B_\varepsilon^c(\mu \otimes \mu)} \left\{ I_\mu^2(1/N f) \right\}.$$

We conclude now in the same way as in our proof of Sanov's theorem, Theorem 2.7.  $\square$

**Lemma 2.14 (Euler Circuits)** *In the setting of Theorem 2.12 and its proof the following holds,*

$$\prod_{x \in E: \bar{f}_x > 0} (\bar{f}_x - 1)! \leq \mathcal{E}(G(f)) \leq \prod_{x \in E: \bar{f}_x > 0} \bar{f}_x!.$$

**Proof.** Upper bound: Suppose we build an Euler circuit by picking the arrows as we go along. Clearly, we cannot make more than  $\bar{f}_x$  different choices where to go from vertex  $x$ . This gives our bound. Lower bound: Pick any Euler circuit  $\mathcal{C}$ , and for each vertex assign the colour red to the outgoing arrow that is used last in the circuit  $\mathcal{C}$ . If we permute after that procedure the non-coloured arrows, of which vertex  $x$  has  $\bar{f}_x - 1$ , then we again get an Euler circuit  $\mathcal{C}'$ . All Euler circuits obtained by such permutations are distinct, i.e.,  $\mathcal{C} \neq \mathcal{C}'$ .  $\square$

### 2.3 Cramer's theorem for finite subsets in $\mathbb{R}$

We now compare Cramér's Theorem for finite sets  $E$  with Sanov's Theorem, Theorem 2.7 for finite sets  $E$ . Suppose that  $(Y_i)_{i \in \mathbb{N}}$  is a sequence of independent, identically distributed  $E$ -valued random variables with law  $\mu \in \mathcal{M}_1(E)$  having support  $E_\mu = E$ .

We shall study the *empirical mean*  $S_N = \frac{1}{N} \sum_{i=1}^N X_i$ , where  $X_i = f(Y_i)$  for some function  $f: E \rightarrow \mathbb{R}$ . Without loss of generality, we assume further that  $E_\mu = E$  and that  $f(a_1) < f(a_2) < \dots < f(a_{|E|})$ . Then  $S_N \in [f(a_1), f(a_{|E|})] =: K$ , and writing  $Y = (Y_1, \dots, Y_N)$  and  $F := (f(a_1), \dots, f(a_{|E|})) \in \mathbb{R}^{|E|}$ , we see that

$$S_N = \sum_{i=1}^{|E|} f(a_i) L_N^Y(a_i) =: \langle f, L_N^Y \rangle,$$

where  $\langle f, \nu \rangle = \sum_{x \in E} f(x) \nu(x)$  is the expectation of  $f$  with respect to  $\nu \in \mathcal{M}_1(E)$ . Thus for every set  $A \subset \mathbb{R}$  and every  $n \in \mathbb{N}$ ,

$$S_N \in A \iff L_N^Y \in \{\nu \in \mathcal{M}_1(E) : \langle f, \nu \rangle \in A\} =: \Gamma. \quad (2.14)$$

**Theorem 2.15 (Cramér's theorem for subsets of  $\mathbb{R}$ )** For any  $A \subset \mathbb{R}$ ,

$$\begin{aligned} -\inf_{x \in A} \{I(x)\} &\leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \in A) \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(S_N \in A) \leq -\inf_{x \in A} \{I(x)\}, \end{aligned}$$

where

$$I(x) = \inf_{\nu \in \mathcal{M}_1(E) : \langle f, \nu \rangle = x} \{H(\nu | \mu)\}.$$

The rate function  $I$  is continuous on the compact set  $K$  and satisfies on  $K$ ,

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}, \quad (2.15)$$

where

$$\Lambda(\lambda) = \log \sum_{i=1}^{|E|} e^{\lambda f(a_i)} \mu(a_i).$$

**Proof.** Suppose that  $f: E \rightarrow \mathbb{R}$  is constant, i.e.,  $f(x) = c \in \mathbb{R}$  for all  $x \in E$ . Then  $X_i = c, S_N = c$ , and hence  $\Gamma = \mathcal{M}_1(E)$  in (2.14). Note that when  $x \neq c$  there is no  $\nu \in \mathcal{M}_1(E)$  with  $\langle f, \nu \rangle = x$ , and thus the infimum in the definition of  $I$  is over an empty set and therefore infinity. Hence,

$$I(x) = \inf_{\nu : \langle f, \nu \rangle = x} \{H(\nu | \mu)\} = \begin{cases} 0 & \text{if } x = c, \\ +\infty & \text{if } x \neq c. \end{cases}$$

The logarithmic moment generating function for  $\hat{S}_n$  is

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{N\lambda S_N}] = \Lambda(\lambda) = \log e^{\lambda c} = \lambda c,$$

and thus

$$\sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\} = \begin{cases} 0 & \text{if } x = c, \\ +\infty & \text{if } x \neq c. \end{cases}$$

Suppose now that  $f$  is not constant. As  $\nu \mapsto \langle f, \nu \rangle$  is continuous, we know that when  $A \subset \mathbb{R}$  is open then so is  $\Gamma \subset \mathcal{M}_1(E)$  defined in (2.14). Then the lower and upper bounds follow from Sanov's theorem, Theorem 2.7. Furthermore, due to (2.14),

$$\inf_{\nu \in \mathring{\Gamma}} \{H(\nu|\mu)\} = \inf_{x \in \mathring{A}} \left\{ \inf_{\nu: \langle f, \nu \rangle = x} \{H(\nu|\mu)\} \right\}.$$

Jensen's inequality yields

$$\Lambda(\lambda) = \log \sum_{x \in E} \mu(x) e^{\lambda f(x)} \geq \sum_{x \in E \cap E_\nu} \nu(x) \log \left( \frac{\mu(x) e^{\lambda f(x)}}{\nu(x)} \right) = \lambda \langle f, \nu \rangle - H(\nu|\mu),$$

with equality for  $\nu_\lambda \in \mathcal{M}_1(E)$  defined as

$$\nu_\lambda(x) = \mu(x) e^{\lambda f(x) - \Lambda(\lambda)}, \quad x \in E.$$

Thus

$$\lambda x - \Lambda(\lambda) \leq \inf_{\nu: \langle f, \nu \rangle = x} \{H(\nu|\mu)\} = I(x)$$

with equality when  $x = \langle f, \nu_\lambda \rangle$ . The function  $\Lambda$  is differentiable with

$$\Lambda'(\lambda) = \langle f, \nu_\lambda \rangle = \mathbb{E}_{\nu_\lambda}[f],$$

and therefore (2.15) holds for all  $x \in \{\Lambda'(\lambda): \lambda \in \mathbb{R}\}$ . An easy computation shows that

$$\Lambda''(\lambda) = \mathbb{E}_{\nu_\lambda}[f^2] - (\mathbb{E}_{\nu_\lambda}[f])^2 = \text{Var}_{\nu_\lambda}(f) > 0$$

as  $f$  is not a constant. Thus  $\Lambda''(\lambda) > 0$  for all  $\lambda \in \mathbb{R}$ ,  $\Lambda$  strictly convex and  $\Lambda'$  strictly increasing. Moreover,

$$f(a_1) = \inf_{\lambda \in \mathbb{R}} \{\Lambda'(\lambda)\} \quad \text{and} \quad f(a_{|E|}) = \sup_{\lambda \in \mathbb{R}} \{\Lambda'(\lambda)\}.$$

Hence, (2.15) holds for all  $x \in \mathring{K}$ . Consider the left endpoint  $x = f(a_1)$  of the compact interval  $K$ , and let  $\nu^*(a_1) = 1$  yielding  $\langle f, \nu^* \rangle = x$ . Then

$$-\log \mu(a_1) = H(\nu^*|\mu) \geq I(x) \geq \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\} \geq \lim_{\lambda \rightarrow -\infty} (\lambda x - \Lambda(\lambda)) = -\log \mu(a_1).$$

The proof for the right endpoint of  $K$  is similar. The continuity of  $I$  follows from the continuity of the relative entropy. □

### 3 General Theory

#### 3.1 Basic theory

In the following we assume that  $(E, d)$  is a Polish space. By default we denote  $\mathcal{B}$  the Borel- $\sigma$ -algebra of  $E$ .

**Definition 3.1 (Weak Large deviation principle)** Suppose that all compact subsets of  $E$  belong to  $\mathcal{B}$ . A sequence  $(\mu_N)_{N \in \mathbb{N}}$  of probability measures is said to satisfy the *weak large deviation principle* if the upper bound in (1.17) holds for every  $\alpha$  and all compact subsets of  $\mathcal{L}_I(\alpha)^c$ , and the lower bound (1.18) holds for all measurable subsets.

**Definition 3.2 (Exponential tightness)** Suppose that all compact subsets of  $E$  belong to the  $\sigma$ -algebra  $\mathcal{B}$ . A sequence  $(\mu_N)_{N \in \mathbb{N}}$  of probability measures  $\mu_N \in \mathcal{M}_1(E, \mathcal{B})$ , is *exponentially tight* if for every  $\alpha < \infty$ , there exists a compact set  $K_\alpha \subset E$  such that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K_\alpha^c) < -\alpha.$$

We now show that one can lift a weak LDP to a standard LDP for exponentially tight sequences.

**Proposition 3.3 (Exponential tightness)** Let  $(\mu_N)_{N \in \mathbb{N}}$  be exponentially tight.

- (a) If the upper bound (1.17) holds for some  $\alpha < \infty$  and all compact subsets of the complement  $\mathcal{L}_I(\alpha)^c$ , then it holds for all measurable sets  $M$  with  $\overline{M} \subset \mathcal{L}_I(\alpha)^c$ . If  $\mathcal{B}(E) \subset \mathcal{B}$  and the upper bound (1.17) holds for all compact sets, then it also holds for all closed sets.
- (b) If the lower bound (1.18) holds (the lower bound in (1.14) when  $\mathcal{B}(E) \subset \mathcal{B}$ ) for all measurable sets (all open sets), then  $I$  is a good rate function.

**Proof.** (a) Pick  $M \in \mathcal{B}$  and  $\alpha < \infty$  such that  $\overline{M} \subset \mathcal{L}_I(\alpha)^c$ , and let  $K_\alpha$  be the compact set in the definition for exponential tightness. Then  $\overline{M} \cap K_\alpha \in \mathcal{B}$  and  $K_\alpha^c \in \mathcal{B}$ .

$$\mu_N(M) \leq \mu_N(\overline{M} \cap K_\alpha) + \mu_N(K_\alpha^c). \quad (3.1)$$

As  $\overline{M} \cap K_\alpha \subset \mathcal{L}_I(\alpha)^c$  we have that

$$\inf_{x \in \overline{M} \cap K_\alpha} \{I(x)\} \geq \alpha.$$

Thus

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \text{R.H.S. of (3.1)} = \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(\overline{M} \cap K_\alpha) \wedge \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K_\alpha^c),$$

and therefore

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(M) \leq -\alpha.$$

(b) We apply the lower bound (1.18) to the open set  $K_\alpha^c$ , and obtain

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K_\alpha^c) \geq - \inf_{x \in K_\alpha^c} \{I(x)\},$$

and thus (noting that  $K_\alpha$  is the compact set from the definition of exponential tightness)  $\inf_{x \in K_\alpha^c} \{I(x)\} > \alpha$ . Therefore,

$$\mathcal{L}_I(\alpha) \subset K_\alpha$$

showing that the level set  $\mathcal{L}_I(\alpha)$  is compact. Hence, the rate function  $I$  is good rate function.  $\square$

**Proposition 3.4 (Rate functions attains infimum over compact sets)** *Suppose that  $I: E \rightarrow [0, \infty]$  is a rate function. Then  $I$  attains its infimum over compact sets, i.e., for all  $K \subset E$  compact there exists  $y \in K$  such that*

$$I(y) = \inf_{x \in K} \{I(x)\}.$$

**Proof.** Suppose  $I$  has no minimum over the compact  $K \subset E$ , and define  $\alpha := \inf_{x \in K} \{I(x)\}$ . Then, for each  $x \in K$ , we have that  $\alpha < I(x)$  and there is  $\varepsilon = \varepsilon(x) > 0$  such that

$$\alpha < I(x) - \varepsilon.$$

As  $I$  is lower semicontinuous, there is an open neighbourhood  $\mathcal{U}(x)$  of  $x$  such that

$$I(x) - \varepsilon < I(y) \quad \text{for all } y \in \mathcal{U}(x).$$

As  $K$  is compact we can extract a finite cover of the set, that is, there are  $x_1, \dots, x_M \in K$ ,  $M \in \mathbb{N}$ , such that

$$K \subset \bigcup_{i=1}^M \mathcal{U}(x_i).$$

Define  $\beta := \min_{1 \leq i \leq M} \{I(x_i) - \varepsilon(x_i)\}$ . Then  $\beta > \alpha$  and  $\beta \leq I(x_k) - \varepsilon(x_k) < I(y)$  for all  $k = 1, \dots, M$ , and for all  $y \in K$ . We thus obtain a contradiction for  $y \in K$  with  $y \in \mathcal{U}(x_k)$  as then  $\beta \leq \inf_{x \in K} \{I(x)\}$ , and our statement follows.  $\square$

We show in the next lemma that the LDP is preserved under suitable inclusions. Hence, in applications, one may first prove an LDP in a space that possesses additional structure (for example, a topological vector space), and then use this lemma to deduce the LDP in the subspace of interest.

**Lemma 3.5 (LDPs for Inclusions)** *Let  $(\mu_N)_{N \in \mathbb{N}}$  be a sequence of probability measures  $\mu_N \in \mathcal{M}_1(E)$ . Suppose that  $\mathcal{E} \subset E$  is a measurable subset with  $\mu_N(\mathcal{E}) = 1$  for all  $N \in \mathbb{N}$ . We equip  $\mathcal{E}$  with the topology induced by  $E$ .*

- (a) *If  $\mathcal{E} \subset E$  is a closed set and if  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP in  $\mathcal{E}$  with rate function  $I$ , then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP in  $E$  with rate function  $I'$  such that  $I' \equiv I$  on  $\mathcal{E}$  and  $I' \equiv +\infty$  on  $\mathcal{E}^c = E \setminus \mathcal{E}$ .*
- (b) *If  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP in  $E$  with rate function  $I$  and  $\mathcal{D}_I \subset \mathcal{E}$ , then the same LDP holds in  $\mathcal{E}$ . If  $\mathcal{E}$  is closed we have that  $\mathcal{D}_I \subset \mathcal{E}$  and the same LDP holds in  $\mathcal{E}$ .*

**Proof.** Note that  $G \cap \mathcal{E}$  are open sets in  $\mathcal{E}$  for every  $G \subset E$  open, likewise,  $F \cap \mathcal{E}$  are closed for all  $F \subset E$  closed. From our assumptions we have that  $\mu_N(\Gamma) = \mu_N(\Gamma \cap \mathcal{E})$  for any measurable set  $\Gamma \subset E$ .

(a) Consider  $\mathcal{E} \subset E$  closed and extend the rate function  $I': \mathcal{E} \rightarrow [0, \infty]$ ,  $I' \equiv I$  on  $\mathcal{E}$ , to  $E$  by setting  $I'(x) = +\infty$  for any  $x \in \mathcal{E}^c$ . Then, for every measurable set  $\Gamma \subset E$ ,

$$\inf_{x \in \Gamma} \{I'(x)\} = \inf_{x \in \Gamma \cap \mathcal{E}} \{I(x)\}.$$

Thus we obtain the large deviation lower and bounds directly for the existing ones with rate function  $I$ .

(b) Suppose the LDP holds in  $E$ . If  $\mathcal{E} \subset E$  is closed, then  $\mathcal{D}_I \subset \mathcal{E}$  by the LDP lower bound for the open set  $\mathcal{E}^c$ ,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(\mathcal{E}^c) \geq - \inf_{x \in \mathcal{E}^c} \{I(x)\}.$$

This implies that

$$\inf_{x \in \Gamma} \{I(x)\} = \inf_{x \in \Gamma \cap \mathcal{E}} \{I(x)\} \quad (3.2)$$

holds for any measurable set  $\Gamma \subset E$ , and henceforth the LDP lower and upper bound follow from the right hand side in (3.2).  $\square$

### 3.2 Contraction principle

This section is on transformations that preserve the LDP, although possibly, changing the rate function.

**Theorem 3.6 (Contraction Principle)** *Let  $(E, d)$  and  $(Y, d_Y)$  be metric spaces and  $T: E \rightarrow Y$  a continuous function. Suppose that  $I: E \rightarrow [0, \infty]$  is a good rate function and define*

$$J(y) := \inf_{x \in E: y = I(x)} \{I(x)\}, \quad y \in Y. \quad (3.3)$$

(a) *Then  $J$  is a good rate function on  $Y$ , where the infimum in (3.3) over the empty set is taken as  $\infty$ .*

(b) *If  $I$  is the rate function for a large deviation principle (LDP) associated with a sequence  $(\mu_N)_{N \in \mathbb{N}}$  of probability measures  $\mu_N \in \mathcal{M}_1(E)$  on  $E$ , then  $J$  controls the LDP on  $Y$  for the sequence  $(\mu_N \circ T^{-1})_{N \in \mathbb{N}}$  of probability measures  $\mu_N \circ T^{-1} \in \mathcal{M}_1(Y)$ .*

**Proof.** (a) By definition we have  $J \geq 0$ . For each point  $y$  in the range  $T(E) \subset Y$  the infimum on the right hand side of (3.3) is attained at some point  $x \in E$ . This follows from the goodness of the rate function  $I$  as for  $y \in T(E)$  the set  $\{x \in E: I(x) = y\}$  is compact and any lower semicontinuous function attains its infimum over a compact set. Thus we obtain for the level sets of  $J$ ,

$$\mathcal{L}_J(\alpha) = \{T(x): I(x) \leq \alpha\} = T(\mathcal{L}_I(\alpha)),$$

where  $\mathcal{L}_I(\alpha)$  are the level sets for  $I$ . As  $\mathcal{L}_I \subset E$  are compact due to the goodness of  $I$ , so are the sets  $\mathcal{L}_J \subset Y$ , and thus  $J$  is a good rate function.

(b) The definition of  $J$  in (3.3) implies that for any  $A \subset Y$ ,

$$\inf_{y \in A} \{J(y)\} = \inf_{x \in T^{-1}(A)} \{I(x)\}. \quad (3.4)$$

Since  $T$  is continuous, the set  $T^{-1}(A)$  is open (closed) subset of  $E$  for any open (closed)  $A \subset Y$ . Therefore, the LDP for  $\mu_N \circ T^{-1}$  follows as a consequence of the LDP for  $\mu_N$  and (3.4). Indeed, pick  $F \subset Y$  closed and write

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N \circ T^{-1}(F) &= \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(T^{-1}(F)) \leq - \inf_{x \in T^{-1}(F)} \{I(x)\} \\ &= - \inf_{y \in F} \inf_{x \in T^{-1}(\{y\})} \{I(x)\} = - \inf_{y \in F} \{J(y)\}. \end{aligned}$$

A similar argument works for  $O \subset Y$  open.  $\square$

The following theorem shows that in the presence of exponential tightness, the contraction principle can be made to work in the reverse direction. This property is extremely useful for strengthening large deviations results from a coarse topology to a finer one.

**Theorem 3.7 (Inverse Contraction Principle)** *Let  $E$  and  $Y$  be Polish spaces. Suppose that  $\psi: Y \rightarrow E$  is a continuous bijection, and that  $(\nu_N)_{N \in \mathbb{N}}$  is an exponentially tight sequence of probability measures  $\nu_N \in \mathcal{M}_1(Y)$ . If  $(\nu_N \circ \psi^{-1})_{N \in \mathbb{N}}$  satisfies the LDP with rate function  $I: E \rightarrow [0, \infty]$ , then  $(\nu_N)_{N \in \mathbb{N}}$  satisfies the LDP with the good rate function  $I' := I \circ \psi$ .*

**Proof.** We first show that  $I'$  is a rate function. By the continuity of  $\psi$ , for any  $\alpha < \infty$ , we see that the level set

$$\mathcal{L}_{I'}(\alpha) = \{y \in Y: I'(y) \leq \alpha\} = \psi^{-1}(\mathcal{L}_I(\alpha))$$

is closed, and thus  $I'$  is lower semicontinuous. Moreover,  $I' \geq 0$ , and hence  $I'$  is a rate function. The exponential tightness allows to prove the LDP upper bound for compact sets  $K \subset Y$ . Hence

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(K) = \limsup_{N \rightarrow \infty} \frac{1}{N} \log \nu_N \circ \psi^{-1}(\psi(K)) \leq - \inf_{x \in \psi(K)} \{I(x)\} = - \inf_{y \in K} \{I'(y)\},$$

which is the upper bound for  $\nu_N$ . We turn to the lower bound which is slightly more involved. Fix  $y \in Y$  with  $I'(y) = I(\psi(y)) = \alpha < \infty$ , and a neighbourhood  $G \ni y$  of  $y$ . For  $\alpha < \infty$ , there exists a compact set  $K_\alpha \subset Y$  such that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(K_\alpha^c) < -\alpha. \quad (3.5)$$

Because  $\psi$  is a bijection,  $K_\alpha^c = \psi^{-1} \circ \psi(K_\alpha^c)$  and  $\psi(K_\alpha^c) = \psi(K_\alpha)^c$ . By the continuity of  $\psi$ , the set  $\psi(K_\alpha)$  is compact, and consequently  $\psi(K_\alpha)^c$  is an open set. We have therefore the large deviation lower bound for  $\nu_N \circ \psi^{-1}$ ,

$$- \inf_{x \in \psi(K_\alpha^c)} \{I(x)\} \leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(K_\alpha^c) < -\alpha.$$

From  $I(\psi(y)) = \alpha$  we know that  $y \in K_\alpha$ . Since  $\psi$  is continuous bijection, it is a homeomorphism between the compact sets  $K_\alpha$  and  $\psi(K_\alpha)$ . Therefore, the set  $\psi(G \cap K_\alpha)$  is a neighbourhood of  $\psi(y)$ . Hence, there exists a neighbourhood  $G'$  of  $g(y)$  in  $E$  such that

$$G' \subset \psi(G \cap K_\alpha) \cup \psi(K_\alpha^c).$$

This implies, for every  $N$ ,

$$\nu_N(G) + \nu_N(K_\alpha^c) \geq \nu_N \circ \psi^{-1}(G'),$$

and thus

$$\begin{aligned} \max\left\{\liminf_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(G), \limsup_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(K_\alpha^c)\right\} &\geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \nu_N \circ \psi^{-1}(G') \\ &\geq -I(\psi(y)) = -I'(y). \end{aligned}$$

Since  $I'(y) = \alpha$ , it follows by combining this inequality with (3.5) that

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \nu_N(G) \geq -I'(y).$$

We are done as the preceding holds for every  $y \in Y$  and every neighbourhood  $G$  of  $y$ .  $\square$

The next result is a direct consequence which holds for general topological Hausdorff spaces  $E$  and concerns the comparison of topologies in terms of LDPs.

**Proposition 3.8 (Different topologies)** *Let  $(\mu_N)_{N \in \mathbb{N}}$  be an exponentially tight sequence of probability measures on  $E$  ( $E$  some topological Hausdorff space) equipped with the topology  $\tau_1$ . If  $(\mu_N)_{N \in \mathbb{N}}$  satisfies an LDP with respect to a Hausdorff topology  $\tau_2$  on  $E$  that is coarser than  $\tau_1$ , that is,  $\tau_2 \subset \tau_1$  (respectively,  $\tau_1$  is finer than  $\tau_2$  when  $\tau_2 \subset \tau_1$ ), then the same LDP holds with respect to the topology  $\tau_1$ .*

**Proof.** We employ Theorem 3.7 for the embedding  $\psi: (E, \tau_1) \rightarrow (E, \tau_2)$ , which is continuous because  $\tau_1$  is finer than  $\tau_2$ . We then conclude with Theorem 3.7. Furthermore, note that, since  $\psi$  is continuous, the measures  $\mu_N$  are well-defined as Borel measures on  $(E, \tau_2)$ .  $\square$

### 3.3 Varadhan's Integral Lemma

**Theorem 3.9 (Varadhan Lemma)** *Suppose that  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with a good rate function  $I: E \rightarrow [0, \infty]$ , and let  $H: E \rightarrow \mathbb{R}$  be a continuous function. Assume that either the tail-condition*

$$\lim_{M \rightarrow \infty} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}_{\mu_N}[e^{NH} \mathbb{1}\{H \geq M\}] = -\infty, \quad (3.6)$$

*or the moment condition for  $\gamma > 1$ ,*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{\gamma NH}] < \infty, \quad (3.7)$$

hold. Then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}_{\mu_N} [e^{NH}] = \sup_{x \in E} \{H(x) - I(x)\}.$$

**Remark 3.10** (a) This theorem is the natural extension of Laplace's method of computing parameter integrals in finite-dimensional spaces to infinite dimensional spaces.

(b) It is clear that any continuous function bounded from above satisfies the tail condition (3.6). The moment condition (3.7) implies the tail condition (3.6) as we see using Hölder's inequality,

$$\begin{aligned} \int_{\{H \geq M\}} e^{NH(x)} \mu_N(dx) &\leq \left( \int e^{\gamma NH(x)} \mu_N(dx) \right)^{1/\gamma} (\mu_N(H \geq M))^{1-\frac{1}{\gamma}} \\ &\leq \left( \int e^{\gamma NH(x)} \mu_N(dx) \right)^{1/\gamma} \left( e^{-\gamma MN} \int e^{\gamma NH(x)} \mu_N(dx) \right)^{1-\frac{1}{\gamma}} \\ &= \exp((1-\gamma)MN) \left( \int e^{\gamma NH(x)} \mu_N(dx) \right). \end{aligned}$$

◇

**Proof of Theorem 3.9.** The proof is an immediate consequence of the following two lemmas and Remark 3.10. □

**Lemma 3.11** *If  $H: E \rightarrow \mathbb{R}$  is lower semicontinuous and the large deviation lower bound holds with  $I: E \rightarrow [0, \infty]$ , then*

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH}] \geq \sup_{x \in E} \{H(x) - I(x)\}.$$

**Proof.** Pick  $x \in E$  and  $\delta > 0$ . Since  $F$  is lower semicontinuous, there exists an open neighbourhood  $G \ni x$  such that  $\inf_{y \in G} \{H(y)\} \geq H(x) - \delta$ . By the large deviation lower bound and the choice of  $G$ ,

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH}] &\geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH} \mathbb{1}_G] \geq \inf_{y \in G} \{H(y)\} + \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(G) \\ &\geq \inf_{y \in G} \{H(y)\} - \inf_{y \in G} \{I(y)\} \geq H(x) - I(x) - \delta. \end{aligned}$$

The statement now follows, since  $\delta > 0$  and  $x \in E$  are arbitrary. □

**Lemma 3.12** *If  $H: E \rightarrow \mathbb{R}$  is an upper semicontinuous for which the tail condition (3.6) holds, and if the large deviation upper bound holds with the good rate function  $I: E \rightarrow [0, \infty]$ , then*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH}] \leq \sup_{x \in E} \{H(x) - I(x)\}.$$

**Proof.** First consider a function  $H$  which is bounded above, i.e.

$$\sup_{x \in E} \{H(x)\} \leq M < \infty.$$

Clearly, this function satisfies the tail condition (3.6). For  $\alpha < \infty$  consider the compact level set  $\mathcal{L}_I(\alpha)$ . For  $x \in \mathcal{L}_I(\alpha)$  there exists a neighbourhood  $A_x$  of  $x$  such that

$$\inf_{y \in \overline{A_x}} \{I(y)\} \geq I(x) - \delta, \quad \sup_{y \in \overline{A_x}} \{H(y)\} \leq H(x) + \delta,$$

where the first inequality follows as  $I$  is lower semicontinuous and the second one is due to upper semicontinuity of  $H$ . From the open cover with the neighbourhoods  $A_x$  we can extract a finite cover of the level set  $\mathcal{L}_I(\alpha) \subset \bigcup_{i=1}^K A_{x_i}$ ,  $K \in \mathbb{N}$ . Therefore,

$$\begin{aligned} \mathbb{E}[e^{NH}] &\leq \sum_{i=1}^K \mathbb{E}[e^{NH} \mathbb{1}_{A_{x_i}}] + e^{NM} \mu_N((\bigcup_{i=1}^K A_{x_i})^c) \\ &\leq \sum_{i=1}^K e^{N(H(x_i) + \delta)} \mu_N(\overline{A_{x_i}}) + e^{NM} \mu_N((\bigcup_{i=1}^K A_{x_i})^c). \end{aligned}$$

We apply now the large deviation upper bound to the sets  $\overline{A_{x_i}}$  and use the fact that  $(\bigcup_{i=1}^K A_{x_i})^c \subset \mathcal{L}_I(\alpha)^c$  and arrive at

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH}] &\leq \max \left\{ \max_{1 \leq i \leq K} \{H(x_i) + \delta - \inf_{y \in \overline{A_{x_i}}} \{I(y)\}\}, M - \inf_{y \in (\bigcup_{i=1}^K A_{x_i})^c} \{I(y)\} \right\} \\ &\leq \max \left\{ \max_{1 \leq i \leq K} \{H(x_i) - I(x_i) + 2\delta\}, M - \alpha \right\} \\ &\leq \max \left\{ \sup_{x \in E} \{H(x) - I(x)\}, M - \alpha \right\} + 2\delta. \end{aligned}$$

Thus, for  $H$  bounded as above, the lemma follows by taking the limits  $\delta \rightarrow 0$  and  $\alpha \rightarrow \infty$ . To treat the general case, we use a cutoff parameter  $M > 0$  and define  $H_M(x) := H(x) \wedge M \leq H(x)$ , and use our arguments above for  $H_M$  to obtain

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH}] &\leq \sup_{x \in E} \{H(x) - I(x)\} \vee \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{NH} \mathbb{1}_{\{H \geq M\}}]. \end{aligned}$$

Now the tail condition (3.6) completes the proof by taking the limit  $M \rightarrow \infty$ .  $\square$

With Varadhan's Lemma we can obtain new LDPs for families of probability measures defined by Radon-Nikodym densities. In application the key is to include dependencies among random variables via densities which cannot be written as the product of single densities.

**Theorem 3.13 (Tilted LDP via Varadhan Lemma)** *Let  $(E, d)$  be a Polish space. Suppose that  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with a good rate function  $I: E \rightarrow [0, \infty]$ , and let  $H: E \rightarrow \mathbb{R}$  be a continuous function that is bounded from above. Then define*

$$Z_N(H) := \int_E e^{NH(x)} \mu_N(dx),$$

and the probability measure  $\mu_N^H \in \mathcal{M}_1(E)$  via the Radon-Nikodym density

$$\frac{d\mu_N^H}{d\mu_N}(x) = \frac{e^{NH(x)}}{Z_N(H)}, \quad x \in E.$$

Then the sequence  $(\mu_N^H)_{N \in \mathbb{N}}$  satisfies the LDP on  $E$  with rate  $n$  and rate function

$$I^H(x) = I(x) - H(x) + \sup_{y \in E} \{H(y) - I(y)\}, \quad x \in E. \quad (3.8)$$

**Proof.** From Theorem 3.9 we know that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Z_N(H) = \sup_{y \in E} \{H(y) - I(y)\}.$$

Then we obtain the large deviation bounds by simply repeating the above arguments in the proof of Theorem 3.9. For example, let  $K \subset E$  be closed, then

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N^H(K) &= \limsup_{N \rightarrow \infty} \frac{1}{N} \log \int_K e^{NH(x)} \mu_N(dx) - \limsup_{N \rightarrow \infty} \frac{1}{N} \log Z_N(H) \\ &\leq \sup_{y \in K} \{H(y) - I(y)\} - \sup_{y \in E} \{H(y) - I(y)\} = - \inf_{y \in K} \{I^H(y)\}, \end{aligned}$$

as  $I^H(x) = I(x) - H(x) - \inf_{y \in E} \{I(y) - H(y)\}$  and

$$- \sup_{y \in E} \{H(y) - I(y)\} = \inf_{y \in E} \{I(y) - H(y)\}.$$

The corresponding lower bound follows similarly. □

### 3.4 Bryc's Inverse Varadhan Lemma

We shall study an inverse to Varadhan's lemma. Suppose  $(\mu_N)_{N \in \mathbb{N}}$  is a sequence of probability measures  $\mu_N \in \mathcal{M}_1(E)$  over the Polish space  $(E, d)$ . In what follows, one can consider more general topological spaces but we content ourselves here with Polish spaces. For each Borel measurable function  $f: E \rightarrow \mathbb{R}$ , define

$$\Lambda_f := \lim_{N \rightarrow \infty} \frac{1}{N} \log \int_E e^{Nf(x)} \mu_N(dx), \quad (3.9)$$

provided the limit exists. In case we have a vector space structure on  $E$ , for example, as we have in Cramér's theorem, we consider (3.9) for linear functionals and called that limit then the logarithmic moment generating function. The key result in this section is that the LDP is a consequence of the exponential tightness and the existence of (3.9) for every  $f \in \mathcal{G}$  for some useful families  $\mathcal{G}$  of function on  $E$ . The minimal requirement on the space  $E$  is that  $E$  is a completely regular topological space, that is,  $E$  is Hausdorff, and for any closed set  $F \subset E$  and  $x \notin F$ , there exists a continuous function  $f: E \rightarrow [0, 1]$  such that  $f(x) = 1$  and  $f(y) = 0$  for all  $y \in F$ . Note that metric spaces and Hausdorff topological vector spaces are completely regular.

We denote  $\mathcal{C}_b(E)$  the space of bounded, real-valued continuous on  $E$ .

**Theorem 3.14 (Bryc)** *Suppose that the sequence  $(\mu_N)_{N \in \mathbb{N}}$  of probability measures  $\mu_N \in \mathcal{M}_1(E)$  over the Polish space is exponentially tight and that the limit in (3.9) exists for all  $f \in \mathcal{C}_b(E)$ . Then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP on  $E$  with rate  $N$  and good rate function*

$$I(x) = \sup_{f \in \mathcal{C}_b(E)} \{f(x) - \Lambda_f\}. \quad (3.10)$$

**Remark 3.15** ◇

**Proof of Theorem 3.14.** We have that  $\Lambda_0 = 0$  and thus we get that  $I \geq 0$ . The function  $I$  is lower semicontinuous since it is a supremum of continuous functions. As exponential tightness is given, it suffices to establish the upper bound for compact sets. We start with the lower bound:

Lower bound: Fix  $x \in E$ . As the metric space  $(E, d)$  is completely regular, there exists a continuous function  $f: E \rightarrow [0, 1]$ , such that  $f(x) = 1$  and  $f(y) = 0$  for all  $y \in G^c$ , where  $G \ni x$  is some open neighbourhood of  $x$ . Denote  $f_m = m(f - 1)$ ,  $m \in \mathbb{N}$ , and note that  $f_m \in \mathcal{C}_b(E)$ . Then

$$\int_E e^{N f_m(x)} \mu_N(dx) \leq e^{-mN} \mu_N(G^c) + \mu_N(G) \leq e^{-mN} + \mu_N(G).$$

We now use the fact that  $f_m \in \mathcal{C}_b(E)$  and that  $f_m(x) = 0$  to obtain the following lower bound,

$$\begin{aligned} \max\left\{\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(G), -m\right\} &\geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \int_E e^{N f_m(y)} \mu_N(dy) = \Lambda_{f_m} \\ &= -(f_m(x) - \Lambda_{f_m}) \geq -\sup_{f \in \mathcal{C}_b(E)} \{f(x) - \Lambda_f\} = -I(x), \end{aligned}$$

and the lower bound follows by letting  $m \rightarrow \infty$ . The reason why this lower bound works is that indicators on open sets are approximated well enough by bounded continuous functions.

Upper bound: Fix a compact set  $K \subset E$  and some  $\delta > 0$ , and define  $I^\delta(x) := \min\{I(x) - \delta, \frac{1}{\delta}\}$ . For any  $x \in K$  there exists  $f_x \in \mathcal{C}_b(E)$  such that

$$f_x(x) - \Lambda_{f_x} \geq I^\delta(x).$$

As  $f_x$  is continuous, there is neighbourhood  $A_x \ni x$  of  $x$ , such that

$$\inf_{y \in A_x} \{f_x(y) - f_x(x)\} \geq -\delta. \quad (3.11)$$

We obtain an upper bound via Chebycheff's inequality.

$$\mu_N(A_x) = \mathbb{E}[\mathbb{1}\{X_N \in A_x\}] \leq \mathbb{P}(f_x(X_N) - f_x(x) \geq -\delta) \leq e^{N\delta} \mathbb{E}[e^{N(f_x(X_N) - f_x(x))}].$$

Thus

$$\frac{1}{N} \log \mu_N(A_x) \leq \delta - \left(f_x(x) - \frac{1}{N} \log \int_E e^{N f_x(y)} \mu_N(dy)\right).$$

We now extract a finite cover,  $\bigcup_{i=1}^M A_{x_i}$ , from the open cover  $\bigcup_{x \in K} A_x \supset K$  of the compact set  $K$ . By the union of events bound,

$$\frac{1}{N} \log \mu_N(K) \leq \frac{1}{N} \log M + \delta - \min_{1 \leq i \leq M} \left\{ f_{x_i}(x_i) - \frac{1}{N} \log \int_E e^{N f_{x_i}(y)} \mu_N(dy) \right\},$$

and thus

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K) &\leq \delta - \min_{1 \leq i \leq M} \{f_{x_i}(x_i) - \Lambda_{f_{x_i}}\} \leq \delta - \min_{1 \leq i \leq M} \{I^\delta(x_i)\} \\ &\leq \delta - \inf_{x \in K} \{I^\delta(x)\}, \end{aligned}$$

and we conclude with the desired upper bound by letting  $\delta \downarrow 0$ . □

## 4 The Gärtner-Ellis theorem

We study various versions of the Gärtner-Ellis theorem, which has two key elements. One is that we consider topological vector spaces, and, secondly study now an approach to include sequences of not necessarily identical distributed and independent random variables. In order to showcase the main ideas we consider only the case for  $E = \mathbb{R}^d$ .

### 4.1 Gärtner-Ellis for $\mathbb{R}^d$

Let  $E = \mathbb{R}^d$ . The vector space structure is crucial for the following results. The set-up is as follows. We let  $(X_N)_{N \in \mathbb{N}}$  be a sequence of  $\mathbb{R}^d$ -valued random variables with law  $\mu_N \in \mathcal{M}_1(\mathbb{R}^d)$ . The *moment generating function* is

$$M_N(\lambda) := \mathbb{E}[e^{\langle \lambda, X_N \rangle}], \quad \lambda \in \mathbb{R}^d, \quad (4.1)$$

and we define  $\Lambda_N(\lambda) := \log M_N(\lambda)$ ,  $\lambda \in \mathbb{R}^d$ .

We also need the following notions which we define for a general topological Hausdorff vector space  $E$  with dual space  $E^*$ .

The *dual space*  $E^*$  of  $E$  consists of all continuous linear functionals on  $E$ . If  $(X_N)_{N \in \mathbb{N}}$  is a sequence of  $E$ -valued random variables  $X_N$  with law  $\mu_N \in \mathcal{M}_1(E)$ , the *logarithmic moment generating function* is defined to be

$$\Lambda_{\mu_N}(\lambda) = \log \mathbb{E}_{\mu_N}[e^{\langle \lambda, X_N \rangle}] = \log \int_E e^{\lambda(x)} \mu_N(dx), \quad \lambda \in E^*, \quad (4.2)$$

where for  $x \in E$  and  $\lambda \in E^*$ ,  $\langle \lambda, x \rangle$  denotes the value of  $\lambda(x) \in \mathbb{R}$ . Let

$$\bar{\Lambda}(\lambda) := \limsup_{N \rightarrow \infty} \frac{1}{N} \Lambda_{\mu_N}(N\lambda), \quad (4.3)$$

using the notation  $\Lambda(\lambda)$  whenever the *limit* exists.

**Definition 4.1** Suppose that  $E$  is a Hausdorff topological vector space with dual  $E^*$ . A point  $x \in E$  is called an *exposed point* of  $\bar{\Lambda}^*$  if there exists an *exposing hyperplane*  $\lambda \in E^*$  such that

$$\langle \lambda, x \rangle - \bar{\Lambda}^*(x) > \langle \lambda, z \rangle - \bar{\Lambda}^*(z), \quad \text{for all } z \neq x. \quad (4.4)$$

**Theorem 4.2 (Gärtner-Ellis for  $\mathbb{R}^d$ )** Suppose that  $(X_N)_{N \in \mathbb{N}}$  is a sequence of  $\mathbb{R}^d$ -valued vectors  $X_N$  and that  $\mu_N \in \mathcal{M}_1(\mathbb{R}^d)$  is the law of  $X_N$ . Assume that the following holds:

$$\Lambda(\lambda) := \lim_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_{\mu_N}(N\lambda) \quad \text{exists as an extended real number for all } \lambda \in \mathbb{R}^d, \quad (4.5)$$

and  $0 \in \mathcal{D}_\Lambda$ . Then the following holds.

(a) For every closed set  $F \subset \mathbb{R}^d$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(F) \leq - \inf_{x \in F} \{\Lambda^*(x)\}.$$

(b) Let  $\mathcal{E}$  be the set of exposed points of  $\Lambda^*$  with an exposing hyperplane  $\lambda \in \mathring{\mathcal{D}}_\Lambda$ . Then, for every open set  $G \subset \mathbb{R}^d$ ,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(G) \geq - \inf_{x \in G \cap \mathcal{E}} \{\Lambda^*(x)\}.$$

(c) If  $\Lambda$  is an essentially smooth, lower semi continuous function, then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with good rate function  $\Lambda^*$ .

**Remark 4.3** A convex function  $\Lambda: \mathbb{R}^d \rightarrow (-\infty, \infty]$  is *essentially smooth* if

(a)  $\mathring{\mathcal{D}}_\Lambda \neq \emptyset$ .

(b)  $\Lambda$  is differentiable in  $\mathring{\mathcal{D}}_\Lambda$ .

(c)  $\Lambda$  is steep, that is,  $\lim_{N \rightarrow \infty} |\nabla \Lambda(\lambda_N)| = \infty$  whenever  $(\lambda_N)_{N \in \mathbb{N}}$  sequence in  $\mathring{\mathcal{D}}_\Lambda$  converging to a point in the boundary  $\partial \mathcal{D}_\Lambda$  of  $\mathcal{D}_\Lambda$ .

In particular, when  $\mathcal{D}_\Lambda = \mathbb{R}^d$ , then  $\Lambda$  is essentially smooth and the LDP holds.  $\diamond$

**Lemma 4.4** Under the assumption (4.5) of Theorem 4.2 the following holds.

(a)  $\Lambda$  is convex and  $\Lambda > -\infty$  everywhere.

(b)  $\Lambda^*$  is a good rate, convex rate function.

**Proof.** (a) Clearly,  $\lambda \mapsto \log \mathbb{E}[e^{\langle X_N, \lambda \rangle}]$  is convex and thus the limit  $\Lambda$  is convex. Now  $\Lambda(0) = 0$  in conjunction with the convexity and  $0 \in \mathcal{D}_\Lambda$  gives  $\Lambda > -\infty$  everywhere.

(b) For all  $x \in \mathbb{R}^d$  we have  $\Lambda^*(x) \geq -\Lambda(0) = 0$  and thus  $\Lambda^* \geq 0$ . Furthermore,  $\Lambda^*$  is convex as a supremum of linear functions. There exists  $\delta > 0$  such that  $B_{2\delta}(0) \subset \mathring{\mathcal{D}}_\Lambda$ . Since  $\Lambda$  is convex, it is continuous on  $\mathring{\mathcal{D}}_\Lambda$  and thus

$$\sup_{\lambda \in B_{2\delta}(0)} \{\Lambda(\lambda)\} = C < \infty,$$

and thus

$$\Lambda^*(x) \geq \sup_{\lambda \in B_{2\delta}(0)} \{\langle x, \lambda \rangle - \Lambda(\lambda)\} \geq \delta|x| - C.$$

Thus  $\Lambda^*$  has bounded level sets. The lower semicontinuity of  $\Lambda^*$  implies that the level sets are also closed  $\in \mathbb{R}^d$  and thus all level sets are compact implying that  $\Lambda^*$  is a good rate function. □

**Proof of Theorem 4.2.** (a) Upper bound:

For  $x \in \mathbb{R}^d$  and  $\delta > 0$  define

$$\Lambda_\delta^*(x) := \min\{\Lambda^*(x) - \delta, \frac{1}{\delta}\}.$$

For all  $x \in \mathbb{R}^d$  there exists a vector  $\lambda_x \in \mathbb{R}^d$  such that

$$\langle x, \lambda_x \rangle - \Lambda(\lambda_x) \geq \Lambda_\delta^*(x),$$

and for this vector there is a neighbourhood  $A_x \ni x$  of  $x$  such that

$$\inf_{y \in A_x} \{\langle y - x, \lambda_x \rangle\} \geq -\delta.$$

We now employ Chebycheff's inequality again like in our proof of Bryc's theorem (Theorem 3.14) to obtain

$$\begin{aligned} \mu_N(A_x) &= \mathbb{P}(X_N \in A_x) \leq \mathbb{P}(\langle X_N - x, \lambda_x \rangle \geq -\delta) \leq e^{N\delta} \mathbb{E}[e^{N\langle X_N - x, \lambda_x \rangle}] \\ &= e^{N\delta} \mathbb{E}[e^{N\lambda_x}] e^{-N\langle x, \lambda_x \rangle}. \end{aligned}$$

Pick  $K \subset \mathbb{R}^d$  compact and extract a finite open cover of neighbourhoods like above,

$$K \subset \bigcup_{i=1}^M A_{x_i}.$$

This way we get

$$\frac{1}{N} \log \mu_N(K) \leq \frac{1}{N} \log M + \delta - \min_{1 \leq i \leq M} \{\langle x_i, \lambda_{x_i} \rangle - \frac{1}{N} \log \mathbb{E}[e^{N\lambda_{x_i}}]\}.$$

Henceforth

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K) &\leq \delta - \min_{1 \leq i \leq M} \{ \langle x_i, \lambda_{x_i} \rangle - \Lambda(\lambda_{x_i}) \} \\ &\leq \delta - \min_{1 \leq i \leq M} \{ \Lambda_\delta^*(x_i) \} \leq \delta - \inf_{y \in K} \{ \Lambda_\delta^*(y) \}, \end{aligned}$$

and we conclude with the upper bound for the compact set  $K$  by letting  $\delta \downarrow 0$ . Now we let  $F \subset \mathbb{R}^d$  be a closed subset. We introduced an approximation parameter  $M \in \mathbb{N}$  such that  $F \cap [-M, M]^d$  is compact for any  $M \in \mathbb{N}$ . Thus

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(F) \leq \max \left\{ - \inf_{y \in F \cap [-M, M]^d} \{ \Lambda^*(y) \}, -K_M \right\},$$

where

$$-K_M := \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(\mathbb{R}^d \setminus [-M, M]^d).$$

If  $K_M \rightarrow \infty$  as  $M \rightarrow \infty$ , our claim follows because

$$\lim_{M \rightarrow \infty} \inf_{y \in F \cap [-M, M]^d} \{ \Lambda^*(y) \} = \inf_{y \in F} \{ \Lambda^*(y) \}.$$

Here is the point where we use our assumption  $0 \in \mathring{\mathcal{D}}_\Lambda$ . This assumption ensures that there exists  $\delta_i > 0, \eta_i > 0; i = 1, \dots, d$ , such that

$$\Lambda(-\eta_i \mathbf{e}_i) < \infty \text{ and } \Lambda(\delta_i \mathbf{e}_i) < \infty, \quad i = 1, \dots, d.$$

We obtain the following estimates for the  $i$ th coordinates using again exponential Chebycheff's inequality,

$$\begin{aligned} \mathbb{P}(X_N^{(i)} \leq -M) &\leq e^{-N\eta_i M} \mathbb{E}[e^{-N\eta_i \mathbf{e}_i}] \\ \mathbb{P}(X_N^{(i)} \geq M) &\leq e^{-N\delta_i M} \mathbb{E}[e^{N\delta_i \mathbf{e}_i}]. \end{aligned}$$

Hence

$$\begin{aligned} -K_M &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(\exists i: X_N^{(i)} \notin [-M, M]) \leq - \min_{1 \leq i \leq d} \{ \min\{\delta_i, \eta_i\} \} M \\ &\quad + \max_{1 \leq i \leq d} \{ \max\{ \Lambda(-\delta_i \mathbf{e}_i), \Lambda(\delta_i \mathbf{e}_i) \} \}, \end{aligned}$$

and finally  $K_M \rightarrow \infty$  as  $M \rightarrow \infty$ . Consequently, by the union of events bound,

$$\lim_{M \rightarrow \infty} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(([-M, M]^d)^c) = -\infty,$$

i.e.,  $(\mu_N)_{N \in \mathbb{N}}$  is exponentially tight.

Lower bound (b): We need to show that for  $y \in \mathcal{E}$ ,

$$\lim_{\delta \rightarrow 0} \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(B_\delta(y)) \geq -\Lambda^*(y). \quad (4.6)$$

Fix  $y \in \mathcal{E}$  and let  $\eta \in \mathring{\mathcal{D}}_\Lambda$  denote the exposing hyperplane for  $y$ . For  $N$  sufficiently large we have that  $\Lambda_{\mu_N}(N\eta) < \infty$  and we can define the new measures  $\tilde{\mu}_N$  via the density,

$$\frac{d\tilde{\mu}_N}{d\mu_N}(z) = \exp(N\langle \eta, z \rangle - \Lambda_{\mu_N}(N\eta)). \quad (4.7)$$

Then we get with some calculation for the change of measure,

$$\begin{aligned} \frac{1}{N} \log \mu_N(B_\delta(y)) &= \frac{1}{N} \Lambda_{\mu_N}(N\eta) - \langle \eta, y \rangle + \frac{1}{N} \int_{B_\delta(y)} e^{N\langle \eta, y-z \rangle} \tilde{\mu}_N(dz) \\ &\geq \frac{1}{N} \Lambda_{\mu_N}(N\eta) - \langle \eta, y \rangle - |\eta|\delta + \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)). \end{aligned}$$

Therefore,

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(B_\delta(y)) &\geq \Lambda(\eta) - \langle \eta, y \rangle + \liminf_{N \rightarrow \infty} \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)) \\ &\geq -\Lambda^*(y) + \liminf_{N \rightarrow \infty} \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)) \end{aligned}$$

The obstacle comes from the missing independence, since the weak law of large numbers no longer applies. The strategy is to utilise the upper bound in (a). For that we analyse the logarithmic moment generating function for  $\tilde{\mu}_N$ . One can easily show that

$$\frac{1}{N} \tilde{\Lambda}_{\tilde{\mu}_N}(N\lambda) \xrightarrow{N \rightarrow \infty} \tilde{\Lambda}(\lambda) = \Lambda(\lambda + \eta) - \Lambda(\eta),$$

where the limiting moment generating function  $\tilde{\Lambda}$  satisfies assumption (4.5) as clearly  $\tilde{\Lambda}(0) = 0$  and  $\tilde{\Lambda} < \infty$  for  $|\lambda|$  small enough. Define

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

Since  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  satisfies the assumptions (4.5), we can apply Lemma 4.4 and part (a) above to show that  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  satisfies a large deviation upper bound with the good rate function  $\tilde{\Lambda}^*$ . Thus, for the closed set  $B_\delta(y)^c$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)^c) \leq - \inf_{x \in B_\delta(y)^c} \{ \tilde{\Lambda}^*(x) \} = \tilde{\Lambda}^*(x_0)$$

for some point  $x_0 \neq y$ . This follows from the compact level sets as a lower semicontinuous function attains its minimum over a compact set. We are left to show that  $\tilde{\Lambda}^*(x_0) > 0$ . At this point we use the property that  $y$  is an exposed point for  $\Lambda^*$  with exposing hyperplane  $\eta$ . First,

$$\Lambda^*(y) \geq \langle \eta, y \rangle - \Lambda(\eta),$$

and thus  $\Lambda(\eta) \geq \langle \eta, y \rangle - \Lambda^*(y)$ . Then

$$\tilde{\Lambda}^*(x_0) = \Lambda^*(x_0) - \langle \eta, x_0 \rangle + \Lambda(\eta) \geq \Lambda^*(x_0) - \langle \eta, x_0 \rangle + \langle \eta, y \rangle - \Lambda^*(y) > 0.$$

Thus, for every  $\delta > 0$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)^c) < 0.$$

This implies that  $\tilde{\mu}_N(B_\delta(y)^c) \rightarrow 0$  as  $n \rightarrow \infty$  and hence  $\tilde{\mu}_N(B_\delta(y)) \rightarrow 1$  as  $N \rightarrow \infty$ , and in particular,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log \tilde{\mu}_N(B_\delta(y)) = 0.$$

(c) We need to show the lower bound without the intersection with the set  $\mathcal{E}$  of exposed points. This requires some deeper results in convex analysis. We need the notation. For every non-empty convex set  $C \subset \mathbb{R}^d$ , the *relative interior* of  $C$ , denoted  $\text{ri}(C)$ , is defined as the set

$$\text{ri}(C) := \{y \in C : x \in C \Rightarrow y - (\varepsilon(x - y)) \in C \text{ for some } \varepsilon > 0\}. \quad (4.8)$$

Then, according to [Roc70], the following holds: If  $\Lambda$  is an essentially smooth, lower semicontinuous, convex function, then  $\text{ri}(\mathcal{D}_{\Lambda^*}) \subset \mathcal{E}$ , where  $\mathcal{E}$  is the set of exposed points. To show that

$$\inf_{y \in G \cap \mathcal{E}} \{\Lambda^*(y)\} = \inf_{y \in G} \{\Lambda^*(y)\},$$

it suffices to show that, for an open set  $G \subset \mathbb{R}^d$ ,

$$\inf_{y \in G \cap \text{ri}(\mathcal{D}_{\Lambda^*})} \{\Lambda^*(y)\} \leq \inf_{y \in G} \{\Lambda^*(y)\}. \quad (4.9)$$

Now (4.9) holds when  $G \cap \text{ri}(\mathcal{D}_{\Lambda^*}) = \emptyset$ . Otherwise, pick  $y \in G \cap \text{ri}(\mathcal{D}_{\Lambda^*})$  and  $z \in \text{ri}(\mathcal{D}_{\Lambda^*})$ . Then, for all  $\delta > 0$  sufficiently small enough,

$$\delta z + (1 - \delta)y \in G \cap \text{ri}(\mathcal{D}_{\Lambda^*}),$$

and thus

$$\inf_{y \in G \cap \text{ri}(\mathcal{D}_{\Lambda^*})} \{\Lambda^*(y)\} \leq \lim_{\delta \downarrow 0} \Lambda^*(\delta z + (1 - \delta)y) \leq \Lambda^*(y).$$

Taking the infimum over  $y \in G \cap \text{ri}(\mathcal{D}_{\Lambda^*})$ , we get the claim (4.9) and thus our statement (c), i.e., the full LDP. □

## 4.2 A general upper bound - topological vector spaces

We finish our basic introduction to the theory of large deviations with considering solely Hausdorff topological vector space  $E$ . The dual of  $E$ , denoted  $E^*$ , is the space of all continuous linear functionals. Suppose that  $(X_N)_{N \in \mathbb{N}}$  is a sequence of  $E$ -valued random variables such that  $X_N$  has law  $\mu_N \in \mathcal{M}_1(E)$ .

We define the *logarithmic moment generating function* for  $\mu_N$  as

$$\Lambda_{\mu_N}(\lambda) := \log \mathbb{E}[e^{\langle \lambda, X_N \rangle}] = \log \int_E e^{\lambda(x)} \mu_N(dx), \quad \lambda \in E^*, \quad (4.10)$$

where for  $x \in E$  and  $\lambda \in E^*$ ,  $\langle \lambda, x \rangle = \lambda(x)$  denotes the value  $\lambda(x) \in \mathbb{R}$ . Furthermore, define

$$\bar{\Lambda}(\lambda) := \limsup_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_{\mu_N}(N\lambda), \quad (4.11)$$

and use the notation  $\Lambda(\lambda)$  when the limit exists. In our current setup, the Fenchel-Legendre transform of a function  $f: E^* \rightarrow [-\infty, \infty]$  is defined as

$$f^*(x) := \sup_{\lambda \in E^*} \{\langle \lambda, x \rangle - f(\lambda)\}, \quad x \in E. \quad (4.12)$$

In the following we denote  $\bar{\Lambda}^*$  the Legendre-Fenchel transform of  $\bar{\Lambda}$ , and  $\Lambda^*$  denotes that of  $\Lambda$  when the latter exists for all  $\lambda \in E^*$ .

**Theorem 4.5 (A General Upper bound)** *Let  $(\mu_N)_{N \in \mathbb{N}}$  be a sequence of probability measures. Then the following holds.*

- (a)  $\bar{\Lambda}$  of (4.11) is convex on  $E^*$  and  $\bar{\Lambda}^*$  is a convex rate function.
- (b) For any compact set  $K \subset E$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(K) \leq - \inf_{x \in K} \{\bar{\Lambda}^*(x)\}. \quad (4.13)$$

**Proof.** (a) Using the linearity of elements in the dual space and applying Hölder's inequality, one can show that the functions  $\Lambda_{\mu_N}(N\lambda)$  are convex. Thus

$$\bar{\Lambda}(\cdot) := \limsup_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_{\mu_N}(N\cdot)$$

is also a convex function. As  $\Lambda_{\mu_N}(0) = 0$  for all  $N \in \mathbb{N}$ , we have that  $\bar{\Lambda}(0) = 0$  and thus  $\bar{\Lambda}^* \geq 0$ . Note that  $g(\lambda) := \langle \lambda, x \rangle - \bar{\Lambda}(\lambda)$  is continuous for every  $\lambda \in E^*$ . Then the lower semicontinuity of  $\bar{\Lambda}^*$  follows from the fact that the supremum over continuous functions is lower semicontinuous. The convexity is shown as in Lemma 1.11.

(b) The upper bound follows exactly the steps in the proof of the upper bound in Theorem 3.14. Actually, the proof here is easier as it uses the continuous linear functions and the logarithmic moment generating function. Details are left for the reader. □

Having now a general upper bound in Theorem 4.5, we turn next to sufficient conditions for the existence of a complementary lower bound. Recall Definition 4.1

about exposed points and exposing hyperplanes in the dual  $E^*$ . The new ingredient in comparison with the Gärtner-Ellis theorem, see Theorem 4.2, is now the assumption that the sequence of probability measures is exponentially tight.

**Theorem 4.6 (Abstract Gärtner-Ellis Theorem)** *Let  $(\mu_N)_{N \in \mathbb{N}}$  be an exponentially tight sequence of probability measures on the Hausdorff topological space  $E$ .*

(a) *For every closed set  $F \subset E$ ,*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(F) \leq - \inf_{x \in F} \{\Lambda^*(x)\}.$$

(b) *Let  $\mathcal{E}$  be the set of exposed points of  $\Lambda^*$  with an exposing hyperplane  $\lambda \in \mathring{\mathcal{D}}_\Lambda$  for which*

$$\Lambda(\lambda) = \lim_{N \rightarrow \infty} \frac{1}{N} \Lambda_{\mu_N}(N\lambda) \text{ exists and } \overline{\Lambda}(\gamma\lambda) < \infty \text{ for some } \gamma > 1.$$

*Then, for every open set  $G \subset \mathbb{R}^d$ ,*

$$\liminf_{n \rightarrow \infty} \frac{1}{N} \log \mu_N(G) \geq - \inf_{x \in G \cap \mathcal{E}} \{\Lambda^*(x)\}.$$

(c) *If for every open set  $G \subset E$ ,*

$$\inf_{x \in G \cap \mathcal{E}} \{\Lambda^*(x)\} = \inf_{x \in G} \{\Lambda^*(x)\},$$

*then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with good rate function  $\overline{\Lambda}^*$ .*

We are not proving this theorem, see [DZ98] for details. The crucial point is to show that (c) holds, and the following statement for Banach spaces summarises frequent approaches to proving large deviation principles. Recall the following definition from analysis and functional analysis.

**Definition 4.7** A function  $f: E^* \rightarrow \mathbb{R}$  is *Gâteaux differentiable* if, for every  $\lambda, \theta \in E^*$ , the function  $f(\lambda + t\theta)$  is differentiable with respect to  $t$  at  $t = 0$ .

**Corollary 4.8** *Let  $(\mu_N)_{N \in \mathbb{N}}$  be an exponentially tight sequence of probability measures on a Banach space  $E$ . Suppose that the function  $\Lambda(\cdot) = \lim_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_{\mu_N}(N\cdot)$  is finite valued, Gâteaux differentiable, and lower semi continuous in  $E^*$  with respect to the weak\* topology. Then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with the good rate function  $\Lambda^*$ .*

**Proof.** The crucial point is to show that (c) in Theorem 4.6 follows under the given assumptions. This is an intricate and delicate proof using a fair amount of variational analysis techniques, and we therefore skip the details here which can be found in [dH00] or [DZ98].

□

### 4.3 Summary: general Cramér's theorem and general Sanov's theorem

We study now study general LDPs for sequences of i.i.d. random variables. We turn to a general Cramér Theorem first. The following assumption formalises the conditions required for our approach the Cramér's theorem.

#### Assumption 4.9

- (a)  $E$  is a locally convex, Hausdorff, topological real vector space.  $\mathcal{E} \subset E$  is a closed, convex subset of  $E$  such that  $\mu(\mathcal{E}) = 1$  and  $\mathcal{E}$  can be made into a Polish space with respect to the topology induced by  $E$ .
- (b) The closed convex hull of each compact  $K \subset \mathcal{E}$  is compact.

**Theorem 4.10 (General Cramér Theorem)** *Let Assumption 4.9 hold. Then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies a weak LDP with rate function  $\Lambda^*$ . Moreover, for every open, convex subset  $A \subset E$ ,*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mu_N(A) = - \inf_{x \in A} \{\Lambda^*(x)\}. \quad (4.14)$$

**Proof.** The proof is quite long and uses sub-additivity property of the law  $\mu_N$  of the empirical mean  $S_N$ . Here,  $\mu_N = \mu^{\otimes N} \circ S_N^{-1}$ , with  $\mu \in \mathcal{M}_1(E)$  the law of the i.i.d. sequence. Details are in Chapter 6 of [DZ98].  $\square$

The following direct corollary of Theorem 4.10 for  $E = \mathcal{E} = \mathbb{R}^d$  is a considerable strengthening of Cramér's theorem (Theorem 1.18 or Theorem 4.2 for i.i.d. sequences), since it dispenses with the requirement that either  $\mathcal{D}_\Lambda = \mathbb{R}^d$  or  $\Lambda$  be steep.

**Corollary 4.11** *The sequence  $(\mu_N)_{N \in \mathbb{N}}$  of the laws  $\mu_N = \mu^{\otimes N} \circ S_N^{-1}$  of the empirical means of  $\mathbb{R}^d$  valued i.i.d. random variables with law  $\mu \in \mathcal{M}_1(\mathbb{R}^d)$  satisfies a weak LDP with convex rate function  $\Lambda^*$ . Moreover, if  $0 \in \mathring{\mathcal{D}}_\Lambda$ , then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the full LDP with the good, convex rate function  $\Lambda^*$ .*

**Proof.** The weak LDP is a direct consequence of Theorem 4.10 as Assumption 4.9 holds. If  $0 \in \mathring{\mathcal{D}}_\Lambda$ , the full LDP follows, since then the sequence  $(\mu_N)_{N \in \mathbb{N}}$  is exponentially tight.  $\square$

We now turn to discuss a general version of Sanov's theorem. Let  $(Y_i)_{i \in \mathbb{N}}$  be an i.i.d. sequence of  $E$ -valued random variables with law  $\mu \in \mathcal{M}_1(E)$ , where  $(E, d)$  is a Polish space. Then the framework is as follows. The random delta measures  $\delta_{Y_k}$  are element in the space of *finite signed measures* denoted  $\mathcal{M}(E)$ . The space  $\mathcal{M}(E)$  is a topological Hausdorff vector space and  $\mathcal{M}_1(E)$  is a closed convex subset of  $\mathcal{M}(E)$ . It turns out that the space  $\mathcal{M}(E)$  is Polish, and so is  $\mathcal{M}_1(E)$ . We skip all

topological details and observe that Theorem 4.10 leads to the following statement for the *empirical measure*  $L_N^Y$ ,

$$L_N^Y = \frac{1}{N} \sum_{k=1}^N \delta_{Y_k}, \quad Y = (Y_1, \dots, Y_N).$$

The *relative entropy* of  $\nu \in \mathcal{M}_1(E)$  with respect to  $\mu \in \mathcal{M}_1(E)$  is denoted  $H(\nu|\mu)$ , and is defined by

$$H(\nu|\mu) = \begin{cases} \int_E f(x) \log f(x) \mu(dx) & , \text{ if } f = \frac{d\nu}{d\mu} \text{ exists,} \\ +\infty & , \text{ otherwise.} \end{cases} \quad (4.15)$$

**Theorem 4.12 (General Sanov Theorem)** (a) *The empirical measures  $L_N^Y$  satisfy a weak LDP in  $\mathcal{M}_1(E)$  with convex rate function*

$$\Lambda^*(\nu) = \sup_{f \in \mathcal{C}_b(E)} \{ \langle f, \nu \rangle - \Lambda(f) \}, \quad (4.16)$$

where for  $f \in \mathcal{C}_b(E)$ ,

$$\Lambda(f) = \log \int_E e^{f(x)} \mu(dx).$$

(b) *The laws of  $L_N^Y$  are exponentially tight.*

(c) *The rate function in (a) is*

$$\Lambda^*(\nu) = H(\nu|\mu), \quad \nu \in \mathcal{M}_1(E). \quad (4.17)$$

**Proof.** See Chapter in [DZ98]. □

## 5 Large deviations for Markov chains

### 5.1 Discrete time finite stater space Markov chains

We now study large deviation principles for sequence of random variables with a dependence structure. The focus is on Markov chains where the index gives the discrete time and the dependence structure is given in terms of the Markov probability respectively the stochastic matrix of transition probabilities. We consider finite state spaces  $E$  throughout this section. We consider sequences  $(Y_i)_{i \in \mathbb{N}}$  of  $E$ -valued random variables  $Y_i$  and denote  $P = (p(x, y))_{x, y \in E}$  the *stochastic matrix* associated with the Markov chain  $(Y_i)_{i \in \mathbb{N}}$ . The entries of the matrix  $P$  are elements in  $[0, 1]$  and their row sums are one. We denote  $P_\sigma$  the Markov probability measure associated with the transition matrix  $P$  and initial state  $\sigma \in E$ , i.e.,

$$P_\sigma(Y_1 = y_1, \dots, Y_N = y_N) = p(\sigma, y_1) \prod_{i=1}^{N-1} p(y_i, y_{i+1}), \quad y_1, \dots, y_N \in E, N \in \mathbb{N}. \quad (5.1)$$

We denote  $E_\sigma$  the expectation with respect to  $P_\sigma$ . A matrix  $B$  with nonnegative entries is called *irreducible*, if for any pair of indices  $i, j$  there exists an  $m = m(i, j)$  such that  $B^m(i, j) > 0$ . Irreducibility is equivalent to the condition that one may find for each  $i, j$  a sequence of indices  $i_1, \dots, i_m$  such that  $i_1 = i, i_m = j$  and  $B(i_k, i_{k+1}) > 0$  for  $k = 1, \dots, m-1$ . We state the following important result from linear algebra of matrices.

**Theorem 5.1 (Perron-Frobenius)** *Let  $B = (B(x, y))_{x, y \in E}$  be an irreducible matrix. Then  $B$  possesses an eigenvalue  $\varrho$ , called the Perron-Frobenius eigenvalue, such that the following holds:*

- (a)  $\varrho > 0$ .
- (b) For any eigenvalue  $\lambda$  of  $B$ ,  $|\lambda| \leq \varrho$ .
- (c) There exist left and right eigenvectors for the eigenvalue  $\varrho$  that have strictly positive coordinates.
- (d) The left and right eigenvectors  $\mu, \theta$  corresponding to the eigenvalue  $\varrho$  are unique up to a constant multiple.
- (e) For every  $x \in E$  and  $\varphi = (\varphi_x)_{x \in E}$  such that  $\varphi_x > 0$  for all  $x \in E$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{y \in E} B^n(x, y) \varphi_y \right) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{y \in E} B^n(y, x) \varphi_y \right) = \log \varrho. \quad (5.2)$$

**Proof.** (a)-(d) are standard in linear algebra and details can be found in the following books specialised on linear algebra for stochastic processes, [Sen81, Nor04, Str05]. To prove (e), we define

$$\alpha := \max_{x \in E} \{\theta_x\}, \beta := \min_{x \in E} \{\theta_x\} > 0, \quad \text{and} \quad \gamma := \max_{x \in E} \{\varphi_x\}, \delta := \min_{x \in E} \{\varphi_x\} > 0, \quad (5.3)$$

where  $\theta$  is the right eigenvector corresponding to  $\varrho$ . Then, for all  $x, y \in E$ ,

$$\frac{\gamma}{\beta} B^n(x, y) \theta_y \geq B^n(x, y) \varphi_y \geq \frac{\delta}{\alpha} B^n(x, y) \theta_y.$$

Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{y \in E} B^n(x, y) \varphi_y \right) &= \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{y \in E} B^n(x, y) \theta_y \right) = \lim_{n \rightarrow \infty} \frac{1}{n} \log (\varrho^n \theta_x) \\ &= \log \varrho, \end{aligned} \quad (5.4)$$

We show in the same way that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{y \in E} \varphi_y B^n(y, x) \right) = \log \varrho.$$

□

We study first additive functionals of Markov chains,

$$Z_N := \frac{1}{N} \sum_{k=1}^N X_k,$$

where  $X_k = f(Y_k)$  for a given deterministic function  $f: E \rightarrow \mathbb{R}^d$ . For any  $\lambda \in \mathbb{R}^d$  we denote  $P_\lambda$  the matrix with entries

$$P_\lambda(x, y) = p(x, y)e^{\langle \lambda, f(y) \rangle}, \quad x, y \in E, \quad (5.5)$$

and Perron-Frobenius eigenvalue  $\varrho(P_\lambda)$ . We see that  $\mathbb{P} - \lambda$  is irreducible if and only if  $P$  is irreducible.

**Theorem 5.2 (LDP for  $(Z_N)_{N \in \mathbb{N}}$ )** *Suppose that the stochastic matrix  $P$  is irreducible. Then the empirical mean sequence  $(Z_N)_{N \in \mathbb{N}}$  satisfies the large deviation principle on  $\mathbb{R}^d$  with rate  $N$  and rate function  $I$ , defined as*

$$I(x) := \sup_{\lambda \in \mathbb{R}^d} \{ \langle \lambda, x \rangle - \log \varrho(P_\lambda) \}. \quad (5.6)$$

**Proof.** We use the Gärtner-Ellis theorem 4.2 and observe that it suffices to show the following:

1.)

$$\Lambda(\lambda) := \lim_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_N(N\lambda) = \lim_{N \rightarrow \infty} \frac{1}{N} \log E_\sigma \left[ e^{N \langle \lambda, Z_N \rangle} \right]$$

exists for every  $\lambda \in \mathbb{R}^d$ .

2.)  $\mathcal{D}_\Lambda = \mathbb{R}^d$  and  $\Lambda$  is differentiable in  $\mathbb{R}^d$ .

3.)  $\Lambda(\lambda) = \log \varrho(P_\lambda)$ .

The first statement follows easily from

$$\begin{aligned} \Lambda_N(N\lambda) &= \log E_\sigma \left[ e^{\langle \lambda, \sum_{k=1}^N X_k \rangle} \right] = \log \left( \sum_{y_1, \dots, y_N} P_\sigma(Y_1 = y_1, \dots, Y_N = y_N) \prod_{k=1}^N e^{\langle \lambda, f(X_k) \rangle} \right) \\ &= \log \left( \sum_{y_1, \dots, y_N} p(\sigma, y_1) e^{\langle \lambda, f(y_1) \rangle} \cdots p(y_{N-1}, y_N) e^{\langle \lambda, f(y_N) \rangle} \right) = \log \sum_{y_N \in E} P_\lambda^N(\sigma, y_N). \end{aligned}$$

$P$  is irreducible and thus so is  $P_\lambda$ . Thus we apply Theorem 5.1 to  $P_\lambda$  and get that  $\Lambda(\lambda) = \log \varrho(P_\lambda)$ . Since  $E$  is finite,  $\varrho(P_\lambda)$ , being an isolated root of the characteristic equation for  $P_\lambda$ , is positive, finite, and differentiable with respect to  $\lambda$ , see [Sen81]. This gives 2.) and 3.), and we obtain that the limiting moment generating function  $\Lambda$  is essentially smooth with  $\mathcal{D}_\Lambda = \mathbb{R}^d$  and thus, according to Theorem 4.2, we conclude with our statement. □

The *empirical measure* for the Markov chain  $(Y_k)_{k \in \mathbb{N}}$  is a random probability measure (vector)  $L_N^Y \in \mathcal{M}_1(E)$  where  $Y = (Y_1, \dots, Y_N)$ , and which is defined as

$$L_N^Y(x) = \frac{1}{N} \sum_{k=1}^N \mathbb{1}_x(Y_k), \quad x \in E. \quad (5.7)$$

Suppose that  $P$  is irreducible and that  $\mu$  is stationary distribution which is the unique left eigenvector according to Theorem 5.1. The ergodic theorem tells us that, when  $P$  is aperiodic and the initial state is distributed according to the stationary distribution  $\mu$ , that

$$L_N^Y \rightarrow \mu \text{ in probability as } N \rightarrow \infty.$$

Thus we can expect some large deviation behaviour away from this convergence. As before, we find a deterministic function and trace our the following result back to our result in Theorem 5.2.

$$f: E \rightarrow \{0, 1\}^E \subset \mathbb{R}^{|E|}, y \mapsto f(y) = (\mathbb{1}_x(y))_{x \in E}. \quad (5.8)$$

Then

$$Z_N = \frac{1}{N} \sum_{k=1}^N f(Y_k) = \frac{1}{N} \sum_{k=1}^N (\mathbb{1}_x(Y_k))_{x \in E} = (L_N^Y(x))_{x \in E}.$$

We embed  $\mathcal{M}_1(E) \subset \mathbb{R}^{|E|}$  and define for any  $q \in \mathcal{M}_1(E)$ ,

$$I(q) := \sup_{\lambda \in \mathbb{R}^d} \{ \langle \lambda, q \rangle - \log \varrho(P_\lambda) \}, \quad (5.9)$$

where

$$P_\lambda(x, y) = P(x, y)e^{\lambda_y}, \quad x, y \in E. \quad (5.10)$$

We get the following large deviation result for  $(L_N^Y)_{N \in \mathbb{N}}$  directly from Theorem 5.2.

**Theorem 5.3 (LDP for  $L_N^Y$  - Sanov's theorem for Markov chains)** *Under the same assumptions as in Theorem 5.2, the large deviation principle holds for the empirical measures  $(L_N^Y)_{N \in \mathbb{N}}$  with respect to the Markov chain distribution with rate  $N$  and rate function given in (5.9).*

We obtain a variational representation of the rate function in Theorem 5.3. For a vector  $u \in \mathbb{R}^E$  we write  $u \gg 0$  when  $u_x > 0$  for all  $x \in E$ .

**Theorem 5.4 (Variational expression for rate function  $I$  in Theorem 5.3)**

$$I(q) = J(q) := \begin{cases} \sup_{u \in \mathbb{R}^E, u \gg 0} \left\{ \sum_{x \in E} q_x \log \frac{u_x}{(uP)_x} \right\} & , \text{ if } q \in \mathcal{M}_1(E), \\ +\infty & , \text{ if } q \notin \mathcal{M}_1(E). \end{cases}$$

**Remark 5.5** If  $(Y_k)_{k \in \mathbb{N}}$  is an i.i.d. sequence of  $E$ -valued random variables  $Y_k$ , then the rows of  $P$  are identical, i.e.,  $p(x, y) = \mu_x, x, y \in E$ . Then

$$J(q) = H(q|\mu),$$

the rate function from Sanov's theorem. ◇

**Proof.** As before, we embed  $\mathcal{M}_1(E) \subset \mathbb{R}^{|E|}$  and note that  $\mathcal{M}_1(E)$  is a closed subset. Thus  $\mathcal{M}_1(E)^c$  is open, and the large deviation lower bound from Theorem 5.3 yields that

$$-\infty = \liminf_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}_\sigma(L_N^Y \in \mathcal{M}_1(E)^c) \geq - \inf_{q \notin \mathcal{M}_1(E)} \{I(q)\}$$

because  $L_N^Y \in \mathcal{M}_1(E)$  by definition. Thus  $I(q) = +\infty$  for  $q \notin \mathcal{M}_1(E)$ .

Let  $q \in \mathcal{M}_1(E)$  and pick  $u \gg 0$  and define

$$\lambda_x := \log \left( \frac{u_x}{(uP)_x} \right), \quad x \in E.$$

We observe that  $uP \gg 0$  as  $u \gg 0$  and  $P$  is irreducible. Recall the function  $f$  from Theorem 5.3,  $f(y) = (\mathbb{1}_x(y))_{x \in E}$ . Then, for every  $y \in E$ ,

$$(uP_\lambda)_y = \sum_{x \in E} u_x P_\lambda(x, y) = \sum_{x \in E} u_x P(x, y) e^{\langle \lambda, f(y) \rangle} = \sum_{x \in E} u_x P(x, y) e^{\lambda_y} = u_y.$$

Hence  $uP_\lambda^n = u$  and according to (e) in Theorem 5.1 we have

$$\log \varrho(P_\lambda) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{x \in E} u_x P_\lambda^n(x, x) \right) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left( \sum_{x \in E} u_x \right) = 0,$$

and thus  $\varrho(P_\lambda) = 1$ . We get therefore a lower bound in (5.9),

$$I(q) \geq \sum_{x \in E} q_x \log \frac{u_x}{(uP)_x},$$

and  $I(q) \geq J(q)$ . To show the reverse inequality, fix a vector  $\lambda \in \mathbb{R}^E$  and let  $u^* \gg 0$  be the left eigenvector for  $\varrho(P_\lambda)$ . Then  $u^* P_\lambda = \varrho(P_\lambda) u^*$ , and

$$\langle \lambda, q \rangle + \sum_{x \in E} q_x \log \frac{(u^* P)_x}{u_x^*} = \sum_{x \in E} q_x \log \frac{(u^* P_\lambda)_x}{u_x^*} = \sum_{x \in E} q_x \log \varrho(P_\lambda) = \log \varrho(P_\lambda).$$

Thus

$$\langle \lambda, q \rangle - \log \varrho(P_\lambda) \leq \sup_{u \in \mathbb{R}^E: u \gg 0} \left\{ \sum_{x \in E} \log \frac{u_x}{(uP)_x} \right\} = J(q),$$

and therefore  $I(q) \leq J(q)$ . □

## 5.2 Pair empirical measures for Markov chains

We now compare our large deviation principle for the empirical pair measure for i.i.d. sequence with the case of a Markov chain. The *pair empirical measure* of the Markov chain  $(Y_k)_{k \in \mathbb{N}}$  is defined as

$$L_N^{2,Y} := \frac{1}{N} \sum_{i=1}^N \delta_{(Y_i, Y_{i+1})} \in \widetilde{\mathcal{M}}_1(E \times E), Y = (Y_1, \dots, Y_N), \text{ and } Y_{N+1} = Y_1. \quad (5.11)$$

We assume in the following that the stochastic transition matrix  $P$  has strictly positive entries, i.e.,  $p(x, y) > 0$  for all  $x, y \in E$ . We let the Markov chain started with initial distribution given by the stationary measure  $\mu$  and use the pair empirical measure to express the Markov chain probability as

$$\begin{aligned} P_\mu(Y_1 = y_1, \dots, Y_N = y_N) &= \mu_{y_1} p(y_1, y_2) \cdots p(y_{N-1}, y_N) \\ &= \frac{\mu_{y_1}}{p(y_{N+1}, y_1)} \exp \left( \sum_{i=1}^N \log p(y_i, y_{i+1}) \right) \\ &= \frac{\mu_{y_1}}{p(y_{N+1}, y_1)} \exp \left( N \sum_{x, y \in E} L_N^{2,Y}(x, y) \log p(x, y) \right), \end{aligned} \quad (5.12)$$

where we use the fact that  $N L_N^{2,Y}(x, y)$  is frequency of transitions  $x \rightarrow y$  of the Markov path. On the other hand we like to compare this probability with the probability given an i.i.d. sequence  $(X_k)_{k \in \mathbb{N}}$  of  $E$ -valued random variables  $X_k$  with law  $\mu \in \mathcal{M}_1(E)$ . We have

$$P_X(y_1, \dots, y_N) := \mathbb{P}(X_1 = y_1, \dots, X_N = y_N) = \prod_{i=1}^N \mu_{y_i} = \exp \left( N \sum_{x, y \in E} L_N^{2,X}(x, y) \log \mu_y \right). \quad (5.13)$$

We see that the Radon-Nikodym density of (5.13) with respect to (5.12) reads as

$$\frac{dP_\mu}{dP_X}(y_1, \dots, y_N) = \frac{\mu_{y_1}}{p(y_N, y_1)} \exp(NF(L_N^{2,y})), \quad y = (y_1, \dots, y_N), \quad (5.14)$$

with

$$F(\nu) = \sum_{x, y \in E} \nu_{x,y} \log \frac{p_{x,y}}{\mu_y}, \quad \nu \in \widetilde{\mathcal{M}}_1(E \times E).$$

Clearly,  $F$  is bounded and continuous.

**Theorem 5.6 (Pair empirical measure LDP for Markov chains)** *Suppose that  $(Y_k)_{k \in \mathbb{N}}$  is a finite state space  $E$  Markov with irreducible transition matrix  $P = (p(x, y))_{x, y \in E}$  with strictly positive entries  $p(x, y) > 0$  for all  $x, y \in E$  and unique stationary measure  $\mu \in \mathcal{M}_1(E)$  with  $\mu_x > 0$  for all  $x \in E$ . Denote  $P_\mu$  the Markov probability measure and define  $\mu_N(\cdot) := P_\mu(L^2 \in \cdot)$ . Then  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the large deviation principle on  $\widetilde{\mathcal{M}}_1(E \times E)$  with rate  $N$  and good rate function*

$$I_P^2(\nu) := \sum_{x, y \in E} \nu_{x,y} \log \frac{\nu_{x,y}}{\bar{\nu}_x p(x, y)}, \quad (5.15)$$

where  $\bar{\nu}_x = \sum_{y \in E} \nu_{x,y}$  is the marginal for  $\nu \in \widetilde{\mathcal{M}}_1(E \times E)$ .

**Proof.** It follows from the Radon-Nikodym density in (5.14) that, for every Borel set  $A \subset \widetilde{\mathcal{M}}_1(E \times E)$ ,

$$\frac{1}{N} \log \mu_N(A) = \left( \frac{1}{N} \right) + \frac{1}{N} \log \int_A e^{NF(\nu)} \nu_N(d\nu), \quad (5.16)$$

where  $\nu_N(\cdot) := P_X(L_N^2 \in \cdot)$  and where  $O(\frac{1}{N})$  accounts for the first factor in (5.14), i.e., all terms  $\frac{\mu_{y_1}}{p(y_N, y_1)}$  for which  $y_1, y_N$  are states in the event  $\Lambda_N^{2,Y} \in A$ .

We know from Theorem 2.12 that  $(\nu_N)_{N \in \mathbb{N}}$  satisfies the LDP on  $\widetilde{\mathcal{M}}_1(E \times E)$  with rate  $N$  and good rate function

$$I_\mu^2(\nu) = \sum_{x,y \in E} \nu_{x,y} \log \frac{\nu_{x,y}}{\bar{\nu}_x \mu_y}, \quad \nu \in \widetilde{\mathcal{M}}_1(E \times E).$$

Secondly, the integral on the right hand side of (5.16) has exactly the form of the tilted large deviation principle in Theorem 3.13 as the function  $F$  is bounded and continuous. Hence,  $(\mu_N)_{N \in \mathbb{N}}$  satisfies the LDP with rate function

$$I_\mu^2(\nu) = I^2, \mu(\nu) - F(\nu) = \sum_{x,y \in E} \nu_{x,y} \log \frac{\nu_{x,y}}{\bar{\nu}_x p(x, y)}.$$

□

**Remark 5.7** (a) Note that (5.15) says that

$$I_\mathbb{P}^2(\nu) = H(\nu | \bar{\nu} \otimes \mathbb{P}),$$

the relative entropy of  $\nu$  with respect to  $\bar{\nu} \otimes \mathbb{P}$ , defined by  $(\bar{\nu} \otimes \mathbb{P})_{x,y} = \bar{\nu}_x p(x, y)$  for all  $x, y \in E$ .

- (b) If the stochastic matrix  $P$  fails to have strictly positive entries in Theorem 5.6 but is irreducible, then Theorem 5.6 stills applies when  $E \times E$  is replaced by  $\{(x, y) \in E \times E : p(x, y) > 0\}$ . The proof can easily be adapted. Also note that it is not relevant that the Markov chain starts in  $\mu$ .

◇

As  $I_\mathbb{P}^2$  is given as a relative entropy, it has the following straightforward properties which we state without proof.

**Lemma 5.8** (a)  $I_\mathbb{P}^2$  is finite, continuous and strictly convex on  $\widetilde{\mathcal{M}}_1(E \times E)$ , except along line segments  $\{t\nu + (1-t)\nu' : t \in [0, 1]\}$  between  $\nu$  and  $\nu'$  satisfying

$$\frac{\nu_{x,y}}{\bar{\nu}_x} = \frac{\nu'_{x,y}}{\bar{\nu}'_x}, \quad \text{for all } x, y \in E.$$

Along such line segments  $I_\mathbb{P}^2$  is affine.

- (b)  $I_\mathbb{P}^2(\nu) \geq 0$  with equality if and only if  $\nu = \mu \otimes P$ .

**Proof.** Left as an exercise. □

Theorem 5.6 allows to deduce rate function and the LDP for the *empirical measure*  $L_N^Y$  via the contraction principle in Theorem 3.6. Clearly, the mapping

$$\mathcal{M}_1(E \times E) \ni \nu \mapsto \nu^{(2)} \in \mathcal{M}_1(E) \text{ with } \nu_y^{(2)} = \sum_{x \in E} \nu_{x,y},$$

is continuous. We thus get the following statement for the function  $J$  in (5.4) which equals the rate function  $I$  in (5.9) and governs the large deviation principle for the empirical measure  $L_N^Y$  in Theorem 5.3.

**Proposition 5.9** *Under the assumptions of Theorem 5.3 and Theorem 5.6, it holds that*

$$J(\nu) = \inf_{q \in \mathcal{M}_1(E \times E): q^{(2)} = \nu} \{I_P^2(q)\}, \quad \nu \in \mathcal{M}_1(E). \quad (5.17)$$

**Proof.** We first note that we do not need to prove (5.17) directly. The idea is to employ the contraction principle in Theorem 3.6 and the uniqueness of the rate function and the identification in Theorem 5.4. For any  $y \in E$  we

$$\sum_{x \in E} L_N^{2,Y}(x, y) = \sum_{x \in E} \frac{1}{N} \sum_{k=1}^N \delta_{(Y_k, Y_{k+1})}(x, y) = L_N^Y(y).$$

Furthermore, for any initial state  $\sigma \in E$  of the Markov chain  $(Y_k)_{k \in \mathbb{N}}$ , we observe that for  $A \subset \mathcal{M}_1(E)$  we have that  $L_N^Y \in A$  if and only if  $L_N^{2,Y} \in \{q \in \mathcal{M}_1(E \times E): q^{(2)} \in A\}$ . The right hand side of (5.17), according to the contraction principle in Theorem 3.6, is a rate function and governs the LDP for the empirical measure  $L_N^Y$ . As the rate function is unique, we obtain the equality with the left hand side of (5.17) via Theorem 5.4.  $\square$

Another observation is that the relation in Theorem 5.4 holds for any nonnegative irreducible matrix  $B = (b(x, y))_{x, y \in E}$  (not necessarily stochastic matrix).

**Exercise 5.10** (a) Show that the relation in Theorem 5.3 holds for any nonnegative irreducible matrix  $B = (b(x, y))_{x, y \in E}$  (not necessarily a stochastic matrix).

(b) Show that for any irreducible, nonnegative matrix  $B = (b(x, y))_{x, y \in E}$ ,

$$\log \varrho(B) = \sup_{\nu \in \mathcal{M}_1(E)} \{-J_B(\nu)\},$$

where  $J_B$  is the function in Theorem 5.3 for the matrix  $B$ .



**Solution.**

(a) Define  $\varphi(x) = \sum_{y \in E} b(x, y)$  for all  $x \in E$ . Clearly,  $\varphi \gg 0$ , and the matrix  $P$  with entries  $p(x, y) = b(x, y)/\varphi(x)$  is a stochastic matrix. Now define  $J_B$  as in Theorem 5.3 for  $B$  and  $I_B$  as in (5.9) for  $B$ . We denote the corresponding functions for the stochastic matrix  $P$  by  $J_P$  and  $I_P$ , respectively. We compute for  $q \in \mathcal{M}_1(E)$ ,

$$\begin{aligned} I_P(q) &= \sup_{u \gg 0} \left\{ \sum_{x \in E} q_x \log \frac{u_x}{(uP)_x} \right\} = \sup_{u \gg 0} \left\{ \sum_{x \in E} q_x \log \frac{u_x}{\left( \sum_{y \in E} u_y b(y, x) \varphi^{-1}(x) u_x \right)} \right\} \\ &= \sup_{u \gg 0} \left\{ \sum_{x \in E} q_x \log \frac{u_x \varphi(x)}{(uB)_x} \right\} = J_B(q) + \sum_{x \in E} q_x \log \varphi(x), \end{aligned}$$

and we get  $J_B(q) = \infty$  for  $q \notin \mathcal{M}_1(E)$  from the property of  $I_P$ . Likewise,

$$I_P(q) = I_B(q) + \sum_{x \in E} q_x \log \varphi(x).$$


(b) Choosing  $\lambda \equiv 0$ , we get  $I_B(q) \geq -\log \varrho(B)$ , and thus

$$-J_B(q) \leq \log \varrho(B).$$

The reversed inequality follows from the proof of Theorem 5.3.

**Exercise 5.11** Deduce by applying Exercise 5.10 (b) and (5.17) in Proposition 5.9 that for any nonnegative irreducible matrix  $B$ ,

$$-\log \varrho(B) = \inf_{q \in \mathcal{M}_1(E_B)} \{I_P^2(q)\}, \quad (5.18)$$

where  $E_B = \{(x, y) \in E \times E : b(x, y) > 0\}$ . 

**Exercise 5.12** Show that for any nonnegative irreducible matrix  $B = (B(x, y))_{x, y \in E}$ ,

$$J_B(q) = \begin{cases} \sup_{u \gg 0} \left\{ \sum_{x \in E} q_x \log \frac{u_x}{(Bu)_x} \right\} & \text{if } q \in \mathcal{M}_1(E), \\ \infty & \text{if } q \notin \mathcal{M}_1(E). \end{cases}$$

**Solution.** This follows from that fact that the eigenvalues for the left and right vector are equal. That is, the matrix  $B_\lambda$  and  $\tilde{B}_\lambda$  have the same eigenvalues, where  $B_\lambda(x, y) = b(x, y)e^{\lambda_y}$  and  $\tilde{B}_\lambda(x, y) = b(x, y)e^{\lambda_x}$  for all  $x, y \in E$ . We then conclude with Theorem 5.3.

### 5.3 Markov process with continuous time and finite state space

## 6 The Gibbs Conditioning principle

We consider dependency structures due to conditions and constraints. We give an example of the one-dimensional Ising model to demonstrate that constraints lead to functionals of physical relevance. In this section we consider finite state spaces  $E$ .

### 6.1 Conditional limit theorem for i.i.d. sequences

Let  $(Y_i)_{i \in \mathbb{N}}$  be a sequence of i.i.d.  $E$ -valued random variables with law  $\mu \in \mathcal{M}_1(E)$  such that  $\mu_x > 0$  for all  $x \in E$ , and define  $X_k = f(Y_k)$  for some deterministic function  $f: E \rightarrow \mathbb{R}$ . We are interested in the fundamental question of statistical mechanics. Given some Borel set  $A \subset \mathbb{R}$  and a constraint of the type  $S_N \in A$ ,  $S_N = \frac{1}{N} \sum_{k=1}^N X_k$ , what is the conditional law<sup>3</sup> of  $Y_1$  when  $N$  is large? We want to know the limit points (accumulation points), as  $N \rightarrow \infty$ , of the conditional probability  $\mu_N^* \in \mathcal{M}_1(E)$  defined as

$$\mu_N^*(x) = \mathbb{P}(Y_1 = x | S_N \in A), \quad x \in E. \quad (6.1)$$

In the following we write  $\mathbf{f} = (f(x))_{x \in E}$  and thus have

$$S_N = \langle \mathbf{f}, L_N^Y \rangle, \quad Y = (Y_1, \dots, Y_N), \quad L_N^Y = \frac{1}{N} \sum_{k=1}^N \delta_{Y_k}.$$

Under the conditioning  $S_N \in A$ , the random variables  $Y_i$  are no longer independent but still identically distributed. Therefore, for every function  $g: E \rightarrow \mathbb{R}$ ,

$$\begin{aligned} \langle g, \mu_N^* \rangle &= \mathbb{E}[g(Y_1) | S_N \in A] = \mathbb{E}[g(Y_2) | S_N \in A] = \mathbb{E}\left[\frac{1}{N} \sum_{k=1}^N g(Y_k) \middle| S_N \in A\right] \\ &= \mathbb{E}[\langle g, L_N^Y \rangle | \langle \mathbf{f}, L_N^Y \rangle]. \end{aligned}$$

With  $\Gamma := \{\nu \in \mathcal{M}_1(E) : \langle \mathbf{f}, \nu \rangle \in A\}$  we write

$$\mu_N^* = \mathbb{E}[L_N^Y | L_N^Y \in \Gamma]. \quad (6.2)$$

With this rewriting, the following characterisation of the limit points of  $(\mu_N^*)_{N \in \mathbb{N}}$  applies to any non-empty set  $\Gamma \subset \mathcal{M}_1(E)$  for which

$$I_\Gamma := \inf_{\nu \in \Gamma} \{H(\nu | \mu)\} = \inf_{\nu \in \bar{\Gamma}} \{H(\nu | \mu)\}. \quad (6.3)$$

**Theorem 6.1 (Gibbs's principle)** *For a given set  $\Gamma \subset \mathcal{M}_1(E)$  satisfying (6.3), define*

$$\mathcal{M} = \{\nu \in \bar{\Gamma} : H(\nu | \mu) = I_\Gamma\}. \quad (6.4)$$

- (a) *All the limit (accumulation) points of  $(\mu_N^*)_{N \in \mathbb{N}}$  belong to  $\overline{\text{co}}(\mathcal{M})$ , the closure of the convex hull of  $\mathcal{M}$ .*
- (b) *When  $\Gamma \subset \mathcal{M}_1(E)$  is a convex set of non-empty interior, the set  $\mathcal{M}$  consists of a single point to which  $\mu_N^*$  converges as  $N \rightarrow \infty$ .*

**Proof of Theorem 6.1.** (a) As  $E$  is a finite set, the set  $\mathcal{M}_1(E)$  is a compact set, which can be identified with the simplex  $\{\nu \in [0, 1]^E : \sum_{x \in E} \nu_x = 1\} \subset [0, 1]^E$ . Thus the

closure  $\bar{\Gamma} \subset \mathcal{M}_1(E)$  is a compact set. For every set  $U \subset \mathcal{M}_1(E)$  we shall estimate the difference of the conditional expectations

$$\begin{aligned} \mathbb{E}[L_N^Y | L_N^Y \in \Gamma] - \mathbb{E}[L_N^Y | L_N^Y \in U \cap \Gamma] &= \mathbb{P}(L_N^Y \in U^c | L_N^Y \in \Gamma) \left( \mathbb{E}[L_N^Y | L_N^Y \in U^c \cap \Gamma] \right. \\ &\quad \left. - \mathbb{E}[L_N^Y | L_N^Y \in U \cap \Gamma] \right). \end{aligned}$$

The condition  $L_N^Y \in U \cap \Gamma$  ensures that  $\mathbb{E}[L_N^Y | L_N^Y \in U \cap \Gamma]$  belongs to  $c(U)$ , while  $\mu_N^* = \mathbb{E}[L_N^Y | L_N^Y \in \Gamma]$ . Thus we can estimate the distance of  $\mu_N^*$  to the convex hull  $c(U)$  as follows, using  $d$  as the total variation metric as well as the distance with respect to this metric of elements on sets of elements,

$$\begin{aligned} d(\mu_N^*, c(U)) &\leq \mathbb{P}(L_N^Y \in U^c | L_N^Y \in \Gamma) d\left(\mathbb{E}[L_N^Y | L_N^Y \in U^c \cap \Gamma], \mathbb{E}[L_N^Y | L_N^Y \in U \cap \Gamma]\right) \\ &\leq \mathbb{P}(L_N^Y \in U^c | L_N^Y \in \Gamma) \end{aligned} \quad (6.5)$$

where the last inequality follows due to the fact that  $d(\cdot, \cdot) \leq 1$ . We define a  $\delta$ -neighbourhood of the set  $\mathcal{M}$ ,

$$\mathcal{M}^\delta := \{\nu \in \mathcal{M} : d(\nu, \mathcal{M}) < \delta\}$$

and we show below that the following holds for all  $\delta > 0$ ,

$$\lim_{N \rightarrow \infty} \mathbb{P}\left(L_N^Y \in \mathcal{M}^\delta \mid L_N^Y \in \Gamma\right) = 1, \quad (6.6)$$

with an exponential rate of convergence. Consequently, (6.5) applied to  $U = \mathcal{M}^\delta$  results in

$$d(\mu_N^*, c(\mathcal{M}^\delta)) \rightarrow 0 \text{ as } N \rightarrow \infty.$$

We conclude now by observing that each point in  $c(\mathcal{M}^\delta)$  is within variational distance  $\delta$  of some point in the convex hull  $c(\mathcal{M})$ . This follows easily from the convexity of the variational distance  $d$  as a mapping on  $\mathcal{M}_1(E) \times \mathcal{M}_1(E)$ . Now, as  $\delta$  is arbitrarily small, limit points of  $(\mu_N^*)_{N \in \mathbb{N}}$  are necessarily in the closure of the convex hull  $c(\mathcal{M})$ .

Proof of (6.6): We now prove (6.6) using large deviation principle and methods. We observe that (6.3) ensures that  $\Gamma$  is an  $I_\mu$ -continuity set of Sanov's theorem, see Theorem 2.7 where  $I_\mu(\nu) = H(\nu|\mu)$ . Thus we get

$$I_\Gamma = - \lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(L_N^Y \in \Gamma) \quad (6.7)$$

and

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(L_N^Y \in (\mathcal{M}^\delta)^c \cap \Gamma) &\leq - \inf_{\nu \in (\mathcal{M}^\delta)^c \cap \Gamma} \{H(\nu|\mu)\} \\ &\leq - \inf_{\nu \in (\mathcal{M}^\delta)^c \cap \bar{\Gamma}} \{H(\nu|\mu)\}. \end{aligned} \quad (6.8)$$

The sets  $(\mathcal{M}^\delta)^c \cap \bar{\Gamma}$  are compact as  $\mathcal{M}^\delta$  are open sets. Thus the lower continuity (in fact in our setting the relative entropy is continuous) of the rate function ensures that the infimum over compact sets is attained, that is, for some  $\hat{\nu} \in (\mathcal{M}^\delta)^c \cap \bar{\Gamma}$ ,

$$\inf_{\nu \in (\mathcal{M}^\delta)^c \cap \bar{\Gamma}} \{H(\nu|\mu)\} = H(\hat{\nu}|\mu) > I_\Gamma. \quad (6.9)$$

Now, our statement (6.6) follows from (6.7), (6.8) and (6.9) because

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(L_N^Y \in (\mathcal{M}^\delta)^c | L_N^Y \in \Gamma) \\ &= \limsup_{N \rightarrow \infty} \left( \frac{1}{N} \log \mathbb{P}(L_N^Y \in (\mathcal{M}^\delta)^c \cap \Gamma) - \frac{1}{N} \log \mathbb{P}(L_N^Y \in \Gamma) \right) < 0. \end{aligned}$$

(b) We prove (b) in our TA class and refer to Example 6.2 below. Further details can be found in [DZ98], Chapter 3.  $\square$

**Example 6.2 (Example for Gibbs parameter)** In the setting of Cramér's theorem for finite subsets of  $\mathbb{R}$  in Section 2.3, Theorem 2.15,  $X_k = f(Y_k)$ ,  $f: E \rightarrow \mathbb{R}$ , and  $E$  finite state space. We define

$$K := [\min_{x \in E} \{f(x)\}, \max_{x \in E} \{f(x)\}]$$

and assume that  $\mathring{K} \neq \emptyset$ . For  $A \neq \emptyset$ , convex, open subset of  $K$  we have

$$S_N \in A \Leftrightarrow L_N^Y \in \{\nu \in \mathcal{M}_1(E) : \langle f, \nu \rangle \in A\} =: \Gamma,$$

and  $\Gamma$  is open when  $A$  is open. By Jensen's inequality,

$$\Lambda(\lambda) \geq \lambda \langle f, \nu \rangle - H(\nu|\mu),$$

with equality holding for  $\nu = \nu_\lambda$  defined by

$$\nu_\lambda(x) = \mu(x) e^{\lambda f(x) - \Lambda(\lambda)}, \quad x \in E.$$

Thus, for all  $\lambda$  and all  $x$ ,

$$\lambda x - \Lambda(\lambda) \leq \inf_{\nu \in \mathcal{M}_1(E) : \langle f, \nu \rangle = x} \{H(\nu|\mu)\} = I(x)$$

with equality holding when  $x = \langle f, \nu_\lambda \rangle$ . The unique limit of  $(\mu_N^*)_{N \in \mathbb{N}}$  is of the form  $\nu_\lambda$  with some chosen  $\lambda \in \mathbb{R}$ , which is called the *Gibbs parameter*. For any  $x \in \mathring{K}$ , the Gibbs parameter associated with the open set  $(x - \delta, x + \delta)$  converges, as  $\delta \rightarrow 0$ , to the unique solution of the equation  $\Lambda'(\lambda) = x$ .



## 6.2 Example: microcanonical ensemble for one-dimensional Ising model

We outline how conditional measures appear in mathematical statistical mechanics. For this we consider the so-called *Ising model*. The model is a specific distribution for families of  $E = \{-1, 1\}$ -valued random variables  $\sigma_x$ , indexed by the inter lattice, i.e.,  $x \in \mathbb{Z}^d$ . The random variable  $\sigma_x \in E$  is called the *spin* at  $x$ . We solely discuss  $d \equiv 1$ . The distribution of  $(\sigma_x)_{x \in \mathbb{Z}}$  is given in terms of so-called *finite-volume distributions* in finite subsets  $\Lambda \subset \mathbb{Z}, |\Lambda| < \infty$ . For given finite set  $\Lambda$  we define the distribution in  $\Lambda$  via an energy function, called *Hamilton function*, which models nearest neighbour interactions in the following way. Any nearest neighbour interaction model need to know values of the spin variables outside of  $\Lambda$ . Alternatively, we may consider so-called periodic boundary conditions. In the following let  $\Lambda_N = \{-N, \dots, 0, 1, \dots, N\} \subset \mathbb{Z}$  and define the Hamilton function for  $\Lambda_N$  for periodic boundary conditions as

$$H_{\Lambda_N}^{(\text{per})}(\sigma) = - \sum_{i=-N}^N \sigma_i \sigma_{i+1} - h \sum_{i \in \Lambda_N} \sigma_i, \quad \sigma_{N+1} = \sigma_{-N}; h \in \mathbb{R}; \sigma \in E^{\mathbb{Z}}. \quad (6.10)$$

Here, the parameter  $h \in \mathbb{R}$  describes an external magnetic field. If we like to consider arbitrary boundary conditions with set  $\sigma \equiv \eta$  outside of  $\Lambda_N$  for a given configuration  $\eta \in E^{\mathbb{Z}}$ . Then the Hamilton function in  $\Lambda_N$  with boundary condition  $\eta \in E^{\mathbb{Z}}$  is defined as

$$H_{\Lambda_N}^{\eta}(\sigma) = - \sum_{i=-N}^{N-1} \sigma_i \sigma_{i+1} - \sigma_N \eta_{N+1} - \eta_{-N-1} \sigma_{-N} - h \sum_{i \in \Lambda_N} \sigma_i. \quad (6.11)$$

The uniform distribution on the state space or spin  $E$  is  $\lambda = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1} \in \mathcal{M}_1(E)$ . The model is then given by the Gibbs distribution  $\gamma_{\Lambda_N}^{\eta} \in \mathcal{M}_1(E^{\Lambda_N})$  in  $\Lambda_N$  with boundary condition  $\eta \in E^{\mathbb{Z}}$  and inverse temperature  $\beta \in (0, \infty)$ , defined as

$$\gamma_{\Lambda_N}^{\eta}(\mathrm{d}\sigma) = \frac{1}{Z_{\Lambda_N}(\beta, \eta)} e^{-\beta H_{\Lambda_N}^{\eta}(\sigma)} \lambda^{\otimes \Lambda_N}(\mathrm{d}\sigma),$$

with the normalisation, also called *partition function*,

$$Z_{\Lambda_N}(\beta, \eta) = \int_{E^{\Lambda_N}} e^{-\beta H_{\Lambda_N}^{\eta}(\sigma)} \lambda^{\otimes \Lambda_N}(\mathrm{d}\sigma).$$

We take a different root here by defining certain type classes as we did for Sanov's theorem. For this it is convenient to switch to the so-called *lattice gas* setting with state space  $\tilde{E} = \{0, 1\}$  and configurations

$$\omega_i = (\sigma_i + 1)/2 \quad \sigma_i = 2\omega_i - 1, \quad i \in \mathbb{Z}, \omega \in \tilde{E}^{\mathbb{Z}}, \sigma \in E^{\mathbb{Z}}.$$

When  $\omega_i = 1$  we say that there is a particle at  $i$ , when  $\omega_i = 0$  then site  $i$  is vacant (empty) with no particle around. The Hamilton function for this *lattice gas* version of the Ising model is

$$H_{\Lambda_N}^{\eta}(\omega) = - \sum_{i=-N}^{N-1} \omega_i \omega_{i+1} - \omega_N \eta_{N+1} - \eta_{-N-1} \omega_{-N} - \mu \sum_{i \in \Lambda_N} \omega_i, \quad \eta \in \tilde{E}^{\mathbb{Z}}, \mu \in \mathbb{R}. \quad (6.12)$$

We condition on lattice gas configurations  $\omega \in \tilde{E}^{\Lambda_N}$  with given average energy and average particle number. This is motivated by our studies of the Gibbs principle and the methods of types when proving Sanov's theorem. Suppose that  $(\varepsilon_N)_{N \in \mathbb{N}}$  and  $(\varrho_N)_{N \in \mathbb{N}}$  are sequences of positive real numbers such that  $\varepsilon_N |\Lambda_N| \in \mathbb{N}_0$ ,  $\varrho_N |\Lambda_N| \in \mathbb{N}_0$  for all but finitely many  $N \in \mathbb{N}$  and that  $\varepsilon_N \rightarrow \varepsilon$  and  $\varrho_N \rightarrow \varrho$  as  $N \rightarrow \infty$  with  $0 < \varepsilon < \varrho < 1$  and  $1 - 2\varrho + \varepsilon > 0$ . We write  $N_{\Lambda_N}(\omega) = \sum_{i \in \Lambda_N} \omega_i$  for the number of particles for the configuration  $\omega \in \tilde{E}^{\Lambda_N}$ . The set

$$\{\omega \in \tilde{E}^{\Lambda_N} : H_{\Lambda_N}^\eta(\omega) = \varepsilon_N |\Lambda_N|; N_{\Lambda_N}(\omega) = \varrho_N |\Lambda_N|\}$$

contains all configuration with average energy  $\varepsilon_N$  and average particle density  $\varrho_N$ . The *microcanonical entropy* for energy density  $\varepsilon_N$  and particle density  $\varrho_N$  if the cardinality of that set and is denoted

$$Z_{\Lambda_N, \varepsilon_N, \varrho_N}^\eta := |\{\omega \in \tilde{E}^{\Lambda_N} : H_{\Lambda_N}^\eta(\omega) = \varepsilon_N |\Lambda_N|; N_{\Lambda_N}(\omega) = \varrho_N |\Lambda_N|\}|.$$

Suppose we consider  $\eta \equiv 0$  on  $\Lambda_N^c$ , then one has an explicit formula for the microcanonical entropy, see [Ada01], namely,

$$Z_{\Lambda_N, \varepsilon_N, \varrho_N}^0 = \binom{\varrho_N |\Lambda_N| - 1}{\varepsilon_N |\Lambda_N|} \binom{|\Lambda_N| - \varrho_N |\Lambda_N| + 1}{\varrho_N |\Lambda_N| - \varepsilon_N |\Lambda_N|}.$$

Furthermore, [Ada01] provide a large deviation proof and analysis of the following statement.

$$\lim_{N \rightarrow \infty} \frac{1}{|\Lambda_N|} \log Z_{\Lambda_N, \varepsilon_N, \varrho_N}^0 = \log \left( \frac{\varrho^\varrho (1 - \varrho)^{1-\varrho}}{\varepsilon^\varepsilon (\varrho - \varepsilon)^{2(\varrho - \varepsilon)} (1 - 2\varrho + \varepsilon)^{1-2\varrho + \varepsilon}} \right) =: s(\varepsilon, \varrho),$$

and the function  $s(\varepsilon, \varrho)$  is called the specific or limiting microcanonical entropy for energy density  $\varepsilon$  and particle density  $\varrho$ . In mathematical statistical mechanics one can show that in thermodynamic equilibrium the inverse temperature and the chemical potential  $\mu = \mu(\varepsilon, \varrho)$  are given as follows, see [Ada01],

$$\begin{aligned} \beta(\varepsilon, \varrho) &= \frac{s}{\partial \varepsilon}(\varepsilon, \varrho) = \log \frac{(\varrho - \varepsilon)^2}{\varepsilon(1 - 2\varrho + \varepsilon)}, \\ \mu(\varepsilon, \varrho) &= \frac{\partial s}{\partial \varrho}(\varepsilon, \varrho) = \log \frac{\varrho(1 - 2\varrho + \varepsilon)^2}{(1 - \varrho)(\varrho - \varepsilon)^2}. \end{aligned}$$

## 7 Sample path large deviations

### 7.1 Mogulskii's theorem

Let  $(X_i)_{i \in \mathbb{N}}$  be an i.i.d. sequence of  $\mathbb{R}^d$ -valued random vectors with law  $\mu \in \mathcal{M}_1(\mathbb{R}^d)$  such that

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}] = \log \int_{\mathbb{R}^d} e^{\lambda x} \mu(dx) < \infty \quad \text{for all } \lambda \in \mathbb{R}^d. \quad (7.1)$$

We shall study large deviations behaviour for random path  $Z_N$ , i.e., families of  $\mathbb{R}^d$ -valued random vector indexed by  $t \in [0, 1]$ ,

$$Z_N(t) = \frac{1}{N} \sum_{k=1}^{\lfloor Nt \rfloor} X_k, \quad 0 \leq t \leq 1. \quad (7.2)$$

Let  $\mu_N$  be the law of the *empirical path*  $Z_N$  in  $L_\infty([0, 1])$ , that is,  $\mu_N = \mu^{\otimes N} \circ Z_N^{-1}$  and  $Z_N = Z_N(X_1, \dots, X_N): [0, 1] \rightarrow \mathbb{R}^d$ . In the following we use the *supremum norm*  $\|\cdot\|$  on  $L_\infty([0, 1])$ ,

$$\|f\| := \sup_{t \in [0, 1]} \{|f(t)|\}, \quad f \in L_\infty([0, 1]), \quad (7.3)$$

and write, throughout,  $|x| = \sqrt{x_1^2 + \dots + x_d^2}$ ,  $x \in \mathbb{R}^d$ , for the Euclidean norm on  $\mathbb{R}^d$ . Recall the definition of the Legendre-Fenchel transform in 1.33,

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

**Theorem 7.1 (Sample path large deviations for random walks, Mogulskii's theorem)** *The sequence  $(\mu_N)_{N \in \mathbb{N}}$  of probability measures  $\mu_N$  satisfies in  $L_\infty([0, 1])$  the LDP with the good rate function*

$$I(h) = \begin{cases} \int_0^1 \Lambda^*(\dot{h}(t)) dt & , \text{ if } h \in \mathcal{AC}, h(0) = 0, \\ \infty & , \text{ otherwise,} \end{cases} \quad (7.4)$$

where  $\mathcal{AC}$  denotes the space of absolutely continuous functions, i.e.,

$$\begin{aligned} \mathcal{AC} = \left\{ h \in \mathcal{C}([0, 1]): \sum_{\ell=1}^k |t_\ell - s_\ell| \rightarrow 0 \text{ as } k \rightarrow \infty, \text{ for } t_\ell, s_\ell \in [0, 1] \text{ with} \right. \\ \left. s_\ell < t_\ell < s_{\ell+1} < t_{\ell+1} \dots \Rightarrow \sum_{\ell=1}^k |h(t_\ell) - h(s_\ell)| \rightarrow 0 \right\}. \end{aligned} \quad (7.5)$$

**Remark 7.2** (a)  $h: [0, 1] \rightarrow \mathbb{R}^d$  absolutely continuous implies that  $h$  is differentiable almost everywhere; in particular, that it is the integral of an  $L_1([0, 1])$  function,  $h(t) = \int_0^t f(s) ds$ ,  $f \in L_1([0, 1])$ . It hold sthat  $\mathcal{AC} = H_1([0, 1])$ .

(b) The measures  $\mu_N$  are supported on the space of functions continuous from the right and having left limits which contains the domain

$$\mathcal{D}_I = \{h \in \mathcal{AC}: h(0) = 0\}.$$

The LDP certainly holds on that (bigger) space, see Lemma 3.5, as well equipped with the supremum norm topology.

Before we begin our proof of Theorem 7.1 we need a new concept in large deviation theory. Namely, suppose that a LDP holds for some sequence which is in some way close, or an approximation of another sequence of measures, then the very same LDP holds for the other sequence. Such sequences of measures are called *exponentially equivalent*.

**Definition 7.3 (Exponential Equivalence)** Suppose  $(E, d)$  is a Polish space. The sequences  $(\mu_N)_{N \in \mathbb{N}}$  and  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  of probability measure on  $E$  are called *exponentially equivalent* if there exist probability spaces  $(\Omega, \mathcal{B}_N, P_N)$  and two sequences  $(X_N)_{N \in \mathbb{N}}, (\tilde{X}_N)_{N \in \mathbb{N}}$  of  $E$ -valued random variables with joint law  $P_N$  and marginals  $\mu_N$  and  $\tilde{\mu}_N$ , respectively, such that the following holds. For each  $\delta > 0$  the set  $\{\omega \in \Omega : (\tilde{X}(\omega), X(\omega)) \in \Gamma_\delta\} \in \mathcal{B}_N$ , and

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N(\Gamma_\delta) = -\infty, \quad (7.6)$$

where

$$\Gamma_\delta = \{(\tilde{x}, x) \in E \times E : d(\tilde{x}, x) > \delta\} \subset E \times E.$$

We cite without proof the following crucial statement that exponentially equivalent sequences share the same LDP.

**Theorem 7.4 (LDP for exponentially equivalent sequences)** Suppose  $(E, d)$  is a Polish space and let  $(\mu_N)_{N \in \mathbb{N}}$  and  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  sequences of probability measure on  $E$ . If an LDP with a good rate function  $I$  holds for the probability measures  $(\mu_N)_{N \in \mathbb{N}}$ , which are exponentially equivalent to the sequence  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$ , then the same LDP holds for  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$ .

**Lemma 7.5 (Empirical profile as linear interpolation)** Let  $\tilde{\mu}_N$  denote the law of the empirical profile  $\tilde{Z}_N$  defined via linear interpolation,

$$\tilde{Z}_N(t) = Z_N(t) + \left(t - \frac{\lfloor Nt \rfloor}{N}\right) X_{\lfloor Nt \rfloor + 1}, \quad 0 \leq t \leq 1. \quad (7.7)$$

Then the sequences  $(\mu_N)_{N \in \mathbb{N}}$  and  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  of probability measures are exponentially equivalent in  $L_\infty([0, 1])$ .

**Proof.** For every  $\delta > 0$  the sets  $\Gamma_\delta = \{\|\tilde{Z}_N - Z_N\| > \delta\}$  are measurable. From its definition we easily have the estimate

$$|\tilde{Z}_N(t) - Z_N(t)| \leq \frac{1}{N} |X_{\lfloor Nt \rfloor + 1}|.$$

For every  $\delta > 0$  and any  $\lambda > 0$ , we get by exponential Chebycheff inequality that

$$\mathbb{P}(\|\tilde{Z}_N - Z_N\| > \delta) \leq N \mathbb{P}(|X_1| > N\delta) \leq N \mathbb{E}[e^{\lambda |X_1|}] e^{-\lambda N \delta},$$

where the first inequality follows from the fact that the empirical profile is a sum of i.i.d. random variables. Since  $\mathcal{D}_\Lambda = \mathbb{R}^d$  according to (7.1), it follows, by considering first  $N \rightarrow \infty$  and then  $\lambda \rightarrow \infty$ , that for any  $\delta > 0$ ,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}(\|\tilde{Z}_N - Z_N\| > \delta) = -\infty,$$

and thus the statement.  $\square$

From Theorem 7.4 and Lemma 7.5 we know that it suffices to prove the LDP for the empirical profile  $\tilde{Z}_N$  instead of  $Z_N$ . The next result for the empirical profile is the crucial step in proving Theorem 7.1.

**Theorem 7.6 (LDP for the empirical profile)** *Let  $X := \{h: [0, 1] \rightarrow \mathbb{R}^d: h(0) = 0\}$ , and equip  $X$  with the topology of pointwise convergence on  $[0, 1]$ . Then the sequence of probability measures  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  of Lemma 7.5 satisfies the LDP in this Hausdorff topological space with the good rate function  $I$  of Theorem 7.1.*

Before proving this major step we state our last ingredient for the proof of Theorem 7.1, which is the exponential tightness. This allows us to prove the LDP in Theorem 7.1 immediately, and we later address the proof of the major step, Theorem 7.14.

**Lemma 7.7 (Exponential tightness)** *The probability measures  $\tilde{\mu}_N$  of Lemma 7.5 are exponentially tight in the space  $\mathcal{C}_0 = \{h \in \mathcal{C}([0, 1]): h(0) = 0\}$  equipped with the supremum norm topology.*

**Proof of Lemma 7.7.** The proof is technical and can be found in [DZ98].  $\square$

The following lemma shows an LDP for any finite collection of distinct times in  $[0, 1]$ . This result is key for the proof of Theorem 7.14.

**Lemma 7.8** *Let  $\mathcal{J}$  denote the collection of all ordered finite subsets of  $(0, 1]$ . For any  $j = \{0 < t_1 < t_2 < \dots < t_{|j|} \leq 1\} \in \mathcal{J}$  and any function  $f: [0, 1] \rightarrow \mathbb{R}^d$ , let*

$$p_j: \mathbf{X} \rightarrow \mathbb{R}^{d|j|}, f \mapsto p_j(f) = (f(t_1), \dots, f(t_{|j|}))$$

*be the projected vector in the finite-dimensional space  $\mathbb{R}^{d|j|}$ , where  $\mathbf{X} = \{f: [0, 1] \rightarrow \mathbb{R}^d\}$ . Then the sequence  $(\mu_N \circ p_j^{-1})_{N \in \mathbb{N}}$  of laws satisfies the LDP in  $\mathbb{R}^{d|j|}$  with the good rate function*

$$I_j(\mathbf{z}) = \sum_{\ell=1}^{|j|} (t_\ell - t_{\ell-1}) \Lambda^* \left( \frac{z_\ell - z_{\ell-1}}{t_\ell - t_{\ell-1}} \right), \quad \mathbf{z} = (z_1, \dots, z_{|j|}) \in \mathbb{R}^{d|j|}, t_0 = 0, z_0 = 0. \quad (7.8)$$

**Proof of Lemma 7.8.** Pick  $j \in \mathcal{J}$ . Then  $\mu_N \circ p_j^{-1}$  is the law of the random vector  $Z_N^j = (Z_N(t_1), \dots, Z_N(t_{|j|})) \in \mathbb{R}^{d|j|}$ . The key idea is to map this random vector to its vector of increments which happens to be a continuous bijective mapping. The vector of increments is

$$Y_N^j = (Y_{N,1}^j, \dots, Y_{N,|j|}^j) = (Z_N(t_1), Z_N(t_2) - Z_N(t_1), \dots, Z_N(t_{|j|}) - Z_N(t_{|j|-1})),$$

and  $T_j: \mathbb{R}^{d|j|} \rightarrow \mathbb{R}^{d|j|}$  defined by  $T_j(Y_N^j) = Z_N^j$  is continuous and one-to-one. Thus it suffices to prove the LDP for the increments (see Contraction Principle in Theorem 3.6). The entries of the increments vector  $Y_N^j$  are independent because  $(X_i)_{i \in \mathbb{N}}$  is an i.i.d. sequence of  $\mathbb{R}^d$ -valued random vectors. For  $\lambda = (\lambda_1, \dots, \lambda_{|j|}) \in \mathbb{R}^{d|j|}$  we get the limiting logarithmic moment generating function

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E}[e^{N \langle \lambda, Y_N^j \rangle}] &= \lim_{N \rightarrow \infty} \frac{1}{N} \log \prod_{\ell=1}^{|j|} \mathbb{E}[e^{N \langle \lambda_\ell, Y_{N,\ell}^j \rangle}] \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \log \prod_{\ell=1}^{|j|} \prod_{k=\lfloor Nt_{\ell-1} \rfloor}^{\lfloor Nt_\ell \rfloor} \mathbb{E}[e^{\langle \lambda_\ell, X_k \rangle}] \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\ell=1}^{|j|} \frac{1}{N} (\lfloor Nt_\ell \rfloor - \lfloor Nt_{\ell-1} \rfloor) \Lambda(\lambda_\ell) = \sum_{\ell=1}^{|j|} (t_\ell - t_{\ell-1}) \Lambda(\lambda_\ell) \\ &=: \Lambda_j(\lambda). \end{aligned}$$

From our assumptions (7.1) we see that  $\Lambda_j$  is finite and differentiable with domain  $\mathcal{D}_{\Lambda_j} = \mathbb{R}^{d|j|}$ . Thus, by the Gärtner-Ellis theorem, see Theorem 4.2, the LDP for  $Y_N^j$  in  $\mathbb{R}^{d|j|}$  follows with the good rate function

$$\begin{aligned} \Lambda_j^*(y) &= \sup_{\lambda \in \mathbb{R}^{d|j|}} \{ \langle \lambda, y \rangle - \Lambda_j(\lambda) \} \\ &= \sup_{\lambda \in \mathbb{R}^{d|j|}} \left\{ \sum_{\ell=1}^{|j|} \langle \lambda_\ell, y_\ell \rangle - (t_\ell - t_{\ell-1}) \Lambda(\lambda_\ell) \right\} \\ &= \sum_{\ell=1}^{|j|} (t_\ell - t_{\ell-1}) \sup_{\lambda_\ell \in \mathbb{R}^d} \left\{ \left\langle \lambda_\ell, \frac{y_\ell}{(t_\ell - t_{\ell-1})} \right\rangle - \Lambda(\lambda_\ell) \right\} \\ &= \sum_{\ell=1}^{|j|} (t_\ell - t_{\ell-1}) \Lambda^* \left( \frac{y_\ell}{(t_\ell - t_{\ell-1})} \right), \quad y \in \mathbb{R}^{d|j|}. \end{aligned} \tag{7.9}$$

□

**Proof of Theorem 7.14. Step 1:** From Lemma 7.8 we have that  $(\tilde{\mu}_N \circ p_j^{-1})_{N \in \mathbb{N}}$  satisfies the LDP in  $\mathbb{R}^{d|j|}$  with the good rate function  $I_j$ , as  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  and  $(\mu_N)_{N \in \mathbb{N}}$  are exponentially equivalent according to Lemma 7.5. This holds for every  $j \in \mathcal{J}$ .

**Step 2:** In this step we use the partial order and the projective limit method. We start by defining the partial order in our setting.

- For  $i, j \in \mathcal{J}$ ,  $i = \{s_1, \dots, s_{|i|}\}$  and  $j = \{t_1, \dots, t_{|j|}\}$ :

$$i \leq j \Leftrightarrow \forall \ell \exists q(\ell) \text{ such that } s_\ell = t_{q(\ell)}.$$

- For  $i \leq j$ ,  $i, j \in \mathcal{J}$ , the projection  $p_{ij}: \mathbb{R}^{d|j|} \rightarrow \mathbb{R}^{d|i|}$  is continuous.

- We define  $Y_j := \mathbb{R}^{d|j|}$ ,  $j \in \mathcal{J}$ , then the projective system  $(Y_j, p_{ij})_{i \leq j \in \mathcal{J}}$  consists of Hausdorff topological spaces  $(Y_j)_{j \in \mathcal{J}}$  and continuous maps  $p_{ij}: Y_j \rightarrow Y_i$  such that  $p_{ik} = p_{ij} \circ p_{jk}$  whenever  $i \leq j \leq k$  and  $p_{jj} = \text{id}$ .
- The projective limit of the projective system is denoted

$$\widetilde{X} = \varprojlim Y_j \subset Y = \prod_{j \in \mathcal{J}} Y_j,$$

consisting of all elements  $x = (y_j)_{j \in \mathcal{J}}$  for which  $y_i = p_{ij}(y_j)$  whenever  $i \leq j$ , equipped with the topology induced by  $Y$ .

- Identification of  $\widetilde{X}$  with  $X$  in Theorem 7.14: Each function  $f \in X$  corresponds to  $(p_j(f))_{j \in \mathcal{J}} \in \widetilde{X}$  since  $p_i(f) = p_{ij}(p_j(f))$  for  $i \leq j \in \mathcal{J}$ . Now each point  $x = (x_j)_{j \in \mathcal{J}} \in \widetilde{X}$  may be identified with  $f: [0, 1] \rightarrow \mathbb{R}^d$  by putting  $f(t) = x_{\{t\}}$  for  $t > 0$  and  $f(0) = 0$ . The topology on  $\widetilde{X}$  is given by the collection  $\{p_j^{-1}(\mathcal{U}_j): \mathcal{U}_j \subset Y_j \text{ open}\}$ , and henceforth it coincides with the topology of pointwise convergence.

We are now in the position to actually prove the required LDP by invoking the so-called Dawson-Gärtner Theorem (see [DZ98], Theorem 4.6.1) for projective limit systems as above. This theorem in conjunction with Theorem 7.14 and our projective system above provides the LDP for  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  in  $X$  with the good rate function

$$I_X(f) = \sup_{0=t_0 < t_1 < \dots < t_k \leq 1, k \in \mathbb{N}} \left\{ \sum_{\ell=1}^k (t_\ell - t_{\ell-1}) \Lambda^* \left( \frac{f(t_\ell) - f(t_{\ell-1})}{t_\ell - t_{\ell-1}} \right) \right\}. \quad (7.10)$$

**Step 3:** Identification  $I_X \equiv I$  with  $I$  from Theorem 7.14 respectively Theorem 7.1. Recall that  $\Lambda^*$  is convex, and thus by Jensen's inequality (with respect to uniform measure on  $[t_{\ell-1}, t_\ell]$ ), for all  $\ell = 1, \dots, k$ ,

$$\frac{1}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} \Lambda^*(\dot{f}(s)) \, ds \geq \Lambda^* \left( \frac{f(t_\ell) - f(t_{\ell-1})}{t_\ell - t_{\ell-1}} \right),$$

and thus we have  $I(f) \geq I_X(f)$ . For the opposite inequality, consider  $f' \text{ in } \mathcal{AC}$  and let  $g(t) := \frac{d}{dt} f(t)$ , then  $g \in L_1([0, 1])$ , and, for  $k \geq 1$ , define

$$g^k(t) = k \int_{[kt]/k}^{([kt]+1)/k} g(s) \, ds, \quad t \in [0, 1], \quad g^k(1) = k \int_{1-1/k}^1 g(s) \, ds.$$

Then

$$I_X(f) \geq \liminf_{k \rightarrow \infty} \sum_{\ell=1}^k \frac{1}{k} \Lambda^*(k(f(\ell/k) - f((\ell-1)/k))) = \liminf_{k \rightarrow \infty} \int_0^1 \Lambda^*(g^k(t)) \, dt.$$

By Lebesgue's theorem,  $\lim_{k \rightarrow \infty} g^k(t) = g(t)$  almost everywhere in  $[0, 1]$ . Hence, by Fatou's Lemma due to the fact that  $\Lambda^*$  is lower semicontinuous.

$$\liminf_{k \rightarrow \infty} \int_0^1 \Lambda^*(g^k(t)) dt \geq \int_0^1 \liminf_{k \rightarrow \infty} \Lambda^*(g^k(t)) dt \geq \int_0^1 \Lambda^*(g(t)) dt = I(f),$$

and hence  $I_X(f) \geq I(f)$ . We are left to show that  $I(f) = \infty$  for  $f \in \mathcal{X}$  but  $f \notin \mathcal{AC}$ . This is left as an exercise. □

We are finally in the position to prove Theorem 7.1 by combining our previous results.

**Proof of Theorem 7.1.** By Theorem 7.14,  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  satisfies the LDP in  $\mathcal{X}$ . Furthermore,  $\mathcal{D}_I \subset \mathcal{C}_0([0, 1])$ , and  $\tilde{\mu}_N(\mathcal{C}_0([0, 1])) = 1$  for all  $N \in \mathbb{N}$  as the empirical profile  $\tilde{Z}_N$  is continuous (actually even piece-wise differentiable) Thus, by Lemma 3.5, the LDP for  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  holds in  $\mathcal{C}_0([0, 1])$  when this space is equipped with the relative (Hausdorff) topology induced by all sets

$$V_{t,x,\delta} = \{g \in \mathcal{C}_0([0, 1]): |g(t) - x| < \delta\}, \quad t \in (0, 1], x \in \mathbb{R}^d, \delta > 0.$$

We observe that the sets  $V_{t,x,\delta}$  are open sets with respect to the supremum norm, and thus this topology on  $\mathcal{C}_0([0, 1])$  is finer (stronger) than the pointwise convergence topology. Now, Lemma 7.7 shows that  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  are exponentially tight with respect to the supremum norm topology. This in turn enables us to strengthen the LDP in  $\mathcal{C}_0([0, 1])$  to the supremum norm topology. This follows with an application of Proposition 3.8. Thus  $(\tilde{\mu}_N)_{N \in \mathbb{N}}$  satisfies the LDP in  $\mathcal{C}_0([0, 1])$  with respect to the supremum norm topology. Since  $\mathcal{C}_0([0, 1]) \subset L_\infty([0, 1])$  is a closed subset, the same LDP holds in  $L_\infty([0, 1])$  by using Lemma 3.5 again (now in the opposite direction). Finally, using Lemma 7.5 in conjunction with Theorem 7.4, we obtain the LDP for  $(\mu_N)_{N \in \mathbb{N}}$  in  $L_\infty([0, 1])$  with respect to the supremum norm topology. □

## 7.2 Schilder's theorem

Mogulskii's theorem Theorem 7.1 can be extended to the laws  $\nu_\varepsilon, \varepsilon > 0$ , of

$$Y_\varepsilon(t) = \varepsilon \sum_{k=1}^{\lfloor \frac{t}{\varepsilon} \rfloor} X_k, \quad 0 \leq t \leq 1. \quad (7.11)$$

Note that  $Z_N$  in Theorem 7.1 corresponds to the special case  $\varepsilon = N^{-1}$ .

**Theorem 7.9** *Assume all the assumptions of Theorem 7.1 above. Then the probability measures  $(\nu_\varepsilon)_{\varepsilon > 0}$  induced on  $L_\infty([0, 1])$  by  $Y_\varepsilon$  in (7.11) satisfy the LDP with rate  $\varepsilon^{-1}$  and with good rate function  $I$  given in (7.4).*

**Proof.** Example Sheet 4. □

Let  $(B(t))_{t \in [0,1]}$  denote standard Brownian motion in  $\mathbb{R}^d$  with *time horizon*  $[0, 1]$  and initial condition  $B(0) = 0$ . For any  $\varepsilon \geq 0$ , define  $B_\varepsilon$  by  $B_\varepsilon(t) = \sqrt{\varepsilon}B(t)$ ,  $t \in [0, 1]$ , and let  $\nu_\varepsilon$  denote the probability measure induced by the random path  $B_\varepsilon$ ,  $B_\varepsilon: [0, 1] \rightarrow \mathbb{R}^d$ , on the space  $\mathcal{C}_0([0, 1])$ , the space of continuous functions  $h: [0, 1] \rightarrow \mathbb{R}^d$  with  $h(0) = 0$ , equipped with the supremum norm topology.

**Question:** Is the process  $B_\varepsilon$  a candidate for an LDP similar to one developed for  $Y_\varepsilon$  in Theorem 7.9 ? Indeed,  $\|B_\varepsilon\| \rightarrow 0$  in probability as  $\varepsilon \downarrow 0$  (actually, almost surely) and exponentially fast in  $1/\varepsilon$  as implied by Lemma 7.10 below.

To formulate our LDP we need some notations. Denote

$$H_1 := \left\{ g: [0, 1] \rightarrow \mathbb{R}^d: g(t) = \int_0^t f(s) ds, t \in [0, 1], f \in L_2([0, 1]) \right\} \quad (7.12)$$

the space of all absolutely continuous functions with square integrable derivative equipped with the norm

$$\|g\|_{H_1} = \left( \int_0^1 |\dot{g}(t)|^2 dt \right)^{1/2}, \quad g \in H_1.$$

**Lemma 7.10** For any  $d \in \mathbb{N}$  and any  $\tau, \varepsilon, \delta >$ ,

$$\mathbb{P} \left( \sup_{0 \leq t \leq \tau} \{|B_\varepsilon(t)| \geq \delta\} \right) \leq 4de^{-\delta^2/2d\tau\varepsilon}. \quad (7.13)$$

The proof of this lemma requires an application of the reflection principle for Brownian motions. We briefly recall the notation and basic facts and refer the reader to either MA4 - Brownian motions module or the excellent book [MP10]. Suppose that  $(B(t))_{t \geq 0}$  is standard Brownian motion in  $\mathbb{R}$  and that  $T$  is a stopping time. The process  $(B^*(t))_{t \geq 0}$  called *Brownian motion reflected at  $T$*  and defined by

$$B^*(t) = B(t)\mathbb{1}\{t \leq T\} + (2B(T) - B(t))\mathbb{1}\{t > T\}$$

is also standard Brownian motion.

We now apply this reflection principle to one-dimensional Brownian motion  $(B(t))_{t \geq 0}$ . Let

$$M(t) := \max_{0 \leq s \leq t} \{B(s)\}, \quad t \geq 0.$$

A priori it is not at all clear what the distribution of this random variable is, but we can determine it as a consequence of the reflection principle. In the following,  $\mathbb{P}_0$  denotes the distribution of the process with initial state at  $0 \in \mathbb{R}^d$ .

**Theorem 7.11 (Maximum process, [MP10])** If  $a > 0$  then

$$\mathbb{P}_0(M(t) > a) = 2\mathbb{P}_0(B(t) > a) = \mathbb{P}_0(|B(t)| > a).$$

**Proof of Lemma 7.10.** In the following we write  $B(t) = (B^{(1)}(t), \dots, B^{(d)}(t))$  and note that the  $d$  coordinates are independent identically distributed one-dimensional Brownian motions.

$$\mathbb{P}\left(\sup_{0 \leq t \leq \tau} \{|B_\varepsilon(t)|\} \geq \delta\right) = \mathbb{P}\left(\sup_{0 \leq t \leq \tau} \{|B(t)|^2\} \geq \varepsilon^{-1} \delta^2\right) \leq d \mathbb{P}\left(\sup_{0 \leq t \leq \tau} \{|B^{(i)}(t)|^2\} \geq \frac{\delta^2}{d\varepsilon}\right)$$

where the inequality is due to the set inclusion

$$\{x \in \mathbb{R}^d : |x|^2 \geq \alpha\} \subset \bigcup_{i=1}^d \{x \in \mathbb{R}^d : |x_i|^2 \geq \frac{\alpha}{d}\}.$$

As the laws of  $B_t$  and  $\sqrt{\tau} B_{t/\tau}$  are identical, we obtain by time rescaling,

$$\mathbb{P}\left(\sup_{0 \leq t \leq \tau} \{|B_\varepsilon(t)|\} \geq \delta\right) \leq d \mathbb{P}\left(\|B_{[0,t]}^{(1)}\| \geq \frac{\delta}{\sqrt{d\tau\varepsilon}}\right).$$

Since  $B^{(1)}$  and  $-B^{(1)}$  possess the same law in  $\mathcal{C}_0([0, 1])$ , by the reflection principle in Theorem 7.11,

$$\mathbb{P}(\|B_{[0,t]}^{(1)}\| \geq \eta) \leq 2\mathbb{P}\left(\sup_{0 \leq t \leq 1} \{B^{(1)}(t)\} \geq \eta\right) = 4\mathbb{P}(B^{(1)}(1) \geq \eta) \leq 4e^{-\frac{\eta^2}{2}},$$

where the last inequality follows by Chebycheff's bound. □

**Theorem 7.12 (Schilder's theorem)** *The family  $(\nu_\varepsilon)_{\varepsilon>0}$  of laws  $\nu_\varepsilon \in \mathcal{M}_1(\mathcal{C}_0([0, 1]))$  satisfies in  $\mathcal{C}_0([0, 1])$  an LDP with rate  $\varepsilon^{-1}$  and good rate function*

$$I_S(h) = \begin{cases} \frac{1}{2} \int_0^1 |\dot{h}(t)|^2 dt & , \text{ if } h \in H_1, \\ \infty & , \text{ otherwise.} \end{cases} \quad (7.14)$$

**Proof of Theorem 7.12.** Observe that  $\widehat{B}_\varepsilon$ , defined as

$$\widehat{B}_\varepsilon(t) := B_\varepsilon(\varepsilon \lfloor t/\varepsilon \rfloor), \quad 0 \leq t \leq 1,$$

is the process  $Y_\varepsilon$  in Theorem 7.9, for the particular choice of the random variables  $X_k$ , namely, the  $X_k$  are standard normally distributed with unit variance and zero mean, i.e.,  $X_k \sim N(0, 1), k \in \mathbb{N}$ . Thus, by Theorem 7.9, the probability laws of  $\widehat{B}_\varepsilon$  satisfy the LDP in  $L_\infty([0, 1])$  with the good rate  $I$  from Theorem 7.1. Similar to Exercise 1(c) on Example Sheet 1, we can compute

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\langle \lambda, x_1 \rangle}] = \frac{1}{2} |\lambda|^2, \quad \lambda \in \mathbb{R}^d,$$

implying

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, x \rangle - \frac{1}{2} |\lambda|^2\} = \frac{1}{2} |x|^2.$$

Hence,  $\mathcal{D}_I = H_1$ , and the rate function  $I$  in Theorem 7.1 specialises to  $I_S$  in Theorem 7.12. We are now left to show the LDP for the scaled process  $B_\varepsilon$ . The idea is to use Theorem 7.4, that is, we shall show that  $\hat{B}$  and  $B_\varepsilon$  are exponentially equivalent. Then, Theorem 7.4 implies our statement. For any  $\delta > 0$ , using Lemma 7.10,

$$\mathbb{P}(\|B_\varepsilon - \hat{B}_\varepsilon\| \geq \delta) \leq (\lfloor 1/\varepsilon \rfloor + 1) \mathbb{P}(\sup_{0 \leq t \leq \varepsilon} \{|B_\varepsilon(t)| \geq \delta\}) \leq 4d\varepsilon^{-1}(1 + \varepsilon)e^{-\delta^2/(2d\varepsilon^2)}.$$

Consequently, by Lemma 7.10,

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \mathbb{P}(\|B_\varepsilon - \hat{B}_\varepsilon\| \geq \delta) = -\infty,$$

and by Theorem 7.4, it follows that  $(\nu_\varepsilon)_{\varepsilon \geq 0}$  satisfies the LDP in  $L_\infty([0, 1])$  with good rate function  $I_S$ . The restriction to the space  $\mathcal{C}_0([0, 1])$  follows from Lemma 3.5, since  $B_\varepsilon \in \mathcal{C}_0([0, 1])$  with probability one.  $\square$

### 7.3 Application: pinning reward for polymer chains and random interfaces

We discuss some direct applications of Mogulskii's theorem (Theorem 7.1), namely scaling limits of random walks and random walk bridges and their relation to random fields over lattices. A key aspect of this study is the fact that we obtain rate functions which have at least two distinct zeros for some choice of critical parameter as dimension, boundary conditions and pinning strength. For this we modify the distributions of our Markov chains with adding some bias towards a subspace of the state space. The general setting is Markov chains in discrete time with continuous state space  $\mathbb{R}^m, m \geq 1$ . Consider a family  $(\varphi_x)_{x \in \mathbb{Z}^d}$  of random variables  $\varphi_x$  taking values in  $\mathbb{R}^m$ . Any probability measure  $\mu \in \mathcal{M}_1((\mathbb{R}^m)^{\mathbb{Z}^d})$  is called a *random field over  $\mathbb{Z}^d$* , or, alternatively, a  *$(d + m)$ -dimensional random field model*. A distribution is given by the collection of all finite-dimensional distribution, that is, for all finite  $\Lambda \subset \mathbb{Z}^d$ , all inverse temperatures  $\beta > 0$ , all boundary conditions  $\psi \in (\mathbb{R}^m)^{\mathbb{Z}^d}$  and all admissible *potential functions*  $V: \mathbb{R}^m \rightarrow \mathbb{R}$ ,

$$\mu_\Lambda^\psi(d\varphi) = \frac{1}{Z_\Lambda(\psi)} \exp\left(-\beta \sum_{\{x,y\} \cap \Lambda \neq \emptyset; |x-y|=1} V((\varphi_x - \varphi_y))\right) \prod_{x \in \Lambda} d\varphi_x \prod_{x \in \Lambda^c} \delta_{\psi_x}(d\varphi_x), \quad (7.15)$$

where  $\Lambda^c = \mathbb{Z}^d \setminus \Lambda$  and

$$Z_\Lambda(\psi) \int \exp\left(-\beta \sum_{\{x,y\} \cap \Lambda \neq \emptyset; |x-y|=1} V((\varphi_x - \varphi_y))\right) \prod_{x \in \Lambda} d\varphi_x \prod_{x \in \Lambda^c} \delta_{\psi_x}(d\varphi_x) \quad (7.16)$$

is the normalisation constant, called *partition function* for  $\Lambda$  and inverse temperature  $\beta > 0$  and boundary condition  $\psi$ . A potential function is admissible if (7.16) is finite. The probability measure  $\mu_\Lambda^\psi$  on  $(\mathbb{R}^m)^\Lambda$  is called the *Gibbs distribution* in  $\Lambda$  with boundary condition  $\psi$ , potential function  $V$  and inverse temperature  $\beta$ . We restrict our discussion to  $d = 1$  and  $m = 1$  in the following. Denote  $\Lambda_N = \{1, \dots, N-1\}$  and its boundary  $\partial\Lambda_N = \{0, N\}$  and closure  $\overline{\Lambda_N}$ . As our Gibbs distribution depends solely on the nearest neighbour gradient of the random field  $(\varphi_x)_{x \in \mathbb{Z}}$ , we need to specify the boundary condition on site 0 and  $N$  only. We consider the following two cases for our boundary condition:

- 1.) Dirichlet boundary condition:  $\psi(0) = aN$  and  $\psi(N) = bN$  for a choice  $a, b \in \mathbb{R}$ .
- 2.) Free boundary condition (right hand side respectively no terminal condition):  
 $\psi(0) = aN, a \in \mathbb{R}$ .

We will see that the corresponding Gibbs distributions, i.e.,

$$\begin{aligned}\mu_N^{a,b}(\mathrm{d}\varphi) &:= \frac{1}{Z_N(a,b)} e^{-\beta \sum_{k=1}^M V(\varphi_k - \varphi_{k-1})} \prod_{k \in \Lambda_N} \mathrm{d}\varphi_k \delta_{aN}(\mathrm{d}\varphi_0) \delta_{bN}(\mathrm{d}\varphi_N), \\ \mu_N^{a,f}(\mathrm{d}\varphi) &:= \frac{1}{Z_N(a)} e^{-\beta \sum_{k=1}^M V(\varphi_k - \varphi_{k-1})} \prod_{k \in \Lambda_N} \mathrm{d}\varphi_k \delta_{aN}(\mathrm{d}\varphi_0),\end{aligned}\tag{7.17}$$

are in fact certain Markov chain bridge respectively Markov chain distributions. The state space of the Markov chains is  $\mathbb{R}^m$ , and the processes have finite (discrete) time horizon  $\{0, 1, \dots, N\}$ . Depending on whether there is a boundary condition on the right hand side or not we identify a terminal condition for the terminal time  $N$  or not. When a process obeys a terminal condition we call the process a *bridge process*. The transition probability density in the state space  $\mathbb{R}^m$  is given by

$$\frac{e^{-V(y-x)}}{Z} \mathrm{d}y, \quad x, y \in \mathbb{R}^m, \quad Z = \int_{\mathbb{R}} e^{-V(x)} \mathrm{d}x,$$

that is, transition probability density from state  $\varphi_{k-1}$  to state  $\varphi_k$  (one time unit) is

$$\frac{e^{-V(\varphi_k - \varphi_{k-1})}}{Z} \mathrm{d}\varphi_k,$$

Hence, our Gibbs distributions in (7.17) are probability measure for the whole path, i.e., they are Markov chain distributions for bridge processes or free processes. We make the following assumptions on the potential function  $V$ .

**Assumption 7.13 (Potential function  $V$ )** (i)

$$\sup_{x \in \mathbb{R}^m} \{e^{\langle \lambda, x \rangle} e^{-V(|x|)}\} < \infty, \quad \text{for all } \lambda \in \mathbb{R}^m.$$

(ii)

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

is finite for all  $x \in \mathbb{R}^m$ , and it satisfies  $\Lambda^* \in \mathcal{C}^3(\mathbb{R}^m)$ .

When  $m = 1$ , the Markov chain path  $\varphi = (\varphi_k)_{k \in \overline{\Lambda_N}}$  may be interpreted as the heights of a random interface of some reference plane. We now study *scaling limits* for the Markov chains. The macroscopic time parameter of the chain, observed after scaling, runs over the interval  $D = [0, 1]$ . The range of the discrete (microscopic) time for the Markov chain is  $D_N = \overline{\Lambda_N}$ . Denote  $h_N = (h_N(t))_{t \in D}$ ,  $h_N: [0, 1] \rightarrow \mathbb{R}^m$ , be the *macroscopic path*, called the *empirical profile*, of the Markov chain determined

from the microscopic path  $\varphi$  with respect to  $\mu_N^{a,b}$  or  $\mu_N^{a,f}$  by linear interpolation of  $(h_N(k/N))_{k \in D_N}$ ,  $h_N(k/N) = 1/N \varphi(k)$ , i.e.,

$$h_N(t) = \frac{\lfloor Nt \rfloor - Nt + 1}{N} \varphi(\lfloor Nt \rfloor) + \frac{Nt - \lfloor Nt \rfloor}{N} \varphi(\lfloor Nt \rfloor + 1), \quad t \in D. \quad (7.18)$$

A direct application of Mogulskii's theorem in Theorem 7.1 gives the following statement.

**Theorem 7.14** *The sequence  $(h_N)_{N \in \mathbb{N}}$  under  $\mu_N^{a,b}$  (respectively,  $\mu_N^{a,f}$ ) satisfies in  $L_\infty([0, 1])$  the LDP with rate  $N$  and good rate function*

$$I(h) = \begin{cases} \int_0^1 \Lambda^*(\dot{h}(t)) dt - \mathcal{N}(a, b) & , \text{if } h \in \mathcal{AC}, h(0) = a, h(1) = b, \\ \infty & , \text{otherwise,} \end{cases} \quad (7.19)$$

with normalisation  $\mathcal{N}(a, b) = \inf_{h: [0,1] \rightarrow \mathbb{R}^m} \{ \int_0^1 \Lambda^*(\dot{h}(t)) dt \}$ , respectively,

$$I^f(h) = \begin{cases} \int_0^1 \Lambda^*(\dot{h}(t)) dt - \mathcal{N}(a) & , \text{if } h \in \mathcal{AC}, h(0) = a, \\ \infty & , \text{otherwise,} \end{cases} \quad (7.20)$$

with normalisation  $\mathcal{N}(a) = \inf_{h: [0,1] \rightarrow \mathbb{R}^m} \{ I^f(h) \}$ .

We now modify the Markov chain distribution by building a bias towards a subspace of the state space  $\mathbb{R}^m$ . We focus on  $\{0\} \subset \mathbb{R}^m$ , that is, for some parameter  $\varepsilon > 0$ , define

$$\mu_{N,\varepsilon}^{a,b}(\mathrm{d}\varphi) = \frac{1}{Z_{N,\varepsilon}(a, b)} e^{-\sum_{k=1}^N V(|\varphi_k - \varphi_{k-1}|)} \prod_{k \in \Lambda_N} (\mathrm{d}\varphi_k + \varepsilon \delta_0(\mathrm{d}\varphi_k)) \delta_{aN}(\mathrm{d}\varphi_0) \delta_{bN}(\mathrm{d}\varphi_N) \quad (7.21)$$

with partition function

$$Z_{N,\varepsilon}(a, b) = \int_{\mathbb{R}^{\Lambda_N}} e^{-\sum_{k=1}^N V(|\varphi_k - \varphi_{k-1}|)} \prod_{k \in \Lambda_N} (\mathrm{d}\varphi_k + \varepsilon \delta_0(\mathrm{d}\varphi_k)) \delta_{aN}(\mathrm{d}\varphi_0) \delta_{bN}(\mathrm{d}\varphi_N),$$

and define  $\mu_{N,\varepsilon}^{a,f}$  similarly. Some well known results show that the limits

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Z_N(a, b) \quad \text{and} \quad \lim_{N \rightarrow \infty} \frac{1}{N} \log Z_{N,\varepsilon}(a, b)$$

exist, and they are called the *limiting free energy* respectively the *limiting pinning free energy*. The difference of these two free energy for zero boundary conditions is denoted

$$\tau(\varepsilon) = \lim_{N \rightarrow \infty} \frac{1}{N} \log \frac{Z_{N,\varepsilon}(0, 0)}{Z_N(0, 0)}. \quad (7.22)$$

Similarly,

$$\tau^f(\varepsilon) = \lim_{N \rightarrow \infty} \frac{1}{N} \log \frac{Z_{N,\varepsilon}(0)}{Z_N(0, f)}$$

exists for the free boundary condition. In the following we write super index  $D$  for Dirichlet boundary conditions and super index  $f$  for the free boundary conditions on the right hand side (no terminal condition). Some facts:

(i) The limits  $\tau^D(\varepsilon)$  and  $\tau^f(\varepsilon)$  exist for all  $\varepsilon \geq 0$ .

(ii)  $\exists$  critical values  $0 \leq \varepsilon_c^D \leq \varepsilon_c^f$  such that

$$\tau(\varepsilon) > 0 \Leftrightarrow \varepsilon > \varepsilon_c^D$$

(iii)

$$\varepsilon_c^D = \begin{cases} > 0 & , m \geq 3, \\ 0 & , m = 1, 2. \end{cases}$$

Note that the pinning measures  $\mu_{N,\varepsilon}^{a,b}, \mu_{N,\varepsilon}^{a,f}$  are distortions of the Markov chain distribution, and we call them  $(1+m)$ -dimensional pinning models. We are interested in LDPs for the empirical profiles under the pinning model measures. We only state the result and outline some key aspects. Ultimately, the pinning reward term with parameter  $\varepsilon$  leads to two distinct zeros of the corresponding rate function. The idea is to write the pinning measure as a sum over possible pinning sets when the field or the state of the Markov chain is exactly zero. This methods allows to write the pinning measures as a weighted sum of pinning free measures with additional internal conditions when the the field assume the value 0. A generalisation of the Binomial expansion leads to the following observation for any measurable function  $f: \mathbb{R}^{\mathbb{Z}} \rightarrow \mathbb{R}$  (we choose  $m = 1$  for simplicity),

$$\begin{aligned} \mathbb{E}_{\mu_{N,\varepsilon}^{a,b}}[f] &= \frac{1}{Z_{N,\varepsilon}(a,b)} \int f(\varphi) \exp\left(-\beta \sum_{k=1}^N V(\varphi_k - \varphi_{k-1})\right) \prod_{k \in \Lambda_N} (\mathrm{d}\varphi_k + \varepsilon \delta_0(\mathrm{d}\varphi_k)) \times \\ &\quad \times \delta_{aN}(\mathrm{d}\varphi_0) \delta_{bN}(\mathrm{d}\varphi_N) \\ &= \sum_{\mathcal{P} \subset \Lambda_N} \varepsilon^{|\mathcal{P}|} \frac{Z_{\Lambda_N \setminus \mathcal{P}}(\psi_a, \psi_b)}{Z_{N,\varepsilon}(a,b)} \mathbb{E}_{\Lambda_N \setminus \mathcal{P}}[f], \end{aligned} \tag{7.23}$$

where

$$\psi_a(x) = \begin{cases} a & , x = 0, \\ 0 & , x \neq 0, \end{cases} \quad \psi_b(x) = \begin{cases} b & , x = N, \\ 0 & , x \neq N, \end{cases}$$

are the boundary conditions adjusted to the pinning site  $\mathcal{P}$  as the expectation  $\mathbb{E}_{\Lambda_N \setminus \mathcal{P}}$  is with respect to the measure  $\mu_{\Lambda_N \setminus \mathcal{P}}$ , which is pinning free but with addition 'internal boundary conditions' when the field assumes the value 0, and which is defined for any pinning set  $\mathcal{P} \subset \Lambda_N$  by

$$\begin{aligned} \mu_{\Lambda_N \setminus \mathcal{P}}(\mathrm{d}\varphi) &= \frac{1}{Z_{\Lambda_N \setminus \mathcal{P}}(\psi_a, \psi_b)} \int \exp\left(-\beta \sum_{k=1}^N V(\varphi_k - \varphi_{k-1})\right) \times \\ &\quad \times \prod_{k \in \Lambda_N} \mathrm{d}\varphi_k \prod_{k \in \mathcal{P}} \delta_0(\mathrm{d}\varphi_k) \delta_{aN}(\mathrm{d}\varphi_0) \delta_{bN}(\mathrm{d}\varphi_N). \end{aligned} \tag{7.24}$$

We observe that the pinning measure  $\mu_{N,\varepsilon}^{a,b}$  (similar statements hold for  $\mu_{N,\varepsilon}^{a,f}$ ) is a convex combination of pinning free probabilities  $\mu_{\Lambda_N \setminus \mathcal{P}}$  which are distributed according to a probability measure on the power set  $\mathcal{P}(\Lambda_N)$  of  $\Lambda_N$ , namely,

$$\Xi_N(\mathcal{P}) = \varepsilon^{|\mathcal{P}|} \frac{Z_{\Lambda_N \setminus \mathcal{P}}(\psi_a, \psi_b)}{Z_{N,\varepsilon}(a, b)}, \quad \mathcal{P} \subset \Lambda_N.$$

We can see the probability measure of  $\Xi_N(\mathcal{P})$  as the percolation probability that the set  $\mathcal{P}$  is 'open'. A large deviation principle can be obtained using the expression above, see [FS04] for the case of potential function  $V(x) = \frac{1}{2}x^2$ , that is, for Gaussian random walks. This can be generalised to the class of potential function defined in Assumption 7.13 using [FO10], which uses different techniques as well.

**Theorem 7.15 (Pinning LDP, [FS04, FO10])** *The sequence  $(h_N)_{N \in \mathbb{N}}$  of empirical profiles satisfies under  $\mu_{N,\varepsilon}^{a,b}$  the LDP in  $L_\infty([0, 1])$  with rate  $N$  and good rate function*

$$I^\varepsilon(h) = \begin{cases} \int_0^1 \Lambda^*(\dot{h}(t)) dt - \tau(\varepsilon) |\{t \in D : h(t) = 0\}| - \mathcal{N}_\varepsilon(a, b) & , \text{ if } h \in \mathcal{AC}, \text{ and} \\ & h(0) = a, h(1) = b, \\ \infty & , \text{ otherwise,} \end{cases} \quad (7.25)$$

where

$$\mathcal{N}_\varepsilon(a, b) = \inf_{h \in \mathcal{AC}, h(0)=a, h(1)=b} \left\{ \int_0^1 \Lambda^*(\dot{h}(t)) dt - \tau(\varepsilon) |\{t \in D : h(t) = 0\}| \right\},$$

is the normalisation of the rate function and where  $|\{t \in D : h(t) = 0\}|$  denotes the Lebesgue measure of the zero set of the function  $h$ .

One can show that the rate function  $I^\varepsilon$  has for certain parameter  $(\tau(\varepsilon), a, b, m)$  two distinct zeros, that is, there exist  $h_1, h_2 \in \mathcal{AC}$  with  $h_i(0) = a, h_i(1) = b, i = 1, 2, h_1 \neq h_2$ , and  $I^\varepsilon(h_1) = I^\varepsilon(h_2) = 0$ . Denote the set of zeros of the rate function  $I^\varepsilon$  by  $\mathcal{M}^\varepsilon = \{h_1, h_2\}$ . Then the LDP in Theorem 7.15 tell us that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \mu_{N,\varepsilon}^{a,b}(\text{dist}_\infty(h_N, \mathcal{M}^\varepsilon) \leq \delta) = 1$$

for every  $\delta > 0$ , where  $\text{dist}_\infty$  denotes the distance under the supremum norm  $\|\cdot\|$ . More precisely, for every  $\delta > 0$  there exists  $c(\delta) > 0$  such that

$$\mu_{N,\varepsilon}^{a,b}(\text{dist}_\infty(h_N, \mathcal{M}^\varepsilon) > \delta) \leq e^{-c(\delta)N},$$

for  $N$  large enough. We say that the two functions  $h_1, h_2 \in \mathcal{M}^\varepsilon$  coexist in the limit  $N \rightarrow \infty$  (scaling limiting regime) under the measure  $\mu_{N,\varepsilon}^{a,b}$  with probabilities  $p_1, p_2 > 0$ ,  $p_1 + p_2 = 1$ , when

$$\lim_{N \rightarrow \infty} \mu_{N,\varepsilon}^{a,b}(\|H_N - h_i\| \leq \delta) = p_i, \quad i = 1, 2,$$

holds for  $\delta > 0$  small enough. We then say that the scaling limit has a non-trivial concentration of measure on the two possible scaling limits. The two zeros are defined as follows.

$$h_1(t) = a + t(b - a), \quad t \in [0, 1],$$

is the affine function connection  $a$  and  $b$  by a straight line. This zero does pick any reward as, depending on the values of  $a$  and  $b$ , the function can have no zeros or exactly one zero which has zero Lebesgue measure. In the other case, the random walk picks up as much of the reward as possible, namely, there are  $\ell = \ell(\varepsilon, a, b) \in (0, 1)$  and  $r = r(\varepsilon, a, b) \in (0, 1)$ ,  $r > \ell$ , such that

$$h_2(t) = \begin{cases} \frac{(\ell-t)}{\ell}a & , t \in [0, \ell], \\ 0 & , t \in [\ell, 1-r], \\ \frac{(t+r-1)}{r}b & , t \in [1-r, 1], \end{cases}$$

which picks up reward in the interval  $[\ell, 1-r] \subset [0, 1]$ .

## References

- [Ada01] S. ADAMS. Complete equivalence of the gibbs ensembles for one-dimensional markov systems. *Journal of Statistical Physics* **105**, no. 5/6, (2001), 879–908.
- [CK81] I. CSISZÁR and J. KÖRNER. *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadaémiai Kiadó, Budapest, 1981.
- [dH00] F. DEN HOLLANDER. *Large Deviations*, vol. 14 of *American Mathematical Society*. AMS, 2000.
- [Dur19] R. DURRETT. *Probability - Theory and Examples*. Cambridge University Press, fifth ed., 2019.
- [DZ98] A. DEMBO and O. ZEITOUNI. *Large Deviations Techniques and Applications*. Springer, 1998.
- [FO10] T. FUNAKI and T. OTOBE. Scaling limits for weakly pinned random walks with two large deviation minimizers. *J. Math. Soc. Japan* **62**, no. 3, (2010), 1005–1041.
- [FS04] T. FUNAKI and H. SAKAGAWA. Large deviations for  $\nabla\varphi$  interface models and derivation of free boundary problems. *Advanced Studies in Pure Mathematics* **39**, (2004), 173–211.
- [Geo12] H.-O. GEORGII. *Stochastics*. De Gruyter Textbook. De Gruyter, 2nd rev. and ext. ed., 2012.
- [MP10] P. MÖRTERS and Y. PERES. *Brownian Motion*. Cambridge University Press, 2010.
- [Nor04] J. NORRIS. *Markov chains*. Cambridge University Press, 2004.
- [Roc70] R. ROCKAFELLAR. *Convex Analysis*. Princeton University Press, 1970.
- [Sen81] E. SENETA. *Non Negative Matrices and Markov Chains*. Springer, 2nd rev. and ext. ed., 1981.
- [Str05] D. STROOCK. *An Introduction to Markov Processes*. Springer, 2005.

## Index

- ( $1+m$ )-dimensional pinning models, 66
- $I$  continuity set, 5
- bridge process, 64
- Brownian motion reflected at  $T$ , 61
- coarser, 27
- cumulative distribution function (CDF), 72
- dual space, 32
- empirical mean, 2
- empirical measure, 15, 41
- empirical path, 55
- empirical profile, 64
- entropy, 14
- Entropy (Shannon), 14
- essential supremum, 74
- essentially smooth, 33
- exponential tightness, 23
- exponentially equivalent, 56
- exposed point, 33
- exposing hyperplane, 33
- finer, 27
- finite signed measures, 40
- finite-volume distributions, 53
- Gâteaux differentiable, 39
- Gibbs distribution, 63
- Gibbs parameter, 52
- good rate function, 4
- irreducible, 42
- Ising model, 53
- Jensen's inequality, 73
- large deviation principle (LDP), 5
- lattice gas, 53
- Legendre-Fenchel transform, 38
- Legendre-Fenchel transform, 6
- level set, 4
- limiting free energy, 65
- logarithmic moment generating function, 2, 6, 38
- macro state, 13
- macroscopic path, 64
- micro state, 13
- microcanonical entropy, 54
- moment generation function, 72
- pair empirical measure, 18
- partition function, 53, 63
- Perron-Frobenius, 42
- Perron-Frobenius eigenvalue, 42
- pinning free energy, 65
- potential functions, 63
- random field, 63
- rate function, 1, 4
- relative entropy, 14
- relative interior, 37
- scaling limits, 64
- spin, 53
- stochastic matrix, 41
- supremum norm, 55
- total variation distance, 15
- type, 14
- type class, 14
- types, 14
- Varadhan Lemma, 27
- weak large deviation principle (LDP), 23

## Appendix A Preliminaries on Probability Theory

We recall some basic concepts and results of probability theory. The reader should be familiar with most of this material some of which is taught in elementary probability courses in the first year. To make these lectures self-contained we review the material mostly without proof and refer the reader to basic chapters of common undergraduate textbooks in probability theory, e.g. [Dur19] and [Geo12]. In Section A.1 we present basic definitions for probability space and probability measure as well as random variables along with expectation, variance and moments. Vital for the lecture will be the review of all classical inequalities in Section A.2.

### A.1 Random variables

A probability space  $(\Omega, \mathcal{F}, P)$  is a triple consisting of a set  $\Omega$ , a  $\sigma$ -algebra  $\mathcal{F}$  and a probability measure  $P$ . We write  $\mathcal{P}(\Omega)$  for the power set of  $\Omega$  which is the set of all subsets of  $\Omega$ .

**Definition A.1 ( $\sigma$ -algebra)** Suppose  $\Omega \neq \emptyset$ . A system  $\mathcal{F} \subset \mathcal{P}(\Omega)$  satisfying

- (a)  $\Omega \in \mathcal{F}$
- (b)  $A \in \mathcal{F} \Rightarrow A^c := \Omega \setminus A \in \mathcal{F}$
- (c)  $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i \geq 1} A_i \in \mathcal{F}$ .

is called  $\sigma$ -algebra (or  $\sigma$ -field) on  $\Omega$ . The pair  $(\Omega, \mathcal{F})$  is then called an event space or measurable space.

**Example A.2 (Borel  $\sigma$ -algebra)** Let  $\Omega = \mathbb{R}^n, n \in \mathbb{N}$  and

$$\mathcal{G} = \left\{ \prod_{i=1}^n [a_i, b_i] : a_i < b_i, a_i, b_i \in \mathbb{Q} \right\}$$

be the system consisting of all compact rectangular boxes in  $\mathbb{R}^n$  with rational vertices and edges parallel to the axes. In honour of Émile Borel (1871–1956), the system  $\mathcal{B}^n = \sigma(\mathcal{G})$  is called the Borel  $\sigma$ -algebra on  $\mathbb{R}^n$ , and every  $A \in \mathcal{B}^n$  a Borel set. Here,  $\sigma(\mathcal{G})$  denotes the smallest  $\sigma$ -algebra generated by the system  $\mathcal{G}$ . Note that the  $\mathcal{B}^n$  can also be generated by the system of open or half-open rectangular boxes, see [Dur19, Geo12].



The decisive point in the process of building a stochastic model is the next step: For each  $A \in \mathcal{F}$  we need to define a value  $P(A) \in [0, 1]$  that indicates the probability of  $A$ . Sensibly, this should be done so that the following holds.

(N) *Normalisation*:  $P(\Omega) = 1$ .

(A)  *$\sigma$ -Additivity*: For pairwise disjoint events  $A_1, A_2, \dots \in \mathcal{F}$  one has

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i).$$

**Definition A.3 (Probability measure)** Let  $(\Omega, \mathcal{F})$  be a measurable space. A function  $P: \mathcal{F} \rightarrow [0, 1]$  satisfying the properties (N) and (A) is called a probability measure or a probability distribution, in short a distribution (or, a little old-fashioned, a probability law) on  $(\Omega, \mathcal{F})$ . Then the triple  $(\Omega, \mathcal{F}, P)$  is called a probability space.

**Theorem A.4 (Construction of probability measures via densities)** (a) *Discrete case: For countable  $\Omega$ , the relations*

$$P(A) = \sum_{\omega \in A} \varrho(\omega) \text{ for } A \in \mathcal{P}(\Omega), \quad \varrho(\omega) = P(\{\omega\}) \text{ for } \omega \in \Omega$$

*establish a one-to-one correspondence between the set of all probability measures  $P$  on  $(\Omega, \mathcal{P}(\Omega))$  and the set of all sequences  $\varrho = (\varrho(\omega))_{\omega \in \Omega}$  in  $[0, 1]$  such that  $\sum_{\omega \in \Omega} \varrho(\omega) = 1$ .*

(b) *Continuous case: If  $\Omega \subset \mathbb{R}^n$  is Borel, then every function  $\varrho: \Omega \rightarrow [0, \infty)$  satisfying the properties*

$$(i) \quad \{x \in \Omega: \varrho(x) \leq c\} \in \mathcal{B}_\Omega^n \text{ for all } c > 0,$$

$$(ii) \quad \int_\Omega \varrho(x) dx = 1$$

*determines a unique probability measure on  $(\Omega, \mathcal{B}_\Omega^n)$  via*

$$P(A) = \int_A \varrho(x) dx \text{ for } A \in \mathcal{B}_\Omega^n$$

*(but not every probability measure on  $(\Omega, \mathcal{B}_\Omega^n)$  is of this form).*

**Proof.** See [Dur19, Geo12]. □

**Definition A.5** A sequence or function  $\varrho$  as in Theorem A.4 above is called a density (of  $P$ ) or, more explicitly (to emphasise normalisation), a probability density (function), often abbreviated as *pdf*. If a distinction between the discrete and continuous case is required, a sequence  $\varrho = (\varrho(\omega))_{\omega \in \Omega}$  as in case (a) is called a discrete density, and a function  $\varrho$  in case (b) a Lebesgue density.

In probability theory one often considers the transition from a measurable space (event space)  $(\Omega, \mathcal{F})$  to a coarser measurable (event) space  $(\Omega', \mathcal{F}')$ . In general such a mapping should satisfy the requirement

$$A' \in \mathcal{F}' \Rightarrow X^{-1}A' := \{\omega \in \Omega: X(\omega) \in A'\} \in \mathcal{F}. \quad (\text{A.1})$$

**Definition A.6** Let  $(\Omega, \mathcal{F})$  and  $(\Omega', \mathcal{F}')$  be two measurable (event) spaces. Then every mapping  $X: \Omega \rightarrow \Omega'$  satisfying property (A.1) is called a *random variable* from  $(\Omega, \mathcal{F})$  to  $(\Omega', \mathcal{F}')$ , or a random element of  $\Omega'$ , or a  $\Omega'$ -valued random variable. Alternatively (in the terminology of measure theory),  $X$  is said to be measurable relative to  $\mathcal{F}$  and  $\mathcal{F}'$ .

In probability theory it is common to write  $\{X \in A'\} := X^{-1}A'$ .

**Theorem A.7 (Distribution of a random variable)** *If  $X$  is a random variable from a probability space  $(\Omega, \mathcal{F}, P)$  to a measurable space  $(\Omega', \mathcal{F}')$ , then the prescription*

$$P'(A') := P(X^{-1}A') = P(\{X \in A'\}) \equiv P(X \in A') \quad \text{for any } A' \in \mathcal{F}'$$

*defines a probability measure  $P'$  on  $(\Omega', \mathcal{F}')$ .*

**Definition A.8** (a) The probability measure  $P'$  in Theorem A.7 is called the *distribution of  $X$  under  $P$* , or the image of  $P$  under  $X$ , and is denoted by  $P \circ X^{-1}$ . (In the literature, one also finds the notations  $P_X$  or  $\mathcal{L}(X; P)$ . The letter  $\mathcal{L}$  stands for the more traditional term law, or loi in French.)

(b) Two random variables are said to be identically distributed if they have the same distribution.

We are considering real-valued or  $\mathbb{R}^n$ -valued random variables in the following and we just call them random variables for all these cases. In basic courses in probability theory, one learns about the two most important quantities associated with a random variable  $X$ , namely the expectation<sup>1</sup> (also called the mean) and variance. They will be noted in this lecture by

$$\mathbb{E}[X] \quad \text{and} \quad \text{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2].$$

The distribution of a real-valued random variable  $X$  is determined by the *cumulative distribution function* (CDF) of  $X$ , defined as

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}((-\infty, t]), \quad t \in \mathbb{R}. \quad (\text{A.2})$$

It is often more convenient to work with the tails of random variables, namely with

$$\mathbb{P}(X > t) = 1 - F_X(t). \quad (\text{A.3})$$

Here we write  $\mathbb{P}$  for the generic distribution of the random variable  $X$  which is given by the context.

For any real-valued random variable the *moment generating function* (MGF) (MGF) is defined

$$M_X(\lambda) := \mathbb{E}[e^{\lambda X}], \quad \lambda \in \mathbb{R}. \quad (\text{A.4})$$

When  $M_X$  is finite for all  $\lambda$  in a neighbourhood of the origin, we can easily compute all moments by taking derivatives (interchanging differentiation and expectation (integration) in the usual way):

$$\mathbb{E}[X^k] = \frac{d^k}{d\lambda^k} \Big|_{\lambda=0} M_X(\lambda), \quad k \in \mathbb{N}. \quad (\text{A.5})$$

<sup>1</sup>In measure theory the expectation  $\mathbb{E}[X]$  of a random variable on a probability space  $(\Omega, \mathcal{F}, P)$  is the Lebesgue integral of the function  $X: \Omega \rightarrow \mathbb{R}$ . This makes theorems on Lebesgue integration applicable in probability theory for expectations of random variables

**Lemma A.9 (Integral Identity)** *Let  $X$  be a real-valued non-negative random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) dt.$$

**Proof.** We can write any non-negative real number  $x$  via the following identity using indicator function <sup>2</sup>:

$$x = \int_0^x 1 dt = \int_0^\infty \mathbb{1}_{\{t < x\}}(t) dt.$$

Substitute now the random variable  $X$  for  $x$  and take expectation (with respect to  $X$ ) on both sides. This gives

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\{t < X\}}(t) dt\right] = \int_0^\infty \mathbb{E}[\mathbb{1}_{\{t < X\}}] dt = \int_0^\infty \mathbb{P}(t < X) dt.$$

To change the order of expectation and integration in the second inequality, we used the Fubini-Tonelli theorem.  $\square$

**Exercise A.10 (Integral identity)** Prove the extension of Lemma A.9 to any real-valued random variable (not necessarily positive):

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) dt - \int_{-\infty}^0 \mathbb{P}(X < t) dt.$$



## A.2 Classical Inequalities

In this section fundamental classical inequalities are presented. Here, classical refers to typical estimates for analysing stochastic limits.

**Proposition A.11 (Jensen's inequality)** *Suppose that  $\Phi: I \rightarrow \mathbb{R}$ , where  $I \subset \mathbb{R}$  is an interval, is a convex function. Let  $X$  be a real-valued random variable. Then*

$$\Phi(\mathbb{E}[X]) \leq \mathbb{E}[\Phi(X)].$$

**Proof.** See [Dur19] or [Geo12] using either the existence of sub-derivatives for convex functions or the definition of convexity with the epi-graph of a function. The epi-graph of a function  $f: I \rightarrow \mathbb{R}$ ,  $I \subset \text{some interval}$ , is the set

$$\text{epi}(f) := \{(x, t) \in \mathbb{R}^2 : x \in I, f(x) \leq t\}.$$

A function  $f: I \rightarrow \mathbb{R}$  is convex if and only if  $\text{epi}(f)$  is a convex set in  $\mathbb{R}^2$ .  $\square$

---

<sup>2</sup>  $\mathbb{1}_A$  denotes the indicator function of the set  $A$ , that is,  $\mathbb{1}_A(t) = 1$  if  $t \in A$  and  $\mathbb{1}_A(t) = 0$  if  $t \notin A$ .

### A.3 $L^p$ -spaces

In the following let  $X$  be a  $\mathbb{R}$ -valued random variable, i.e., there is a probability space  $(\Omega, \mathcal{F}, P)$  such that  $X: \Omega \rightarrow \mathbb{R}$  is a measurable function. By default, we equip the real line  $\mathbb{R}$  with its Borel- $\sigma$ -algebra. We begin with the definition of the *essential supremum* of  $X$ .

**Definition A.12 (Essential supremum)** Let  $X$  be  $\mathbb{R}$ -valued random variable. The *essential supremum* of  $X$ , written  $\text{ess-sup}(X)$ , is the smallest number  $\alpha \in \mathbb{R}$  such that the set  $\{x \in \Omega: X(x) > \alpha\}$  has measure zero, that is,

$$P(\{x \in \Omega: X(x) > \alpha\}) = 0.$$

If no such number exists we define  $\text{ess-sup}(X) = \infty$ .

To understand this definition better we shall check the following example.

**Example A.13 (Essential supremum being infinity)** Suppose that  $\Omega = (0, 1)$ ,  $\mathcal{F} = \mathcal{B}((0, 1))$ , and let  $P$  be the uniform measure on  $(0, 1)$ . This measure has constant probability density,

$$P(A) = \int_{\Omega} \mathbb{1}_A(t) dt = b - a, \quad \text{for any } A = (a, b) \text{ with } 0 \leq a < b \leq 1.$$

Define  $X: (0, 1) \rightarrow \mathbb{R}, x \mapsto \frac{1}{x}$ . Then  $X$  is continuous function and therefore measurable. Then  $\text{ess sup}(X) = \infty$ . To see this, pick any  $\alpha \in \mathbb{R}_+$ . Then

$$\{x \in (0, 1): \frac{1}{x} > \alpha\} = (0, \frac{1}{\alpha})$$

and

$$P((0, \frac{1}{\alpha})) = \frac{1}{\alpha} > 0.$$

As this holds for all  $\alpha > 0$ , we have that  $\text{ess-sup}(X) = \infty$ . ♣

**Definition A.14** Let  $(\Omega, \mathcal{F}, P)$  be a probability space. Given two measurable functions  $f, g: [0, \infty]$ , we say that  $f$  is *equivalent to*  $g$ , written  $f \sim g$ , if

$$f(x) = g(x) \quad \text{for } P - a.e. x \in \Omega,$$

that is,

$$P(\{x \in \Omega: f(x) \neq g(x)\}) = 0.$$

We shall identify - with an abuse of notation - identify a measurable function  $f$  with its equivalence class  $[f]$ .

**Definition A.15** Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $1 \leq p < \infty$ .

$$L^p \equiv L^p(\Omega, \mathcal{F}, P) := \{f: \Omega \rightarrow [-\infty, \infty]: f \text{ measurable and } \|f\|_{L^p} < \infty\},$$

where

$$\|f\|_{L^p} := \left( \int_{\Omega} |f|^p dP \right)^{\frac{1}{p}} = \left( \int_{\Omega} |f(x)|^p P(dx) \right)^{\frac{1}{p}}.$$

If  $p = \infty$ , then

$$L^\infty \equiv L^\infty(\Omega, \mathcal{F}, P) := \{f: \Omega \rightarrow [-\infty, \infty]: f \text{ measurable and } \|f\|_{L^\infty} < \infty\},$$

where


$$\|f\|_{L^\infty} := \text{ess-sup}(|f|),$$

and we write  $\|f\|_\infty \equiv \|f\|_{L^\infty}$  occasionally.

A consequence of Jensen's inequality is that  $\|X\|_{L^p}$  is an increasing function in the parameter  $p$ , i.e.,

$$\|X\|_{L^p} \leq \|X\|_{L^q} \quad 0 \leq p \leq q \leq \infty. \quad (\text{A.6})$$

This follows from the convexity of  $\Phi(x) = x^{\frac{q}{p}}$  when  $q \geq p$ .

**Exercise A.16** Show that (A.6) holds. 

**Proposition A.17 (Minkowski's inequality)** For  $p \in [1, \infty]$ , let  $X, Y \in L^p$ , then

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}.$$

**Proposition A.18 (Cauchy-Schwarz inequality)** For  $X, Y \in L^2$ ,

$$|\mathbb{E}[XY]| \leq \|X\|_{L^2} \|Y\|_{L^2}.$$

**Proposition A.19 (Hölder's inequality)** For  $p, q \in (1, \infty)$  with  $1/p + 1/q = 1$  let  $X \in L^p$  and  $Y \in L^q$ . Then

$$\mathbb{E}[XY] \leq \mathbb{E}[|XY|] \leq \|X\|_{L^p} \|Y\|_{L^q}.$$

**Lemma A.20 (Linear Markov's inequality)** For non-negative random variables  $X$  and  $t > 0$  the tail probability is bounded as

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

**Proof.** Pick  $t > 0$ . Any positive number  $x$  can be written as

$$x = x \mathbb{1}_{\{X \geq t\}} + x \mathbb{1}_{\{X < t\}}.$$

As  $X$  is non-negative, we insert  $X$  into the above expression in place of  $x$  and take the expectation (integral) to obtain

$$\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}_{\{X \geq t\}}] + \mathbb{E}[X \mathbb{1}_{\{X < t\}}] \geq \mathbb{E}[t \mathbb{1}_{\{X \geq t\}}] = t \mathbb{P}(X \geq t).$$

□

This is one particular version of the Markov inequality which provides linear decay in  $t$ . In the following proposition we obtain the general version which will be used frequently throughout the lecture.

**Proposition A.21 (Markov's inequality)** *Let  $Y$  be a real-valued random variable and  $f: [0, \infty) \rightarrow [0, \infty)$  an increasing function. Then, for all  $\varepsilon > 0$  with  $f(\varepsilon) > 0$ ,*

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[f \circ |Y|]}{f(\varepsilon)}.$$

**Proof.** Clearly, the composition  $f \circ |Y|$  is a positive random variable such that

$$f(\varepsilon)\mathbb{1}_{\{|Y| \geq \varepsilon\}} \leq f \circ |Y|.$$

Taking the expectation on both sides of that inequality gives

$$f(\varepsilon)\mathbb{P}(|Y| \geq \varepsilon) = \mathbb{E}[f(\varepsilon)\mathbb{1}_{\{|Y| \geq \varepsilon\}}] \leq \mathbb{E}[f \circ |Y|].$$

□

The following version of the Markov inequality is often called Chebyshev's inequality.

**Corollary A.22 (Chebyshev's inequality, 1867)** *For all  $Y \in L^2$  with  $\mathbb{E}[Y] \in (-\infty, \infty)$  and  $\varepsilon > 0$ ,*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\text{Var}(Y)}{\varepsilon^2}.$$

## Appendix B Modes of Convergence

We shall review in this chapter the basic modes of convergence of random variables. Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of random variables taking values in some metric space  $(E, d)$ , that is, each  $X_n: \Omega \rightarrow E$  is a measurable map between a given probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and the range or target space  $(E, d)$  where one equips the metric space  $E$  with its Borel- $\sigma$ -field (algebra)  $\mathcal{B}(E)$ . Let  $X$  be a random variable taking values in  $(E, d)$ .

**Definition B.1 (always surely or almost everywhere or with probability 1 or strongly)** The sequence  $(X_n)_{n \in \mathbb{N}}$  converges almost surely or almost everywhere or with probability 1 or strongly towards  $X$  if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = \mathbb{P}\left(\{\omega \in \Omega: \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

This means that the values of  $X_n$  approach the value of  $X$ , in the sense that events for which  $X_n$  does not converge to  $X$  have probability 0. We write  $X_n \xrightarrow{\text{a.s.}} X$  for almost sure convergence.

**Definition B.2 (Convergence in probability)** The sequence  $(X_n)_{n \in \mathbb{N}}$  converges in probability to  $X$  if

$$\lim_{n \rightarrow \infty} \mathbb{P}(d(X_n, X) > \varepsilon) = 0, \quad \text{for all } \varepsilon > 0.$$

We write  $X_n \xrightarrow{P} X$  for convergence in probability.

**Proposition B.3 (Markov's inequality)** Let  $Y$  be a real-valued random variable and  $f: [0, \infty) \rightarrow [0, \infty)$  an increasing function. Then, for all  $\varepsilon > 0$  with  $f(\varepsilon) > 0$ ,

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[f \circ |Y|]}{f(\varepsilon)}.$$

**Corollary B.4 (Chebyshev's inequality, 1867)** For all  $Y \in L^2$  and  $\varepsilon > 0$ ,

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\text{Var}(Y)}{\varepsilon^2}.$$

By Chebyshev's inequality the convergence in probability is equivalent to  $\mathbb{E}[d(X_n, X) \wedge 1] \rightarrow 0$  as  $n \rightarrow \infty$ . This is related to the almost sure convergence as follows.

**Lemma B.5 (Subsequence criterion)** Let  $X, X_1, X_2, \dots$  be random variables in  $(E, d)$ . Then  $(X_n)_{n \in \mathbb{N}}$  converges to  $X$  in probability if and only if every subsequence  $N' \subset \mathbb{N}$  has a further subsequence  $N'' \subset N'$  such that  $X_n \rightarrow X$  almost surely along  $N''$ . In particular,  $X_n \xrightarrow{\text{a.s.}} X$  implies that  $(X_n)_{n \in \mathbb{N}}$  converges to  $X$  in probability.

**Definition B.6 (Convergence in distribution)** We say that  $X_n$  converges in distribution to  $X$ , if, for every bounded continuous function  $f: E \rightarrow \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f].$$

We write  $X_n \xrightarrow{d} X$  for convergence in distribution.

**Remark B.7** (a)  $X_n \xrightarrow{d} X$  is equivalent to weak convergence of the distributions.

(b) if  $X_n \xrightarrow{d} X$  and  $g: E \rightarrow \mathbb{R}$  continuous, then  $g(X_n) \xrightarrow{d} g(X)$ . But note that, if  $E = \mathbb{R}$  and  $X_n \xrightarrow{d} X$ , this does not imply that  $\mathbb{E}[X_n]$  converges to  $\mathbb{E}[X]$ , as  $g(x) = x$  is not a bounded function on  $\mathbb{R}$ .

(c) Suppose  $E = \{1, \dots, m\}$  is finite and  $d(x, y) = 1 - \mathbb{1}_{x=y}$ . Then  $X_n \xrightarrow{d} X$  if and only if  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$  for all  $k \in E$ .

(d) Let  $E = [0, 1]$  and  $X_n = 1/n$  almost surely. Then  $X_n \xrightarrow{d} X$ , where  $X = 0$  almost surely. However, note that  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = 0) = 0 \neq \mathbb{P}(X = 0)$ .

◇

## Appendix C Law of large numbers and the central limit theorem

**Definition C.1 (Variance and covariance)** Let  $X, Y \in L^2$  be real-valued random variables.

(a)

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

is called the **variance**, and  $\sqrt{\text{Var}(X)}$  the **standard deviation** of  $X$  with respect to  $\mathbb{P}$ .

(b)

$$\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

is called the **covariance** of  $X$  and  $Y$ . It exists since  $|XY| \leq X^2 + Y^2$ .

(c) If  $\text{cov}(X, Y) = 0$ , then  $X$  and  $Y$  are called **uncorrelated**.

**Theorem C.2 (Weak law of large numbers,  $L^2$ -version)** Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of uncorrelated (e.g. independent) real-valued random variables in  $L^2$  with bounded variance, in that  $v := \sup_{n \in \mathbb{N}} \text{Var}(X_n) < \infty$ . Then for all  $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq \varepsilon\right) \leq \frac{v}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0,$$

and thus  $\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \xrightarrow{\mathbb{P}} 0$ . In particular, if  $\mathbb{E}[X_i] = \mathbb{E}[X_1]$  for all  $i \in \mathbb{N}$ , then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1].$$

We now present a second version of the weak law of large numbers, which does not require the existence of the variance. To compensate we must assume that the random variables, instead of being pairwise uncorrelated, are even pairwise independent and identically distributed.

**Theorem C.3 (Weak law of large numbers,  $L^1$ -version)** Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of pairwise independent, identically distributed real-valued random variables in  $L^1$ . Then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1].$$

**Theorem C.4 (Strong law of large numbers)** If  $(X_n)_{n \in \mathbb{N}}$  is a sequence of pairwise uncorrelated real-valued random variables in  $L^2$  with  $v := \sup_{n \in \mathbb{N}} \text{Var}(X_n) < \infty$ , then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \rightarrow 0 \text{ almost surely as } n \rightarrow \infty.$$

**Theorem C.5 (Central limit theorem; A.M. Lyapunov 1901, J.W. Lindeberg 1922, P. Le  vy 1922)**

Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of independent, identically distributed real-valued random variables in  $L^2$  with  $\mathbb{E}[X_i] = m$  and  $\text{Var}(X_i) = v > 0$ . Then,

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - m}{\sqrt{v}} \xrightarrow{d} N(0, 1).$$

The normal distribution is defined in the following section.

## Appendix D Normal distribution

A real-valued random variable  $X$  is **normally** distributed with mean  $\mu$  and variance  $\sigma^2 > 0$  if

$$\mathbb{P}(X > x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_x^\infty e^{-\frac{(u-\mu)^2}{2\sigma^2}} du, \quad \text{for all } x \in \mathbb{R}.$$

We write  $X \sim N(\mu, \sigma^2)$ . We say that  $X$  is standard normal distributed if  $X \sim N(0, 1)$ .

A random vector  $X = (X_1, \dots, X_n)$  is called a **Gaussian random vector** if there exists an  $n \times m$  matrix  $A$ , and an  $n$ -dimensional vector  $b \in \mathbb{R}^n$  such that  $X^T = AY + b$ , where  $Y$  is an  $m$ -dimensional vector with independent standard normal entries, i.e.  $Y_i \sim N(0, 1)$  for  $i = 1, \dots, m$ . Likewise, a random variable  $Y = (Y_1, \dots, Y_m)$  with values in  $\mathbb{R}^m$  has the  $m$ -dimensional standard Gaussian distribution if the  $m$  coordinates are standard normally distributed and independent. The covariance matrix of  $X = AY + b$  is then given by

$$\text{cov}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T] = AA^T.$$

**Lemma D.1** If  $A$  is an orthogonal  $n \times n$  matrix, i.e.  $AA^T = \mathbb{I}$ , and  $X$  is a  $n$ -dimensional standard Gaussian vector, then  $AX$  is also a  $n$ -dimensional standard Gaussian vector.

**Lemma D.2** Let  $X_1$  and  $X_2$  be independent and normally distributed with zero mean and variance  $\sigma^2 > 0$ . Then  $X_1 + X_2$  and  $X_1 - X_2$  are independent and normally distributed with mean 0 and variance  $2\sigma^2$ .

**Proposition D.3** If  $X$  and  $Y$  are  $n$ -dimensional Gaussian vectors with  $\mathbb{E}[X] = \mathbb{E}[Y]$  and  $\text{cov}(X) = \text{cov}(Y)$ , then  $X$  and  $Y$  have the same distribution.

**Corollary D.4** A Gaussian random vector  $X$  has independent entries if and only if its covariance matrix is diagonal. In other words, the entries in a Gaussian vector are uncorrelated if and only if they are independent.

**Lemma D.5 (Inequalities)** Let  $X \sim N(0, 1)$ . Then for all  $x > 0$ ,

$$\frac{x}{x^2 + 1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \leq \mathbb{P}(X > x) \leq \frac{1}{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

## Appendix E Gaussian integration formulae

For any  $a > 0$ ,

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\pi/a}.$$

For  $b \in \mathbb{C}$  and  $a > 0$ ,

$$I(b) = \int_{-\infty}^{\infty} e^{-a/2x^2 + bx} dx = e^{b^2/2a} \sqrt{2\pi/a}.$$

Let  $A \in \mathbb{R}^{n \times n}$ ,  $A = A^T > 0$  (i.e. all eigenvalues of  $A$  are positive), and define  $C = A^{-1}$  and write  $\langle \varphi, \psi \rangle$  for the scalar product of  $\varphi, \psi \in \mathbb{R}^n$ .

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}\langle \varphi, A\varphi \rangle} \prod_{i=1}^n d\varphi_i = (2\pi)^{n/2} \det(A^{-\frac{1}{2}}) = \det(2\pi C)^{\frac{1}{2}}.$$

For any  $J \in \mathbb{C}^n$  we obtain

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}\langle \varphi, A\varphi \rangle + \langle J, \varphi \rangle} \prod_{i=1}^n d\varphi_i = \det(2\pi C)^{\frac{1}{2}} e^{\frac{1}{2}\langle J, C J \rangle}.$$

Let  $C \in \mathbb{R}^{n \times n}$  be invertible matrix and  $C > 0$ . The probability measure  $\mu_C \in \mathcal{M}_1(\mathbb{R}^n)$  defined by

$$\mu_C(d\varphi) = \frac{1}{\sqrt{\det(2\pi C)}} e^{-1/2\langle \varphi, C^{-1}\varphi \rangle} \prod_{i=1}^n d\varphi_i,$$

is called the **Gaussian measure** on  $\mathbb{R}^n$  with mean zero and covariance matrix  $C$ .

**The covariance splitting formula:** Let  $C_i = C_i^T, i = 1, 2$ , be positive invertible matrices. Define  $C = C_1 + C_2$ . Then for all  $F \in \mathcal{L}(\mu_C)$ ,

$$\begin{aligned} \int_{\mathbb{R}^n} F(\varphi) \mu_C(d\varphi) &= \int_{\mathbb{R}^n} \mu_{C_1}(d\varphi_1) \int_{\mathbb{R}^n} \mu_{C_2}(d\varphi_2) F(\varphi_1 + \varphi_2) \\ &= \int_{\mathbb{R}^n} \mu_{C_1}(d\varphi) \int_{\mathbb{R}^n} \mu_{C_2}(d(\varphi - \varphi_1)) F(\varphi). \end{aligned}$$

In other words, if  $C = C_1 + C_2$ , the Gaussian random variable  $\varphi$  is the sum of two independent (see above) Gaussian random variables,  $\varphi = \varphi_1 + \varphi_2$ , and the Gaussian measure factors, i.e.  $\mu_C = \mu_{C_1} \otimes \mu_{C_2}$ .

The characteristic function of a Gaussian vector  $X = (X_1, \dots, X_n)$  with mean  $\mu \in \mathbb{R}^n$  and covariance matrix  $C$  reads as

$$\varphi_X(t) = \mathbb{E} \left[ e^{i\langle t, \mu \rangle - \frac{1}{2}\langle t, C t \rangle} \right], \quad t \in \mathbb{R}^n.$$

An  $\mathbb{R}^n$ -valued stochastic process  $X = \{X_t: t \geq 0\}$  is called **Gaussian** if, for any integer  $k \geq 1$  and real numbers  $0 \leq t_1 < t_2 < \cdots < t_k < \infty$ , the random vector  $(X_{t_1}, \dots, X_{t_k})$  has a joint normal distribution. If the distribution of  $(X_{t+t_1}, \dots, X_{t+t_n})$  does not depend on  $t$ , we say that the process is stationary. The finite-dimensional distributions of a Gaussian process  $X$  are determined by its expectation vector  $m(t) := \mathbb{E}[X(t)], t \geq 0$ , and its covariance matrix

$$\varrho(s, t) := \mathbb{E}[(X_s - m(s))(X_t - m(t))^T], \quad s, t \geq 0.$$

If  $m(t) = 0$  for all  $t \geq 0$ , we say that  $X$  is a zero-mean Gaussian process.