MA946 - Introduction to Graduate Probability

Stefan Adams

2020, update 22.12.2020

Contents

1 Basic probability theory **1**
1.1 Probability spaces 1
1.2 Probability measures 3
1.3 Random variables 5
1.4 Conditional Probability and Independence 9
1.5 Classical inequalities 12

2 Limit Theorems **13**
2.1 Weak Law of Large Numbers (WLLN) 13
2.2 Kolmogorov's zero-one law and Borel-Cantelli Lemma 14
2.3 Strong Law of Large Numbers (SLLN) 16
2.4 The Central Limit Theorem (CLT) 18
2.5 Convergence of Probability Measures 26

3 Random walks and their scaling limit **32**
3.1 The simple random walk on $\mathbb{Z}^d$ 32
3.2 The Local central Limit Theorem (LCLT) 34
3.3 Strong Markov property and basic potential theory 37
3.4 Discrete Heat Equation and its scaling limit 44
3.5 Brownian motion 47

4 Large Deviation Theory **53**
4.1 Introduction and Definition 53
4.2 Combinatorial Techniques for finite sample spaces 57
4.3 Cramér Theorem, Varadhan Lemma, and basic principles 61

5 Random Fields **85**
5.1 Setting and definitions 85
5.2 The discrete Gaussian Free Field (DGFF) 85
5.3 Scaling limits 94

Appendix A Modes of Convergence **95**

Appendix B Law of large numbers and the central limit theorem **96**

Appendix C Normal distribution **97**

Appendix D Gaussian integration formulae **98**

Appendix E Some useful properties of the weak topology of probability measures **99**

Introduction

Preface

Introduction

1 Basic probability theory

1.1 Probability spaces

If one aims at describing any effects of chance, the first question is: What can happen in a given situation? And which part of this is relevant. All possibilities that seem natural to distinguish are collected into a set Ω.

Example 1.1 (Rolling a die) Rolling a die once we may take $\Omega = \{1, \ldots, 6\}$. Rolling a die n-times, $\Omega = \{1, \ldots, 6\}^n$, for $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ and $1 \leq i \leq n$, ω_i represents the number showing at the i-th row. ♣

Example 1.2 (Coin tossing) Tossing a coin infinitely often we take $\Omega = \{0, 1\}^{\mathbb{N}}$, where $\omega_i = 1$ means toss shows 'head'. ♣

Notation 1.3 Ω is the set of outcomes or the *sample space*.

We are often only interested in the occurrence of an event consisting of a certain selection of single outcomes, that is, we shall identify events with some system of subsets of Ω. The next examples represents an event as a subset of Ω.

Example 1.4 (Coin tossing with k-times heads) The event 'In n coin flips, heads shows at least k times' corresponds to the subset

$$A = \{\omega \in \Omega \colon \sum_{i=1}^{n} \omega_i \geq k\}$$

of the sample space $\Omega = \{0, 1\}^n$.

♣

We want to assign to each event A a probability $P(A) \in [0, 1]$ for all $A \in \mathcal{F}$ where $\mathcal{F}$ is some system of events (subsets of Ω). Why not simply taking $\mathcal{F} = \mathcal{P}(\Omega)$ with $\mathcal{P}(\Omega)$ being the power set of Ω. We shall check the following 'no-go theorem'.

Theorem 1.5 (Vitali, 1905) *The power set is too large. Let $\Omega = \{0, 1\}^{\mathbb{N}}$. Then there is* no *mapping $P \colon \mathcal{P}(\Omega) \to [0, 1]$ with the following properties:*

(N) Normalisation: $P(\Omega) = 1$.

(A) σ-Additivity: If $A_1, A_2, \ldots \subset \Omega$ are pairwise disjoint, then

$$P\Big(\bigcup_{i \geq 1} A_i\Big) = \sum_{i \geq 1} P(A_i)\,.$$

(I) Flip invariance: For all $A \subset \Omega, n \geq 1$, one has

$$P(T_n(A)) = P(A)\,,$$

where

$$T_n(\omega) = (\omega_1, \ldots, \omega_{n-1}, 1 - \omega_n, \omega_{n+1}, \ldots)\,, \quad \omega \in \Omega\,.$$

Proof. Exercise.

□

Definition 1.6 (σ-algebra) Let $\Omega \neq \varnothing$. A system $\mathcal{F} \subset \mathcal{P}(\Omega)$ of subsets satisfying

(a) $\Omega \in \mathcal{F}$,

(b) $A \in \mathcal{F} \Rightarrow A^{\mathrm{c}} := \Omega \setminus A \in \mathcal{F}$,

(c)

$$A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i \geq 1} A_i \in \mathcal{F}\,,$$

is called a *σ-algebra on* Ω. The pair $(\Omega, \mathcal{F})$ is called *event space* or *measurable space*.

Note that due to

$$\bigcap_{i \in I} = \Big(\bigcup_{i \in I} A_i^{\mathrm{c}}\Big)^{\mathrm{c}}\,,$$

it follows that $\mathcal{F}$ is closed under countable intersections.

Remark 1.7 (Generating σ-algebras) If $\Omega \neq \varnothing$ and $\mathcal{G} \subset \mathcal{P}(\Omega)$ is arbitrary, then there is a unique σ-algebra $\mathcal{F} = \sigma(\mathcal{G})$ on Ω such that $\mathcal{F} \supset \mathcal{G}$. This $\mathcal{F}$ is called the *σ-algebra generated by* $\mathcal{G}$, and $\mathcal{G}$ is called a *generator of* $\mathcal{F}$.

◇

Example 1.8 (i) **The power set**. Suppose that Ω is countable and $\mathcal{G} = \{\{\omega\} \colon \omega \in \Omega\}$ the system containing the singleton sets of Ω. Then, $\sigma(\mathcal{G}) = \mathcal{P}(\Omega)$. Indeed, since every $A \in \mathcal{P}(\Omega)$ is countable, it follows that $A = \bigcup_{\omega \in A}\{\omega\} \in \sigma(\mathcal{G})$.

(ii) **The Borel σ-algebra**. Let $\Omega = \mathbb{R}^n$ and

$$\mathcal{G} = \Big\{ \prod_{i=1}^{n} [a_i, b_i] \colon a_i < b_i, a_, b_i \in \mathbb{Q} \Big\}\,.$$

The system $\mathcal{B}^n := \sigma(\mathcal{G})$ is called the *Borel σ-algebra on* $\mathbb{R}^n$. For $n = 1$ we simply write $\mathcal{B}^1 = \mathcal{B}$.

♣

Remark and Definition 1.9 (Product σ-algebra) Suppose $\Omega = \prod_{i \in I}$ for some index set $I \neq \varnothing$; $\mathcal{E}_i$ a $\sigma-$algebra on E_i and

$$X_i \colon \Omega \to E_i$$

the projection onto the ith coordinate.

$$\mathcal{G} = \{X_i^{-1} A_i \colon i \in I, A_i \in \mathcal{E}_i\}$$

is the collection of all sets in Ω specified by an event in a single coordinate. Then

$$\bigotimes_{i\in I} \mathcal{E}_i := \sigma(\mathcal{G})$$

is called the *product σ-algebra* of the $\mathcal{E}_i$ on Ω. If $E_i = E$ and $\mathcal{E}_i = \mathcal{E}$ for all i, we write $\mathcal{E}^{\otimes I}$ instead of $\bigotimes_{i\in I} \mathcal{E}_i$. ◇

1.2 Probability measures

The crucial step in building a stochastic model is to assign a probability $P(A) \in [0,1]$ for each event $A \in \mathcal{F}$ in such a way that the following holds:

(N) **Normalisation:** $P(\Omega) = 1$.

(A) **σ-additivity:** For pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$ (that is, $A_i \cap A_j = \varnothing$ for $i \neq j$) one has

$$P\Big(\bigcup_{i\geq 1} A_i\Big) = \sum_{i\geq 1} P(A_i)\,.$$

Definition 1.10 Let $(\Omega, \mathcal{F})$ be a measurable space. A function $P\colon \mathcal{F} \to [0,1]$ satisfying properties (N) and (A) is called a *probability measure* (or probability distribution). The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*. The set of all probability measures on $(\Omega, \mathcal{F})$ is denoted $\mathcal{M}_1(\Omega, \mathcal{F})$. We write sometimes $\mathcal{M}_1(\Omega)$ when the σ-algebra is clear from the context.

Example and Definition 1.11 Let $(\Omega, \mathcal{F})$ be a measurable space and $\xi \in \Omega$, then

$$\delta_\xi(A) = \begin{cases} 1 & \text{if } \xi \in A\,, \\ 0 & \text{otherwise} \end{cases}, \qquad A \in \mathcal{F}\,,$$

defines a probability measure δ_ξ on $(\Omega, \mathcal{F})$. The measure $\delta_\xi \in \mathcal{M}_1(\Omega, \mathcal{F})$ is called the *Dirac distribution* or the *unit mass at the point ξ*.

Theorem 1.12 *Every probability measure P on a measurable space $(\Omega, \mathcal{F})$ has the following properties, for arbitrary events $A, B, A_1, B_1, \ldots \in \mathcal{F}$.*

(a) $P(\varnothing) = 0$.

(b) Finite additivity:

$$P(A \cup B) + P(A \cap B) = P(A) + P(B)\,,$$

and so in particular $P(A) + P(A^{\mathrm{c}}) = 1$.

(c) Monotonicity: If $A \subset B$, then $P(A) \leq P(B)$.

(d) *σ-Subadditivity:*

$$P\Big(\bigcup_{i\in I} A_i\Big) \le \sum_{i\in I} P(A_i)$$

for any countable index set I.

(e) *σ-Continuity: If either $A_n \uparrow A$ or $A_n \downarrow A$ (that is, the A_n are either increasing with union A, or decreasing with intersection A), then*

$$P(A_n) \to P(A) \text{ as } n \to \infty.$$

Proof. Exercise. □

Example 1.13 (Lebesgue measure) The mapping

$$\lambda^n \colon \mathcal{B}^n \to [0,\infty]$$

that assigns to each Borel set $A \in \mathcal{B}^n$ its n-dimensional volume

$$\lambda^n(A) := \int \mathbb{1}_A(x)\,\mathrm{d}x$$

satisfies the σadditivity, the σ-continuity, and the monotonicity, and $\lambda^n(\varnothing) = 0$. The mapping λ^n is a 'measure' on $(\mathbb{R}^n, \mathcal{B}^n)$ and is called the *n-dimensional Lebesgue measure*. For $\Omega \in \mathcal{B}^n$, the restriction λ^n_Ω of λ^n to $\mathcal{B}_\Omega := \{B \cap \Omega \colon B \in \mathcal{B}^n\}$ is called the *Lebesgue measure on Ω*.

◇

Theorem 1.14 (Construction of probability measures via densities) *(a) Discrete case. For countable sample spaces Ω, the relations*

$$P(A) = \sum_{\omega\in A} \varrho(\omega)\,, \quad A \in \mathcal{F}, \quad \varrho(\omega) = P(\{\omega\}) \text{ for } \omega \in \Omega$$

establish a one-to-one correspondence between all $P \in \mathcal{M}_1(\Omega, \mathcal{P}(\Omega))$ and the set of all sequences $\varrho = (\varrho(\omega))_{\omega\in\Omega}$ in $[0,1]$ with $\sum_{\omega\in\Omega} \varrho(\omega) = 1$.

(b) Continuous case. Let $\Omega \in \mathcal{B}^n$ be a Borel set. Then every function $\varrho \colon \Omega \to [0,\infty)$ satisfying

(i) $\{x \in \Omega \colon \varrho(x) \le c\} \in \mathcal{B}^n_\Omega$ for all $c > 0$.

(ii) $\int_\Omega \varrho(x)\,\mathrm{d}x = 1$

determines a probability measure $P \in \mathcal{M}_1(\Omega, \mathcal{B}^n_\Omega)$ via

$$P(A) = \int_A \varrho(x)\,\mathrm{d}x \quad A \in \mathcal{B}^n_\Omega\,.$$

Definition 1.15 A sequence or function ϱ as in Theorem 1.14 is called a *density for P* or a *probability density function*, often abbreviated as *pdf*.

1.3 Random variables

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

$$A' \in \mathcal{F}' \Rightarrow X^{-1}A' := \{\omega \in \Omega \colon X(\omega) \in A'\} \in \mathcal{F}. \tag{1.1}$$

Definition 1.16 Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable (event) spaces. The every mapping $X \colon \Omega \to \Omega'$ satisfying property (1.1) is called a random variable from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$, or a random element of Ω'. 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 1.17 (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*

$$P'(A') := P(X^{-1}A') = P(\{X \in A\}) = P(X \in A) \quad \textit{for } A' \in \mathcal{F}'$$

defines a probability measure P' on $(\Omega', \mathcal{F}')$.

Definition 1.18 (a) The probability measure P' in Theorem 1.17 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.

In the following when X is a random variable on some probability space we often write $\mathbb{P} = P \circ X^{-1}$

Definition 1.19 Let $X \colon \Omega \to \mathbb{R}$ be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, P)$. The distribution of a real-valued random variable X is determined by the *cumulative distribution function (CDF)* of X , defined as

$$F_X(t) := P(X \leq t) = \mathbb{P}((-\infty, t]), \quad t \in \mathbb{R}. \tag{1.2}$$

It is often more convenient to work with the tails of random variables, namely with

$$P(X > t) = 1 - F_X(t). \tag{1.3}$$

The *moment generating function (MGF)* is defined

$$M_X(\lambda) := \mathbb{E}[\mathrm{e}^{\lambda X}], \quad \lambda \in \mathbb{R}. \tag{1.4}$$

Remark 1.20 When M_X is finite for all λ in a neighbourhood of the origin, we can easily compute all moments by taking derivatives (interchanging differentiation and expectation (integration) in the usual way). $\diamond$

The expectation of a real-valued random variable is defined in two stages. First we define it for the case of random variables taking at most countably many different values. The general case is then the usual limiting procedure.

Definition 1.21 Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X\colon \;\to \mathbb{R}$ a real-valued random variable. X is called *discrete* if its range $X(\Omega) := \{X(\omega)\colon \omega \in \Omega\}$ is at most countable. A discrete random variable X has an expectation if

$$\sum_{x\in X(\Omega)} |x| P(X = x) < \infty\,.$$

Then the sum

$$\mathbb{E}[X] := \sum_{x\in X(\Omega)} x P(X = x)$$

is well-defined and is called the *expectation of* X, and one writes $X \in L^1(P)$, or $X \in L^1$. WE write sometimes $\mathbb{E}_P[X]$ instead $\mathbb{E}[X]$ to highlight the underlying probability measure.

Remark 1.22 If X is discrete and non-negative, then the sum $\sum_{x\in X(\Omega)} xP(X = x)$ is always well-defined, but it might be infinite. Discrete random variables always have an expectation, as long as we admit the value $+\infty$. Clearly, $X \in L^1(P)$ if and only if $\mathbb{E}[|X|] < \infty$. $\diamond$

Theorem 1.23 (Expectation rules) *Let $(\Omega, \mathcal{F}, P)$ be given and let $X, Y, X_n, Y_n\colon \Omega \to \mathbb{R}$ be discrete random variables in L^1. Then the following holds.*

(a) Monotonicity. If $X \le Y$, then $\mathbb{E}[X] \le \mathbb{E}[Y]$.

(b) Linearity. For every $c \in \mathbb{R}$, we have $\mathbb{E}[cX] = c\mathbb{E}[X]$, and $X + Y \in L^1$ and $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(c) σ-additivity and monotone convergence. If every $X_n \ge 0$ and $X = \sum_{n\ge 1} X_n$, then $\mathbb{E}[X] = \sum_{n\ge 1} \mathbb{E}[X_n]$. If $Y_n \uparrow Y$ for $n \to \infty$, it follows that $\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[Y_n]$.

Proof. Exercise. $\square$

For the general case of real-valued random variables $X\colon \Omega \to \mathbb{R}$ on some probability space $(\Omega, \mathcal{F}, P)$ we shall consider the $1/n$ discretisation

$$X_{(n)} := \frac{\lfloor nX \rfloor}{n}$$

of X. Thus

$$X_{(n)}(\omega) = \frac{k}{n} \text{ if } \frac{k}{n} \le X(\omega) < \frac{k+1}{n}\,.$$

We can easily show the following properties.

Lemma 1.24 *(a) For all $n \in \mathbb{N}$, the inequalities $X_{(n)} \leq X < X_{(n)} + \frac{1}{n}$ are valid.*

(b) If $X_{(n)} \in L^1$ for some $n \in \mathbb{N}$, then $X_{(n)} \in L^1$ for every $n \in \mathbb{N}$, and $(\mathbb{E}[X_{(n)}])_{n\in\mathbb{N}}$ is a Cauchy sequence.

Definition 1.25 Let $X\colon \Omega \to \mathbb{R}$ be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, P)$. Then X has an expectation if $X_{(n)} \in L^1(P)$ for some $n \in \mathbb{N}$. In this case,

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_{(n)}]$$

is called the *expectation of* X, and one says that X belongs to $L^1 = L^1(P)$.

Proposition 1.26 (Expectation Rules) *The calculation rules (a)-(c) in Theorem 1.23 carry over from discrete to general real-valued random variables.*

Lemma 1.27 (Integral Identity) *Let X be a real-valued non-negative random variable on some probability space $(\Omega, \mathcal{F}, P)$. Then*

$$\mathbb{E}[X] = \int_0^\infty P(X > t)\,\mathrm{d}t\,.$$

Proof. We can write any non-negative real number x via the following identity using indicator function [1]:

$$x = \int_0^x 1\,\mathrm{d}t = \int_0^\infty \mathbb{1}_{\{t<x\}}(t)\,\mathrm{d}t\,.$$

Substitute now the random variable X for x and take expectation (with respect to X) on both sides. This gives

$$\mathbb{E}[X] = \mathbb{E}\Big[\int_0^\infty \mathbb{1}_{\{t<X\}}(t)\,\mathrm{d}t\Big] = \int_0^\infty \mathbb{E}[\mathbb{1}_{\{t<X\}}]\,\mathrm{d}t = \int_0^\infty P(t < X)\,\mathrm{d}t\,.$$

To change the order of expectation and integration in the second inequality, we used the Fubini-Tonelli theorem. □

Exercise 1.28 (Integral identity) Prove the extension of Lemma 1.27 to any real-valued random variable (not necessarily positive):

$$\mathbb{E}[X] = \int_0^\infty P(X > t)\,\mathrm{d}t - \int_{-\infty}^0 P(X < t)\,\mathrm{d}t\,.$$

[1] $\mathbb{1}_A$ 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 $t \notin A$.

Remark 1.29 (Expectation depends solely on distribution) Note that for any real-valued random variable $X\colon \Omega \to \mathbb{R}$ on some probability space $(\Omega, \mathcal{F}, P)$ that

$$X \in L^1(P) \Leftrightarrow \mathrm{id}_{\mathbb{R}} \in L^1(P \circ X^{-1}),$$

where $\mathrm{id}_{\mathbb{R}}$ is the identity map of the real line $\mathbb{R}$. Then

$$\mathbb{E}_P[X] = \mathbb{E}_{P\circ X^{-1}}[\mathrm{id}_{\mathbb{R}}],$$

and thus the expectation of a random variable depends only on its distribution.

◇

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X\colon \Omega \to \mathbb{R}$ a real-valued random variable. If $X^r \in L^1(P)$ for some $r \in \mathbb{N}$, the expectation $\mathbb{E}[X^r]$ is called the *r-th moment of X*, and one writes $X \in L^r = L^r(P)$. Note that $L^s \subset L^r$ for $r < s$, since $|X|^r \leq 1 + |X|^s$.

Definition 1.30 (Variance and Covariance) Let $X, Y \in L^2$ be real-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$.

(a) The *variance of X* is defined as

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The square root $\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation of X with respect to P*.

(b) The *covariance* of X and Y relative to P is defined as

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

It exits since $|XY| \leq X^2 + Y^2$.

(c) X and Y are *uncorrelated with respect to P* if $\mathrm{Cov}(X, Y) = 0$.

Theorem 1.31 (Rules) *Let $X, Y, X_1, X_2, \ldots \in L^2$ be real-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$ and $a, b, c, d \in \mathbb{R}$.*

(a)

$$\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y),$$

and thus

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

(b)

$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

(c) $\sum_{i=1}^n X_i \in L^2$ *and*

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j).$$

If $X_1, \dots, X_n$ are pairwise uncorrelated, then

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,.$$

1.4 Conditional Probability and Independence

Proposition 1.32 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $A \in \mathcal{F}$ with $P(A) > 0$. Then there is a unique probability $P_A \in \mathcal{M}_1(\Omega, \mathcal{F})$ satisfying*

(i) $P_A(A) = 1$.

(ii) There exists $c_A > 0$ such that for all $B \in \mathcal{F}, B \subset A$, $P_A(B) = c_A P(B)$.

which is defined by

$$P_A(B) := \frac{P(A \cap B)}{P(A)} \quad \textit{for } B \in \mathcal{F}\,.$$

Proof. Clearly the defined probability measure P_A satisfies properties (i) and (ii). Now suppose that P_A satisfies (i) and (ii). We shall show that then P_A necessarily is given by the above formula. For every $B \in \mathcal{F}$ we have

$$P_A(B) = P_A(A \cap B) + P_A(B \setminus A) = c_A P(A \cap B),$$

where we used (ii) and the fact that $P_A(B \setminus A) = 0$ by (i). For $B = A$ it simply follow that $1 = P_A(A) = c_A P(A)$. Hence $c_A = 1/P(A)$ and P_A has the required form. □

Definition 1.33 In the setting of Proposition 1.32, for every $B \in \mathcal{F}$, the expression

$$P(B|A) := \frac{P(A \cap B)}{P(A)}$$

is called the *conditional probability of B given A with respect to P.*

Theorem 1.34 (Case-distinction and Bayes' formula) *Let $(\Omega, \mathcal{F})$ be measurable space with a countable partition of $\Omega = \bigcup_{i \in I}$ into pairwise disjoint events $(B_i)_{i \in I}$. The the following holds.*

(a) For all $A \in \mathcal{F}$,

$$P(A) = \sum_{i \in I} P(B_i) P(A|B_i)\,.$$

(b) Bayes' formula (1763). For all $A \in \mathcal{F}$ with $P(A) > 0$ and every $k \in I$,

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i \in I} P(B_i) P(A|B_i)}\,.$$

Proof. Exercise. □

From our intuition the independence of two events A and B is given when the probability one assigns to A is not influenced by the information that B has occurred. and likewise the occurrence of A doe snot lead to any re-weighting of the probability of B. In mathematical language this means that

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B) \quad \text{whenever } P(A), P(B) > 0\,.$$

Definition 1.35 Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events $A, B \in \mathcal{F}$ are called (stochastically) *independent with respect to* P if $P(A \cap B) = P(A)P(B)$.

Example 1.36 (Independence despite causality) Rolling two distinguishable dice, we set $\Omega = \{1, \ldots, 6\}^2$, $\mathcal{F} = \mathcal{P}(\Omega)$ and P the uniform distribution $P(\{\omega\}) = \frac{1}{36}\,, \omega \in \Omega$. Let $A = \{(k, \ell) \in \Omega \colon k + \ell = 7\}$ and $B = \{(k, \ell) \in \Omega \colon k = 6\}$. Then $|A| = |B| = 6$ ($|A| = \sharp A$ number of elements in A) and $|A \cap B| = 1$. Thus

$$P(A \cap B) = \frac{1}{6^2} = P(A)P(B)\,.$$

♣

Remark 1.37 The last example shows that ' independence means a proportional overlap of probabilities and does not necessarily involve any causality'. Furthermore note that $A \in \mathcal{F}$ is independent of itself if $P(A) \in \{0, 1\}$.

◇

Definition 1.38 Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\varnothing \neq I$ be an index set. The family $(A_i)_{i \in I}$ of events $A_i \in \mathcal{F}$ is called *independent with respect to* P if, for every finite subset $\varnothing \neq J \subset I$, we have

$$P\Big(\bigcap_{i \in J} A_i\Big) = \prod_{i \in J} P(A_i)\,.$$

The next example shows that there can be dependence despite pairwise independence.

Example 1.39 We are tossing a fair coin twice, i.e., $\Omega = \{0, 1\}^2, \mathcal{F} = \mathcal{P}(\Omega)$ and P uniform measure $P(\{\omega\}) = 1/4\,, \omega \in \Omega$. Define

$$\begin{aligned}
A &= \{1\} \times \{0, 1\} = \{\text{first toss 'heads'}\} \\
B &= \{0, 1\} \times \{1\} = \{\text{second toss 'heads'}\} \\
B &= \{(0, 0), (1, 1)\} = \{\text{both tosses give same result}\}\,.
\end{aligned}$$

Then

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$
$$P(A \cap C) = \frac{1}{4} = P(A)P(C)$$
$$P(B \cap C) = \frac{1}{4} = P(B)P(C)\,,$$

and thus A, B, C are pairwise independent. However,

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C)\,,$$

and thus the triple A, B, C is not independent. Note that $C = (A \cap B) \cup (A^c \cap B^c)$. ♣

We shall add the independence definition for families of random variables.

Definition 1.40 Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\varnothing \neq I$ be an index set. For all $i \in I$ let $Y_i \colon \Omega \to \Omega_i$ be a random variable for some measurable space $(\Omega_i, \mathcal{F}_i)$. The family $(Y_i)_{i \in I}$ of random variables Y_i is *independent with respect to* P if, for an arbitrary choice of events $B_i \in \mathcal{F}_i$, the family of events $(\{Y_i \in \mathbb{B}_i\})_{i \in I}$ is independent, i.e.,

$$P\Big(\bigcap_{i \in J} \{Y_i \in B_i\}\Big) = \prod_{i \in J} P(Y_i \in B_i)$$

holds for any finite $\varnothing \neq J \subset I$.
If X, Y are random variables defined on $(\Omega, \mathcal{F}, P)$ are independent we sometimes write $X \perp Y$.

Corollary 1.41 (Independence of finitely many random variables) *Let* $(Y_i)_{1 \le i \le n}$ *be a finite family of random variables defined on some probability space* $(\Omega, \mathcal{F}, P)$*.*

(a) If each Y_i *is* Ω_i*-valued with* Ω_i *at most countable. Then*

$$(Y_i)_{1 \le i \le n} \text{ independent } \Leftrightarrow P(Y_1 = \omega_1, \ldots, Y_n = \omega_n) = \prod_{i=1}^n P(Y_i = \omega_i)\,,\ \omega_i \in \Omega_i.$$

(b) Suppose each Y_i *is* $\mathbb{R}$*-valued. Then*

$$(Y_i)_{1 \le i \le n} \text{ independent } \Leftrightarrow P(Y_1 = c_1, \ldots, Y_n \le c_n) = \prod_{i=1}^n P(Y_i \le c_i)\,, c_i \in \mathbb{R}\,.$$

Note that independent random variables are uncorrelated and that the converse statement is not true in general.

Corollary 1.42 *If* X, Y *are random variables defined on* $(\Omega, \mathcal{F}, P)$ *are independent, i.e.,* $X \perp Y$*, then* X *and* Y *are uncorrelated.*

Example 1.43 (Uncorrelated but not independent) Suppose $\Omega = \{1,2,3\}$ with P being the uniform distribution. Define two random variables X and Y by $X = (1,0,-1)$ and $Y = (0,1,0)$, respectively. Then $XY = 0$ and $\mathbb{E}[X] = 0$, and thus $\mathrm{Cov}(X,Y) = 0$. However,

$$P(X = 1, Y = 1) = 0 \neq \frac{1}{9} = P(X = 1_P(Y = 1)$$

showing that X and Y are not independent. ♣

1.5 Classical inequalities

In this section fundamental classical inequalities are presented. Here, classical refers to typical estimates for analysing stochastic limits.

Proposition 1.44 (Jensen's inequality) *Suppose that $\Phi\colon I \to \mathbb{R}$, 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 $f\colon I \to \mathbb{R}, I \subset$ some interval, is the set

$$\mathrm{epi}(f) := \{(x, f(x)) \in \mathbb{R}^2 \colon x \in I\}.$$

A function $f\colon I \to \mathbb{R}$ is convex if, and only if $\mathrm{epi}(f)$ is a convex set in $\mathbb{R}^2$. □

A consequence of Jensen's inequality is that $\|X\|_{L^p}$ is an increasing function in the parameter p, i.e.,

$$\|X\|_{L^p} \leq \|X\|_{L^q} \qquad 0 \leq p \leq q \leq \infty.$$

This follows form the convexity of $\Phi(x) = x^{\frac{q}{p}}$ when $q \geq p$.

Proposition 1.45 (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 1.46 (Cauchy-Schwarz inequality) *For $X, Y \in L^2$,*

$$|\mathbb{E}[XY]| \leq \|X\|_{L^2}\|Y\|_{L^2}.$$

Proposition 1.47 (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 1.48 (Linear Markov's inequality) *For non-negative random variables X on some probability space $(\Omega, \mathcal{F}, P)$ and $t > 0$ the tail is bounded as*

$$P(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

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 and take the expectation (integral) to obtain

$$\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}_{\{X \geq t\}}] + \mathbb{E}[X\mathbb{1}_{\{X < t\}}] \geq \mathbb{E}[t\mathbb{1}_{\{X \geq t\}}] = tP(X \geq t).$$

□

This is one version of the Markov inequality which provides linear decay in t. In the following proposition we obtain the general version .

Proposition 1.49 (Markov's inequality) *Let Y be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, P)$ and $f\colon [0,\infty) \to [0,\infty)$ be an increasing function. Then, for all $\varepsilon > 0$ with $f(\varepsilon) > 0$,*

$$P(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[f \circ |Y|]}{f(\varepsilon)}.$$

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

$$f(\varepsilon)P(|Y| \geq \varepsilon) = \mathbb{E}[f(\varepsilon)\mathbb{1}_{\{|Y| \geq \varepsilon\}}] \leq \mathbb{E}[f \circ |Y|].$$

□

The following version of the Markov inequality is often called Chebyshev's inequality.

Corollary 1.50 (Chebyshev's inequality, 1867) *For all $Y \in L^2$ with $\mathbb{E}[Y] \in (-\infty, \infty)$ and $\varepsilon > 0$,*

$$P\Big(|Y - \mathbb{E}[Y]| \geq \varepsilon\Big) \leq \frac{\text{Var}(Y)}{\varepsilon^2}.$$

2 Limit Theorems

2.1 Weak Law of Large Numbers (WLLN)

If we briefly consider again tossing a fair coin we are inclined to try to prove that the empirical average converges to $1/2$. Suppose $P(X_i = 1) = \frac{1}{2} = P(X_i = 0)$, then $\mathbb{E}[X_i] = \frac{1}{2}$ for all $i \in \mathbb{N}$. Using Stirling's formula we can find a constant $C > 0$ such that

$$P\Big(\frac{1}{2n}\sum_{i=1}^{2n} X_i = \frac{1}{2}\Big) = \binom{2n}{n} 2^{-2n} \leq \frac{C}{\sqrt{\pi n}} \to 0 \quad \text{as } n \to \infty\,.$$

We thus see that we need another convergence criterium. We first consider convergence in probability.

Definition 2.1 (Convergence in probability) Suppose $(Y_n)_{n\in\mathbb{N}}$ and Y are real-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. We say that the sequence $(Y_n)_{n\in\mathbb{N}}$ *converges to* Y *in probability*, written as $Y_n \xrightarrow[n\to\infty]{\text{P}} Y$, if

$$P\Big(|Y_n - Y| \leq \varepsilon\Big) \to 1 \text{ as } n \to \infty \text{ for all } \varepsilon > 0\,.$$

Theorem 2.2 (Weak Law of Large Numbers (WLLN) - L^2 - version) *Let $(X_i)_{i\in\mathbb{N}}$ be pairwise uncorrelated random variables in L^2 for some probability space $(\Omega, \mathcal{F}, P)$ with uniformly bounded variance, $v := \sup_{i\in\mathbb{N}} \mathrm{Var}(X_i) < \infty$. Then*

$$Y_n := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow[n\to\infty]{\text{P}} 0\,.$$

Proof. From our assumptions we have that $Y_n \in L^2$ and $\mathbb{E}[Y_n] = 0$. Furthermore,

$$\mathrm{Var}(Y_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \leq \frac{v}{n}\,.$$

We thus conclude with Chebyshev's inequality 1.50. □

2.2 Kolmogorov's zero-one law and Borel-Cantelli Lemma

In the following we are dealing with sequences of random variables.

Theorem 2.3 *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of real-valued random variables on some probability space $(\Omega, \mathcal{F}, P)$. Then*

$$\inf_{i\in\mathbb{N}} X_i\,;\ \sup_{i\in\mathbb{N}} X_i\,; \limsup_{n\to\infty} X_n \ \textit{and}\ \liminf_{n\to\infty} X_n$$

are also random variables.

Proof. The infimum of a sequence is $< a$ if and only if some entry of that sequence is $< a$. Thus

$$\{\inf_{i\in\mathbb{N}} X_i < a\} = \bigcup_{i\in\mathbb{N}}\{X_i < a\} \in \mathcal{F}\,.$$

Similarly,

$$\{\sup_{i\in} > a\} = \bigcup_{i\in\mathbb{N}}\{X_i > a\} \in \mathcal{F}\,.$$

Furthermore,

$$\liminf_{n\to\infty} X_n = \sup_{n\in\mathbb{N}}\{\inf_{m\geq n} X_m\}$$
$$\limsup_{n\to\infty} X_n = \inf_{n\in\mathbb{N}}\{\sup_{m\geq n} X_m\}\,.$$

Note also that $Y_n := \inf_{m\geq n} X_m$ is a random variable. □

When analysing limits of random variables we often encounter the following events which lead naturally to another convergence criterion. In the setting of Theorem 2.3 the set

$$\Omega_0 := \{\omega \in \Omega\colon \lim_{n\to\infty} X_n \text{ exists}\} = \{\omega \in \Omega\colon \limsup_{n\to\infty} X_n = \liminf_{n\to\infty} X_n\} \tag{2.1}$$

is a measurable set.

Definition 2.4 (Almost sure convergence) If $P(\Omega_0) = 1$ for the set (2.1), we say that the sequence $(X_n)_{n\in\mathbb{N}}$ of random variables defined on $(\Omega, \mathcal{F}, P)$ *converges almost surely (almost everywhere) to the random variable* $X_\infty := \limsup_{n\to\infty}$ (X_∞ may take the value ∞).

The following tail events will be important when studying stochastic limits.

Definition 2.5 (Tail events) Let $(\Omega, \mathcal{F}, P)$ and $(\Omega_k, \mathcal{F}_k), k \in \mathbb{N}$, be given. Suppose that $(Y_k)_{k\in\mathbb{N}}$ is a sequence of Ω_k valued random variables Y_k defined on $(\Omega, \mathcal{F}, P)$. An event $A \in \mathcal{F}$ is called an *asymptotic or tail event* if, for every $n \in \mathbb{N}_0$, A depends only on $(Y_k)_{k>n}$, in that there exists an event $B_n \in \bigotimes_{k>n} \mathcal{F}_k$ such that

$$A = \{(Y_k)_{k>n} \in B_n\}\,.$$

We denote $\mathcal{T}(Y_k\colon k \in \mathbb{N})$ the collection of all such tail events.

Example 2.6 (Existence of long-term averages) Let $(\Omega_k, \mathcal{F}_k) = (\mathbb{R}, \mathcal{B})$ for all $k \in \mathbb{N}$ and $a < b$. Then

$$A = \Big\{\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^N Y_k \ \text{ exists and } \ \in [a, b]\Big\}$$

is a tail event for the sequence $(Y_k)_{k\in\mathbb{N}}$.

Proof. Denote $X_i\colon \prod_{k>n} \mathbb{R} \to \mathbb{R}$ the ith projection and define

$$B_n := \Big\{\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^N X_{n+k} \ \text{ exists and } \ \in [a, b]\Big\}\,.$$

Then $A\{(Y_k)_{k>n} \in B_n\}$ for every $n \in \mathbb{N}$, and thus A is a tail-event..

♣

Theorem 2.7 (Kolmogorov's zero-one law) *Let* $(Y_k)_{k\in\mathbb{N}}$ *be an independent sequence of random variables* Y_k *defined on some probability space* $(\Omega, \mathcal{F}, P)$ *and taking values in* $(\Omega_k, \mathcal{F}_k)$*. Then, for every* $A \in \mathcal{T}(Y_k\colon k \in \mathbb{N})$*, either* $P(A) = 0$ *or* $P(A) = 1$.

Proof. Elementary result from measure theory [BB01, Coh13] and undergraduate probability theory ([Kal02, Str93, Geo12, Bil12, Dur19].

□

Theorem 2.8 (Borel-Cantelli Lemma; 1909,1917) *Let $(A_k)_{k\in\mathbb{N}}$ be a sequence of events in $(\Omega, \mathcal{F}, P)$ and consider*

$$A = \{\omega \in \Omega\colon \omega \in A_k \text{ for infinitely many } k\} = \limsup_{k\to\infty} A_k\,.$$

Then the following statements hold.

(a) If $\sum_{k\in\mathbb{N}} P(A_k) < \infty$, then $P(A) = 0$.

(b) If $\sum_{k\in\mathbb{N}} P(A_k) = \infty$ and $(A_k)_{k\in\mathbb{N}}$ independent, then $P(A) = 1$.

Proof. *(a) We have that*

$$A \subset \bigcup_{k\geq m} A_k$$

and thus

$$P(A) \leq \sum_{k\geq m} P(A_k) \text{ for all } m.$$

The sum on the right hand side is the tail of a convergent series, and thus tends to 0 as $m \to \infty$.

(b)

$$A^{\mathrm{c}} = \bigcup_{m\in\mathbb{N}} \bigcap_{k\geq m} A_k^{\mathrm{c}}\,.$$

Thus, using that the independence of the A_k's implies the independence of the A_k^{c}'s and $1 - x \leq \mathrm{e}^{-x}$,

$$\begin{aligned} P(A^{\mathrm{c}}) &\leq \sum_{m\in\mathbb{N}} P\Big(\bigcap_{k\geq m} A_k^{\mathrm{c}}\Big) = \sum_{m\in\mathbb{N}} \lim_{n\to\infty} P\Big(\bigcap_{k=m}^{n} A_k^{\mathrm{c}}\Big) = \sum_{m\in\mathbb{N}} \lim_{n\to\infty} \prod_{k=m}^{n} (1 - P(A_k)) \\ &\leq \sum_{m\in\mathbb{N}} \lim_{n\to\infty} \exp\Big(-\sum_{k=m}^{n} P(A_k)\Big) = 0\,. \end{aligned}$$

□

2.3 Strong Law of Large Numbers (SLLN)

The weak law of large numbers (WLLN) in Theorem 2.2 alone doe snot quite fulfil our expectations. For example, if we flip a fair coin 100 times, then with some small probability it may happen that the relative frequency differs strongly from $\frac{1}{2}$, but this deviation should vanish gradually, provided we continue flipping the coin for long enough. This intuition is based on the notion of almost sure convergence. Recall from Definition 2.4 that a sequence $(Y_n)_{n\in\mathbb{N}}$ of $\mathbb{R}$-valued random variables converges to the random variable Y, all defined on some probability space $(\Omega, \mathcal{F}, P)$, P-almost surely if

$$P(\{\omega \in \Omega\colon Y_n(\omega) \to Y) = 1\,.$$

Exercise 2.9 Show that almost sure convergence implies convergence in probability but the converse is not true.

☕☕

Theorem 2.10 (Strong Law of Large Numbers (SLLN) - L^2 - version) *If $(X_i)_{i\in\mathbb{N}}$ is a sequence of pairwise uncorrelated $\mathbb{R}$-valued L^2 random variables on some probability space $(\Omega, \mathcal{F}, P)$ with $v := \sup_{i\in\mathbb{N}} \mathrm{Var}(X_i) < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \to 0\, P - \textit{almost surely if}\, n \to \infty\,.$$

Remark 2.11 For an L^1-version of the SLLN see [Dur19]. ◇

Proof. Without loss of generality we may assume that $\mathbb{E}[X_i] = 0$ for all $i \in \mathbb{N}$. Write $Y_n := 1/n\sum_{i=1}^n X_i$.

Step 1: Show that $Y_{n^2} \to 0$ almost surely as $n \to \infty$. For any $\varepsilon > 0$, Theorem 2.2 implies that

$$P(|Y_{n^2}| > \varepsilon) \leq v/n^2\varepsilon^2$$

and thus

$$\sum_{n\in\mathbb{N}} P(|Y_{n^2}| > \varepsilon) < \infty\,.$$

Now Borel-Cantelli Lemma, Theorem 2.8, implies that

$$P(\limsup_{n\to\infty}|Y_{n^2}| > \varepsilon) \leq P(|Y_{n^2}| > \varepsilon \text{ for infinitely many } n) = 0\,.$$

Thus we conclude with $P(\limsup_{n\to\infty}|Y_{n^2}| \not\to 0) = 0$.

Step 2: For $m \in \mathbb{N}$ let $n := n(m) \in \mathbb{N}$ be such that $n^2 \leq m < (n+1)^2$. We shall compare Y_m and Y_{n^2}. We write $S_k := kY_k = \sum_{i=1}^k X_i$. Chebychev's inequality 1.50 implies

$$P(|S_m - S_{n^2}| > \varepsilon n^2) \leq \varepsilon^{-2}n^{-4}\mathrm{Var}\Big(\sum_{n^2<i\leq m} X_i\Big) \leq \frac{v}{(n-m)}\varepsilon^2 n^4$$

and

$$\sum_{m\in\mathbb{N}} P(|S_m - S_{n^2}| > \varepsilon n^2\Big) \leq \frac{v}{\varepsilon^2}\sum_{n\in\mathbb{N}}\sum_{m=n^2}^{(n+1)^2-1}\Big(\frac{m-n^2}{n^4}\Big) = \frac{v}{\varepsilon^2}\sum_{n\in\mathbb{N}}\sum_{k=1}^{2n}\frac{k}{n^4}$$
$$= \frac{v}{\varepsilon^2}\sum_{n\in\mathbb{N}}\frac{2n(2n+1)}{2n^4} < \infty\,,$$

and so once again with Theorem 2.8, one obtains

$$P\Big(\Big|\frac{S_m}{n(m)^2} - Y_{n(m)^2}\Big| \underset{m\to\infty}{\longrightarrow} 0\Big) = 1\,,$$

and with Step 1 above we conclude that $P(S_n/n^2 \underset{m\to\infty}{\longrightarrow} 0) = 1$. Since $|Y_m| \leq |S_m|n(m)^2$, it follows that $P(Y_m \underset{m\to\infty}{\to} 0) = 1$. □

2.4 The Central Limit Theorem (CLT)

Let $(X_i)_{i\in\mathbb{N}}$ be a Bernoulli sequence with parameter $p \in (0,1)$. $S_n = \sum_{i=1}^n X_i$ represents the number of successes in n experiments. How much do the S_n fluctuate around their expectation np, i.e., what is the order of magnitude of the deviations $S_n - np$ in the limit $n \to \infty$?

Exercise 2.12 Suppose that $(a_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}_+$. Using Stirling's formula and Chebychev's inequality, show that

$$P\Big(\Big|S_n - np\Big| \leq a_n\Big) \underset{n\to\infty}{\longrightarrow} \begin{cases} 1 & \text{if } a_n/\sqrt{n} \to \infty \text{ as } n \to \infty\,, \\ 0 & \text{if } a_n/\sqrt{n} \to 0 \text{ as } n \to \infty\,. \end{cases}$$

Definition 2.13 (Convergence in distribution) Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}$-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. *Y_n converges in distribution to a real-valued random variable Y if $F_{Y_n}(c) \underset{n\to\infty}{\longrightarrow} F_Y(c)$ for all points $c \in \mathbb{R}$ at which F_Y is continuous.* We write $Y_n \overset{\mathrm{d}}{\underset{n\to\infty}{\longrightarrow}} Y$.

Proposition 2.14 *Under the above assumptions, the following statements are equivalent.*

(a) $Y_n \overset{\mathrm{d}}{\underset{n\to\infty}{\longrightarrow}} Y$.

(b) $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)]$ *as* $n \to \infty$ *for all* $f \in \mathcal{C}_\mathrm{b}(\mathbb{R};\mathbb{R})$.

If F_Y is continuous, then the following is also equivalent:

(c) F_{Y_n} *converges uniformly to* F_Y*, i.e.,* $\|F_{Y_n} - F_Y\| \to 0$.

Proof. Exercise. □

Exercise 2.15 (Fatou's lemma) Prove the following statement. Suppose $g\colon \mathbb{R} \to \mathbb{R}_+$ is continuous. If $Y_n \overset{\mathrm{d}}{\underset{n\to\infty}{\longrightarrow}} Y$, then

$$\liminf_{n\to\infty} \mathbb{E}[g(Y_n)] \geq \mathbb{E}[g(Y)]\,.$$

Theorem 2.16 *Let $Y, Y_1, Y_2, \ldots$ be $\mathbb{R}$-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. Then the following statements are equivalent.*

(a) $Y_n \overset{\mathrm{d}}{\underset{n\to\infty}{\longrightarrow}} Y$.

(b) For all open sets $G \subset \mathbb{R}$: $\liminf_{n\to\infty} P(Y_n \in G) \geq P(Y \in G)$.

(c) For all closed sets $K \subset \mathbb{R}$: $\limsup_{n\to\infty} P(Y_n \in K) \le P(Y \in K)$*.*

(d) For all $A \in \mathcal{B}$ with $P(Y \in \partial A) = 0$,

$$\lim_{n\to\infty} P(Y_n \in A) = P(Y \in A)\,.$$

The proof of Theorem 2.16 relies on the following theorem which we only cite here (this is standard result in measure theory). A sequence $(F_n)_{n\in\mathbb{N}}$ of probability distribution functions converges weakly to a limit F (written $F_n \underset{n\to\infty}{\overset{\text{w}}{\Rightarrow}} F$ or $F_n \overset{\text{w}}{\Rightarrow} F$) if $F_n(y) \to f(y)$ as $n \to \infty$ for all y that are continuity points of F. Notice that this convergence is equivalent to the convergence in distribution for the corresponding random variables.

Theorem 2.17 *If $F_n \overset{\text{w}}{\Rightarrow} F$, then there are random variables $Y, Y_n, n \in \mathbb{N}$, with distribution function F_n so that $Y_n \to Y$ almost surely as $n \to \infty$.*

Proof. See book by Durrett [Dur19], or Billlingsley [Bil99] or Bauer [BB01]. □

Proof of Theorem 2.16. (a) ⇒ (b): Let X_n have the same distribution as Y_n and $X_n \to X$ almost surely as $n \to \infty$. Since G is open,

$$\liminf_{n\to\infty} \mathbb{1}_G(X_n) \ge 1_G(X)\,,$$

so Fatou's Lemma implies

$$\liminf_{n\to\infty} P(X_n \in G) \ge P(X \in G)\,.$$

(b) ⇔ (c): This follows from: A is open ⇔ A^{c} is closed and $P(A) + P(A^{\text{c}}) = 1$.
(b) & (c) ⇒ (d): Let $K = \overline{A}$ and $G = \text{int}(A)$ be the closure and interior of A respectively. The boundary $\partial A = \overline{A} \setminus \text{int}(A)$ and $P(Y \in \partial A) = 0$ so

$$P(Y \in K) = P(Y \in A) = P(Y \in G)\,.$$

Using (b) & (c),

$$\limsup_{n\to\infty} P(Y_n \in A) \le \limsup_{n\to\infty} P(Y_n \in K) \le P(Y \in K) = P(Y \in A)$$
$$\liminf_{n\to\infty} P(Y_n \in A) \ge \liminf_{n\to\infty} P(Y_n \in G) \ge P(Y \in G) = P(Y \in A)\,.$$

(d) ⇒ (a): Let x be such that $P(Y = x) = 0$ and consider $A = (-\infty, x]$. □

Theorem 2.18 (Continuous mapping theorem) *Let $g\colon \mathbb{R} \to \mathbb{R}$ be measurable and define $D_g := \{x \in \mathbb{R} \colon g \text{ is discontinuous at } x\}$ and let $Y, Y_1, Y_2, \ldots$ be $\mathbb{R}$-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. If $Y_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} Y$ and $P(Y \in D_g) - 0$ then*

$$g(Y_n) \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} g(Y)\,.$$

If in addition g is bounded, then

$$\lim_{n\to\infty} \mathbb{E}[g(Y_n)] = \mathbb{E}[g(Y)]\,.$$

Proof. Let $X_n \stackrel{\mathrm{d}}{=} Y_n$ (i.e., X_n has same distribution as Y_n) with $X_n \to X$ almost surely as $n \to \infty$.If f is continuous and bounded then $D_{f\circ g} \subset D_g$ so $P(X \in D_{f\circ g}) = 0$ and it follows that $f(g(X_n)) \to f(g(X))$ almost surely as $n \to \infty$. Since $f \circ g$ is bounded the bounded convergence theorem implies

$$\lim_{n\to\infty} \mathbb{E}[f(g(X_n))] = \mathbb{E}[f(g(X))]\,.$$

As this holds for all $f \in \mathcal{C}_\mathrm{b}(\mathbb{R};\mathbb{R})$, we conclude with the statement. The second one is proved by taking $f(x) = x$ and consider cutoff parameter like, i.e., repeat the above arguments for $f_M = f \wedge f$, $M > 0$, followed by the limit $M \to \infty$.

□

Definition 2.19 (a) $m \in \mathbb{R}, v > 0$. The probability measure $\mathsf{N}(m, v) \in \mathcal{M}_1(\mathbb{R}, \mathcal{B})$ with density function

$$\varphi_{m,v}(x) = \frac{1}{\sqrt{2\pi v}}\mathrm{e}^{-(x-m)^2/2v}\,, \quad , x \in \mathbb{R}\,.$$

is called *the normal distribution or the Gauss distribution with mean* m *and variance* v We write $\varphi \equiv \varphi_{0,1}$.

(b) For $c \in \mathbb{R}$ denote

$$\Phi(c) = \int_{-\infty}^{c} \varphi(x)\,\mathrm{d}x = \mathsf{N}(0,1)((-\infty, c])$$

the *cummulative distribution function of* $\mathsf{N}(0,1)$.

Theorem 2.20 (The Central Limit Theorem (CLT)) *Let* $(X_i)_{i\in\mathbb{N}}$ *be a sequence of independent, identically distributed* $\mathbb{R}$*-valued random variables in* L^2 *on some probability space* $(\Omega, \mathcal{F}, P)$ *with* $\mathbb{E}[X_i] = m$, $\mathrm{Var}(X_i) = v > 0$. *Then,*

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}} \xrightarrow[n\to\infty]{\mathrm{d}} Z \sim \mathsf{N}(0,1)\,,$$

where $Z \sim N(0,1)$ *means the random variable* Z *is normally distributed with* $\mathsf{N}(0,1)$*, meaning that* $\|F_{S_n^*} - \Phi\| \to 0$ *as* $n \to \infty$.

Definition 2.21 (Convolution) Let $Q_1, Q_2 \in \mathcal{M}_1(\mathbb{R})$ probability measures with densities (Radon-Nikodym density with respect to the Lebesgue measure) ϱ_1 and ϱ_2, respectively, then the *convolution* $Q_1 * Q_2 \in \mathcal{M}_1(\mathbb{R})$ has the density (with respect to the Lebesgue measure)

$$\varrho_1 * \varrho_2(x) = \int \varrho_1(y)\varrho_2(x-y)\,\mathrm{d}y\,, \quad x \in \mathbb{R}\,.$$

The convolution $Q_1 * Q_2$ is defined as the image measure under the addition mapping $A\colon \mathbb{R}^2 \to \mathbb{R}, (x_1, x_2) \mapsto A(x_1, x_2) = x_1 + x_2$.

$$Q_1 * Q_2 = (Q_1 \otimes Q_2) \circ A^{-1}$$

is called *the convolution of* Q_1 *and* Q_2.

Exercise 2.22 Show that

$$\mathsf{N}(m_1, v_1) * \mathsf{N}(m_2, v_2) = \mathsf{N}(m_1 + m_2, v_1 + v_2)\,.$$

Proof of Theorem 2.20. Without loss of generality let $m = 0, v = 1$. We shall show that

$$\mathbb{E}[f \circ S_n^*] \to \mathbb{E}_{\mathsf{N}(0,1)}[f] \quad \text{as } n \to \infty \text{ for all } f \in \mathcal{C}_\mathrm{b}(\mathbb{R}; \mathbb{R})\,.$$

For this endeavour we can assume that $f \in \mathcal{C}^2(\mathbb{R}; \mathbb{R})$ with bounded and uniformly continuous derivatives f' and f'' as in the proof of the convergence in distribution one can approximate the indicator function $1\!\!1_{(-\infty,c]}$ by such functions f. For our proof pick another random sequence, that is, let $(Y_i)_{i\in\mathbb{N}}$ be an independent identically distributed sequence of standard normal random variables $Y_i \sim \mathsf{N}(0,1)$, and assume that this sequence is independent of $(X_i)_{i\in\mathbb{N}}$. Then it is easy to see that

$$T_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \sim \mathsf{N}(0,1)\,,$$

so we shall show that

$$\lim_{n\to\infty} |\mathbb{E}[f \circ S_n^* - f \circ T_n^*]| = 0\,.$$

For this we write the difference $f \circ S_n^* - f \circ T_n^*$ as a telescoping sum with the following notations: $X_{i,n} := X_i/\sqrt{n}, Y_{i,n} := Y_i/\sqrt{n}$ and

$$W_{i,n} := \sum_{j=1}^{i-1} Y_{j,n} + \sum_{j=i+1}^{n} X_{i,n}\,.$$

Then

$$f \circ S_n^* - f \circ T_n^* = \sum_{i=1}^n \Big(f(W_{i,n} + X_{i,n}) - f(W_{i,n} + Y_{i,n}) \Big)\,.$$

Taylor approximation:

$$f(W_{i,n}+X_{i,n}) = f(W_{i,n}) + f'(W_{i,n})X_{i,n} + \frac{1}{2}f''(W_{i,n})X_{i,n}^2 + R_{X,i,n}$$

with remainder term

$$R_{X,i,n} := \frac{1}{2}X_{i,n}^2\Big(f''(W_{i,n}+\theta X_{i,n}) - f''(W_{i,n})\Big)\,; \qquad \text{for some } \theta\in[0,1]\,.$$

To estimate the remainder term note first that

$$|R_{X,i,n}| \le X_{i,n}^2\|f^{''}\|\,.$$

As $f^{''}$ is uniformly continuous, for every $\varepsilon>0$ one can find $\delta>0$ such that $|R_{X,i,n}|\le X_{i,n}^2\varepsilon$ for $|X_{i.n}|\le\delta$. This yields

$$|R_{X,i,n}| \le X_{i,n}^2(\varepsilon 1\!\!1_{\{|X_{i,n}|\le\delta\}} + \|f^{''}\|1\!\!1_{\{|X_{i,n}|>\delta\}})\,.$$

A similar Taylor approximation holds for all $f(W_{i,n}+Y_{i,n})$. We insert our Taylor approximations and take the expectation, and using that

$$\mathbb{E}[X_{i,n}] = \mathbb{E}[Y_{i,n}] = 0; \mathbb{E}[X_{i,n}^2] = \frac{1}{n} = \mathbb{E}[Y_{i,n}^2]$$
$$\mathbb{E}[f^{''}(W_{i,n})(X_{i,n}^2 - Y_{i,n}^2)] = \mathbb{E}[f^{''}(W_{i,n})]\mathbb{E}[X_{i,n}^2 - Y_{i,n}^2] = 0\,,$$

we obtain

$$\begin{aligned}|\mathbb{E}[f\circ S_n^* = f\circ T_n^*]| &\le \sum_{i=1}^n \mathbb{E}[|R_{X,in}| + |R_{Y,i,n}|]\\ &\le \sum_{i=1}^n\Big(\varepsilon\mathbb{E}[X_{i,n}^2+Y_{i,n}^2] + \|f^{''}\|\mathbb{E}[X_{i,n}^2 1\!\!1_{\{|X_{i,n}|>\delta\}} + Y_{i,n}^2 1\!\!1_{\{|Y_{i,n}|>\delta\}}\Big)\\ &= 2\varepsilon + \|f^{''}\|\mathbb{E}[X_1^2 1\!\!1_{\{|X_1|>\delta\sqrt{n}\}} + Y_1^2 1\!\!1_{\{|Y_1|>\delta\sqrt{n}\}}]\,.\end{aligned}$$

Note that

$$\mathbb{E}[X_1^2 1\!\!1_{\{|X_1|>\delta\sqrt{n}\}}] = 1 - \mathbb{E}[X_1^2 1\!\!1_{\{|X_1|\le\delta\sqrt{n}\}}] \;\to 0 \text{ as } n\to\infty\,,$$

and similarly,

$$\mathbb{E}[Y_1^2 1\!\!1_{\{|X_1|>\delta\sqrt{n}\}}] \to 0 \text{ as } n\to\infty\,.$$

Hence

$$\limsup_{n\to\infty}|\mathbb{E}[f\circ S_n^* = f\circ T_n^*]| \le 2\varepsilon\,.$$

□

We now introduce characteristic function and Fourier transform for probability measures. At the end of this we shall come up with an alternative proof of the CLT.

Definition 2.23 (Characteristic function) Let X be a $\mathbb{R}$-valued random variable defined on some probability space $(\Omega,\mathcal{F},P)$.

$$\varphi_X(t) := \mathbb{E}[\mathrm{e}^{\mathrm{i}tX}] \quad t\in\mathbb{R}\,,$$

is called the *characteristic function of* X.

Proposition 2.24 (Properties) *Let X be a $\mathbb{R}$-valued random variable defined on some probability space $(\Omega, \mathcal{F}, P)$. Then the following properties hold.*

(a) $\varphi_X(0) = 1$.

(b) $\varphi_X(-t) = \overline{\varphi_X(t)}$.

(c) $|\varphi_X(t)| \le 1$.

(d)
$$\mathbb{E}[\mathrm{e}^{\mathrm{i}t(aX+b)}] = \mathrm{e}^{\mathrm{i}tb}\varphi_X(at)\,, \quad a, b \in \mathbb{R}\,.$$

(e) $\varphi_{-X} = \varphi_X$.

(f) Suppose and Y are independent $\mathbb{R}$-valued random variables. Then

$$\varphi_{X+Y} = \varphi_X \varphi_Y\,.$$

Proof. These elementary properties are straightforward to prove and are left as an exercise. See e.g. [Dur19, Bil12]. □

Lemma 2.25 $(\mu_n)_{n\in\mathbb{N}}, \mu_n \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$. *Assume that there is a $\mu \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$ such that the μ_n's converge in the sense that*

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} f \,\mathrm{d}\mu_n = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu \quad \text{for all } f \in \mathcal{C}_{\mathrm{c}}^\infty(\mathbb{R}^d; \mathbb{C})\,.$$

Then the following holds.

(a) For any $f \in \mathcal{C}(\mathbb{R}^d; [0, \infty))$ one has

$$\int_{\mathbb{R}^d} f(y)\,\mu(\mathrm{d}y) \le \liminf_{n\to\infty} \int_{\mathbb{R}^d} f(y)\,\mu_n(\mathrm{d}y)\,. \tag{2.2}$$

(b) If $f \in \mathcal{C}(\mathbb{R}^d; \mathbb{C})$ satisfies

$$\lim_{R\to\infty} \sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} |f(y)| \mathbb{1}_{\{|f|\ge R\}}(y)\,\mu_n(\mathrm{d}y) = 0\,, \tag{2.3}$$

then f is μ - integrable and

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} f \,\mathrm{d}\mu_n = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu\,. \tag{2.4}$$

(c) (2.4) *holds for any $f \in \mathcal{C}(\mathbb{R}^d; \mathbb{C})$ satisfying*

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} |f|^{1+\alpha} \,\mathrm{d}\mu_n < \infty \quad \text{for some } \alpha \in (0, \infty)\,.$$

Proof. We first show (2.4) for every $f \in \mathcal{C}_b(\mathbb{R}^d;\mathbb{C})$. To this end, we choose $\varrho \in \mathcal{C}_c^\infty(B(0,1);[0,\infty))$ so that $\int_{\mathbb{R}^d} \varrho(y)\,dy = 1$. Here, $B(0,1) \subset \mathbb{R}^d$ is the open ball around 0 with radius 1. Define

$$f_k(x) := k \int_{|y|\le k} \varrho(k(x-y)) f(y)\,dy \qquad \text{for } k \in \mathbb{N}\,.$$

Clearly, $f_k \in \mathcal{C}_c^\infty(\mathbb{R}^d;\mathbb{C})$ and $\|f_k\|_\infty \le \|f\|_\infty$. Thus

$$\begin{aligned}\limsup_{n\to\infty} \left| \int f\,d\mu_n - \int f\,d\mu \right| &\le \limsup_{k\to\infty} \limsup_{n\to\infty} \int |f - f_k|\,d\mu_n \\ &\le 2\|f\|_\infty \limsup_{n\to\infty} \mu_n(B(0,R)^c)\end{aligned}$$

for every $R \in (0,\infty)$. For any such $R \in (0,\infty)$ we can choose a function $g_R \in \mathcal{C}_c^\infty(B(0,R);[0,1])$ so that $g_R = 1$ on the ball $B(0,R/2)$, and therefore,

$$\limsup_{n\to\infty} \mu_n(B(0,R)^c) \le 1 - \liminf_{n\to\infty} \int g_R\,d\mu_n \le \mu(B(0,R/2)^c) \to 0 \text{ as } R \to \infty\,.$$

We have now proved (2.4) for every $f \in \mathcal{C}_b(\mathbb{R}^d;\mathbb{C})$.
We now show (2.2): We simply set $f_R = f \wedge R$, $R \in (0,\infty)$, for any nonnegative continuous function f on $\mathbb{R}^d$. Then by the Monotone Convergence Theorem,

$$\int f\,d\mu = \lim_{R\uparrow\infty} \int f_R\,d\mu = \lim_{R\uparrow\infty} \lim_{n\to\infty} \int f_R\,d\mu_n \le \liminf_{n\to\infty} \int f\,d\mu_n\,.$$

To prove (2.4) for $f \in \mathcal{C}(\mathbb{R}^d;\mathbb{C})$ satisfying (2.3), it suffices to handle the case when the function f is nonnegative. For nonnegative f satisfying (2.3) we know from (2.2) that f is μ-integrable. Hence, we any $\varepsilon > 0$, we can choose an $R \in (0,\infty)$ so that for the continuous and bounded function $f_R := f \wedge R$ we get

$$\sup_{n\in\mathbb{N}} \int |f(y) - f_R(y)|\,\mu_n(dy) \vee \int |f(y) - f_R(y)|\,\mu(dy) < \varepsilon\,.$$

Thus

$$\begin{aligned}\left| \int f\,d\mu_n - \int f\,d\mu \right| &= \left| \int (f - f_R)\,d\mu_n - \int (f - f_R)\,d\mu - \int (f_R - f_R)\,d\mu_n \right| \\ &\le 2\varepsilon + \left| \int (f_R\,d\mu_n - \int f_R\,d\mu \right|,\end{aligned}$$

and (2.4) for the f_R's simply implies (2.4) for the given f. □

As Lemma 2.25 makes explicit, to test whether (2.4) holds for all $f \in \mathcal{C}_b(\mathbb{R}^d;\mathbb{C})$ requires only that we test it for all $f \in \mathcal{C}_c^\infty(\mathbb{R}^d;\mathbb{C})$. In conjunction with elementary Fourier analysis, this means that we need only test it for f's which are imaginary exponential functions.

Definition 2.26 Let $\mu \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$. The *characteristic function $\widehat{\mu}$ of μ is its Fourier transform* given by

$$\widehat{\mu}(k) := \int_{\mathbb{R}^d} \exp\left(\mathrm{i}\langle k, x\rangle\right) \mu(\mathrm{d}x)\,, \quad k \in \mathbb{R}^d\,.$$

For $f \in L^1(\mathbb{R}^d; \mathbb{C})$ we use

$$\widehat{f}(k) := \int_{\mathbb{R}^d} \exp\left(\mathrm{i}\langle , x\rangle\right) f(x)\, \mathrm{d}x\,, \quad k \in \mathbb{R}^d\,.$$

to denote its Fourier transform.

Remark 2.27 Clearly, the Fourier transform $\widehat{\mu}$ of $\mu \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$ is a continuous function which is bounded by one. Furthermore, for $f \in \mathcal{C}_c^\infty(\mathbb{R}^d; \mathbb{C})$, $\widehat{f} \in \mathcal{C}^\infty(\mathbb{R}^d; \mathbb{C})$ and $\widehat{f}$ as well as all its derivatives are rapidly decreasing. ◇

Lemma 2.28 *Let $\mu \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$. Then the following holds.*

(a) For every $f \in \mathcal{C}_b(\mathbb{R}^d; \mathbb{C}) \cap L^1(\mathbb{R}^d; \mathbb{C})$ with $\widehat{f} \in L^1(\mathbb{R}^d; \mathbb{C})$,

$$\int_{\mathbb{R}^d} f \,\mathrm{d}\mu = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(k)\widehat{\mu}(-k)\, \mathrm{d}k\,. \tag{2.5}$$

(b) Let $(\mu_n)_{n\in\mathbb{N}}, \mu_n \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}^d)$, be given. Then (2.4) *holds for every $f \in \mathcal{C}(\mathbb{R}^d; \mathbb{C})$ satisfying* (2.3) *if and only if*

$$\widehat{\mu}_n(k) \to \widehat{\mu}(k) \text{ as } n \to \infty \text{ for every } k \in \mathbb{R}^d\,.$$

Proof. (a) We shall use a mollifier as follows. Pick an even function $\varrho \in \mathcal{C}_c^\infty(\mathbb{R}^d; [0, \infty))$ with $\int_{\mathbb{R}^d} \varrho(x)\, \mathrm{d}x = 1$, and set $\varrho_\varepsilon(x) := \varepsilon^{-d}\varrho(\varepsilon^{-1}x)$ for $\varepsilon \in (0, \infty)$. We now define the convolution of our measure μ with the chosen mollifier,

$$\psi_\varepsilon(x) := \int_{\mathbb{R}^d} \varrho_\varepsilon(x - y)\, \mu(\mathrm{d}y)\,, \quad x \in \mathbb{R}^d\,.$$

Then it is easy to see that $\psi_\varepsilon \in \mathcal{C}_b(\mathbb{R}^d, \mathbb{R})$ and $\|\psi_\varepsilon\|_{L^1(\mathbb{R}^d)} = 1$ for every $\varepsilon \in (0, \infty)$. By Fubini's Theorem (see [Dur19, BB01, Coh13]) we immediately see that $\widehat{\psi_\varepsilon}(k) = \widehat{\varrho}(\varepsilon k)\widehat{\mu}(k)$. For any $f \in \mathcal{C}_b(\mathbb{R}^d; \mathbb{C}) \cap L^1(\mathbb{R}^d; \mathbb{C})$ write $f_\varepsilon := \varrho_\varepsilon * f$ for the convolution of ϱ_ε with f. Thus, Fubini's Theorem followed by the Classical Parseval Identity yields

$$\int_{\mathbb{R}^d} f_\varepsilon \,\mathrm{d}\mu = \int_{\mathbb{R}^d} f(x)\psi_\varepsilon(x)\, \mathrm{d}x = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{\varrho}(\varepsilon k)\widehat{f}(k)\widehat{\mu}(-k)\, \mathrm{d}k\,.$$

Since, as $\varepsilon \to 0$, $f_\varepsilon \to f$ while $\widehat{\varrho}(\varepsilon k) \to 1$ boundedly and pointwise, (2.5) now follows from Lebesgue's Dominated Convergence Theorem.

(b) By Lemma 2.25, we need only check (2.4) when $f \in \mathcal{C}_c^\infty(\mathbb{R}^d; \mathbb{C})$. For such f, $\widehat{F}$ is smooth and rapidly decreasing, and therefore the result follows immediately from (a) together with Lebesgue's Dominated Convergence Theorem. □

Remark 2.29 (Alternative proof of the CLT) Suppose $m = 0$ and $v = 1$ in our setting of Theorem 2.20 and denote the distribution of S_N^* by μ_N. Then,

$$\widehat{\mu}_N(k) = \Big(\widehat{\mu}\Big(\frac{k}{\sqrt{N}}\Big)\Big)^N = \Big(1 - \frac{k^2}{2N} + o(1/N)\Big)^N \to \mathrm{e}^{-\frac{k^2}{2}} \text{ as } N \to \infty\,, \quad k \in \mathbb{R}\,.$$

Then simply observe that $\varphi_X(k) = \mathrm{e}^{-k^2/2}$ when $X \sim \mathsf{N}(0, 1)$.

◇

2.5 Convergence of Probability Measures

We shall develop convergence theory on the set of probability measures over some measurable space $(E, \mathcal{E})$. We first start with a standard approach to convergence, namely via bounded measurable functions. We then come up with a notion which reflects better the topology over the underlying space E. In the following let $\mathcal{B}(E; \mathbb{R}) \equiv \mathcal{B}((E, \mathcal{E}); \mathbb{R})$ be the space of bounded, $\mathbb{R}$-valued, $\mathcal{E}$-measurable functions on E, use $\mathcal{M}_1(E) \equiv \mathcal{M}_1(E, \mathcal{E})$ to denote the space of all probability measures on $(E, \mathcal{E})$. The duality between $\mathcal{B}(E; \mathbb{R})$ and $\mathcal{M}_1(E)$ is given by

$$\langle f, \mu\rangle := \int_E f \,\mathrm{d}\mu\,, \quad \mu \in \mathcal{M}_1(E), f \in \mathcal{B}(E; \mathbb{R})\,.$$

In the following, let $\mathcal{B}_1(E) := \{f \in \mathcal{B}(E; \mathbb{R})\colon \|f\|_\infty \le 1\}$ the ball of bounded measurable functions with supremum norm less equal to 1, where $\|f\|_\infty := \sup_{x\in E}\{|f(x)|\}$. For every $\mu \in \mathcal{M}_1(E)$ a neighbourhood basis of μ is given by the sets

$$U(\mu, \delta) = \Big\{\nu \in \mathcal{M}_1(E)\colon \sup_{f\in\mathcal{B}_1(E)} |\langle f, \nu\rangle - \langle f, \mu\rangle| < \delta\Big\}\,,$$

where $\delta \in (0, \infty)$. The topology defined by these sets is called the *uniform topology on* $\mathcal{M}_1(E)$. The next lemma shows that this definition is leading to a metric.

Lemma 2.30 *Define*

$$\|\mu - \nu\|_{\text{var}} := \sup\Big\{|\langle f, \mu\rangle - \langle f, \nu|\colon f \in \mathcal{B}_1\Big\}\,.$$

Then $(\mu, \nu) \in \mathcal{M}_1(E)^2 \mapsto \|\mu - \nu\|_{\text{var}}$ *is a metric on* $\mathcal{M}_1(E)$ *which is compatible with the uniform topology.*

Proof. The interested reader may check [BB01, Bil99]. □

Remark 2.31 Suppose E is uncountable, $\{x\} \in \mathcal{E}$, then the point masses $\delta_x, x \in E$, form an uncountable subset of $\mathcal{M}_1(E)$. Furthermore,

$$\|\delta_x - \delta_y\|_{\text{var}} = 2 \qquad \text{for } x \neq y\,.$$

Hence, in this case $\mathcal{M}_1(E)$ cannot be covered by a countable collection of open $\|\cdot\|_{\text{var}}$-balls of radius 1. In addition, we shall find a topology for which the point masses are close when the corresponding point are close in the underlying space. ◇

In a first step one can eliminate the uniformity in the definition of the uniform topology. For $\mu\mathcal{M}_1(E,\mathcal{E})$ a neighbourhood basis is given by the sets

$$S(\mu,\delta;f_1,\ldots,f_n) := \Big\{\nu\in\mathcal{M}_1(E)\colon \max_{1\le k\le n}|\langle f_k,\nu\rangle-\langle f_k,\mu\rangle|<\delta\Big\}\,, \qquad (2.6)$$

$\delta>0$, $n\in\mathbb{N}$, $f_1,\ldots,f_n\in|Bcal)E;\mathbb{R})$. The topology define d by these neighbourhoods is called the *strong topology* or the τ*-topology* . Here, a net $\{\mu_\alpha\colon\alpha\in A\}$ converges to μ if and only if

$$\lim_\alpha\langle f,\mu_\alpha\rangle=\langle f,\mu\rangle\,.$$

Again, this topology cannot distinguish point masses δ_x and δ_y when x and y are very close. The next idea is for metric spaces $(E,\mathrm{d}_)$ (net convergence can then be replaced by convergence) is consider (2.6) for test functions $f\in\mathcal{C}_\mathrm{b}(E;\mathbb{R})$.

Definition 2.32 (a) Let (E,d) be a metric space and $\mathcal{E}$ the Borel-σ-algebra, $(\mu_n)_{n\in\mathbb{N}}$ a sequence of probability measures $\mu\in\mathcal{M}_1(E)$. The $(\mu_n)_{n\in\mathbb{N}}$ converges *weakly* to $\mu\in\mathcal{M}_1(E)$ as $n\to\infty$, in symbols , $\mu_n \overset{\mathrm{w}}{\underset{n\to\infty}{\Rightarrow}} \mu$ or $\mu_n \underset{n\to\infty}{\Rightarrow} \mu$, if

$$\langle f,\mu_n\rangle\to\langle f,\mu\rangle \;\text{ as }\; n\to\infty \;\text{ for all } f\in\mathcal{C}_{rmb}(E;\mathbb{R})\,.$$

(b) Let $(X_n)_{n\in\mathbb{N}}$ and X be $\mathbb{R}$-valued random variables defined on some probability space $(\Omega,\mathcal{F},P)$. X_n converges in distribution to X as $n\to\infty$, in symbols $X_n \overset{\mathrm{d}}{\underset{n\to\infty}{\Rightarrow}} X$, if

$$\mathbb{E}[f(X_n)]\to\mathbb{E}[f(X)] \quad \text{as } n\to\infty \text{ for all } f\in\mathcal{C}_\mathrm{b}(\mathbb{R};\mathbb{R})\,.$$

Exercise 2.33 Show that $\delta_y \overset{\mathrm{w}}{\Rightarrow} \delta_x$ if and only if $y\to x$ in (E,d).

Definition 2.34 **[Tight set]** Let (E,d) be a separable metric space. A subset $M\subset\mathcal{M}_1(E)$ is *tight* if, for every $\varepsilon>0$, there exists a compact set $K\subset E$ so that

$$\mu(K)\ge 1-\varepsilon \qquad \text{for every } \mu\in M\,.$$

Definition 2.35 (Tight family) A family $(X_t)_{t\in T}$ of $\mathbb{R}^d$-valued random variables defined on some probability space $(\Omega,\mathcal{F},P)$ is *tight* if

$$\lim_{r\to\infty}\sup_{t\in T}P(|X_t|>r)=0\,.$$

A sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^d$-valued random variables is tight if

$$\lim_{r\to\infty}\limsup_{n\to\infty}P(|X_n|>r)=0\,.$$

Lemma 2.36 *Suppose $(X_n)_{n\in\mathbb{N}}$ and X are $\mathbb{R}^d$-valued random variables on some probability space $(\Omega, \mathcal{F}, P)$ and that $X_n \underset{n\to\infty}{\overset{d}{\Rightarrow}} X$. Then $(X_n)_{n\in\mathbb{N}}$ is tight.*

Proof. Fix $r > 0$, and define the continuous and bounded function $f(x) := (1 - (r - |x|)_+)_+ \, , x \in \mathbb{R}$. Then, using the definition of f and taking expectation, we get

$$\limsup_{n\to\infty} P(|X_n| > r) \leq \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \leq P(|X| > r - 1)\,.$$

Now letting $r \to \infty$ to conclude with the statement.

□

Lemma 2.37 *Let $(X_n)_{n\in\mathbb{N}}$ be $\mathbb{R}^d$-valued random variables on some probability space $(\Omega, \mathcal{F}, P)$. Then*

$$(X_n)_{n\in\mathbb{N}} \textit{ is tight } \Leftrightarrow c_n X_n \underset{n\to\infty}{\overset{\mathrm{P}}{\longrightarrow}} 0$$

for any sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n \geq 0$ and $c_n \to 0$ as $n \to \infty$.

Proof. Suppose that $(X_n)_{n\in\mathbb{N}}$ is tight and fix $r, \varepsilon > 0$. Then there is $n_0 \in \mathbb{N}$ such that $c_n r \leq \varepsilon$ for all $n \geq n_0$. Thus

$$\limsup n \to \infty P(|c_n X_n| > \varepsilon) \leq \limsup_{n\to\infty} P(|X_n| > r)$$

and $P(|c_n X_n| > \varepsilon) \to 0$ as $r \to \infty$. Conversely, if $(X_n)_{n\in\mathbb{N}}$ is not tight, we may choose $(n_k)_{k\in\mathbb{N}}$ so that $\inf_{k\in\mathbb{N}} P(|X_{n_k}| > k) > 0$. Rhen $c_n := \sup\{k^{-1} \colon n_k \geq n\}$ defines a zero sequence but $P(|c_n X_{n_k}| > 1) \not\to 0$ as $n \to \infty$. □

The following theorem provides useful conditions equivalent to the weak convergence of probability measures.

Definition 2.38 Suppose (E, d) is a metric space with Borel σ-algebra $\mathcal{E}$ and let $\mu \in \mathcal{M}_1(E)$. A set $A \in \mathcal{E}$ whose boundary ∂A satisfies $\mu(\partial A) = 0$ is called *μ-continuity set.*

Proposition 2.39 (Weak convergence - The Portmanteau Theorem, 1956) *Let (E, d) be a separable metric space and $\mathcal{B}(E)$ the Borel-σ-algebra. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures $\mu_n \in \mathcal{M}_1(E)$, $\mu \in \mathcal{M}_1(E)$.Then the following statements are equivalent.*

(a) $\mu_n \Rightarrow \mu$ *as* $n \to \infty$.

(b) For all $F \subset E$ closed, $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$.

(c) For all $G \subset E$ open, $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$.

(d) For every upper semicontinuous function $f \colon E \to \mathbb{R}$ which is bounded above,

$$\limsup_{n\to\infty} \langle f, \mu_n \rangle \leq \langle f, \mu \rangle\,.$$

(e) For every lower semicontinuous function $f\colon E \to \mathbb{R}$ which is bounded below,

$$\liminf_{n\to\infty} \langle f, \mu_n \rangle \geq \langle f, \mu \rangle .$$

(f) For every $f \in \mathcal{B}(E;\mathbb{R})$ which is continuous at μ-almost every $x \in E$, $\langle f, \mu_n \rangle \to \langle f, \mu \rangle$ as $n \to \infty$.

(g) $\mu_n(A) \to \mu(A)$ as $n \to \infty$ for all μ-continuity sets $A \in \mathcal{B}(E)$.

Remark 2.40 (a) If (E, d) is metric space, then convergence is net convergence when the space is not separable.

(b) Suppose that $x_n \to x_0$ in E as $n \to \infty$, so that $\delta_{x_n} \Rightarrow \delta_{x_0}$ as $n \to \infty$. Furthermore, suppose all x_n are all distinct from x_0 (e.g., $x_0 = 0, x_n = 1/n$). Then the inequality in (b) is strict if $F = \{x_0\}$, and the inequality in (c) is strict if $g = \{x_0\}^{\mathrm{c}}$. If $A = \{x_0\}$, then convergence does not hold in (g); but this does not contradict the theorem, because the limit measure of $\partial\{x_0\} = \{x_0\}$ is 1, not 0.

◇

Proof of Proposition 2.39. (a) ⇒ (b): Let $F \subset E$ be closed and define

$$\psi_k(x) = 1 - \Big(\frac{\mathrm{d}(x,F)}{1+\mathrm{d}(x,F)}\Big)^{\frac{1}{k}}, \quad k \in \mathbb{N}, x \in E .$$

Then ψ_k is uniformly continuous and bounded and

$$1 \geq \psi_k(x) \searrow 1_F(x) \quad \text{as } k \to \infty \quad \text{for each } x \in E .$$

Thus, countable additivity followed by (a) imply that

$$\mu(F) = \lim_{k\to\infty} \langle \psi_k, \mu \rangle = \lim_{k\to\infty} \lim_{n\to\infty} \langle \psi_k, \mu_n \rangle \geq \limsup_{n\to\infty} \mu_n(F) .$$

(b) ⇔ (c) and (d) ⇔ (e) and (f) ⇒ (a) are all trivial and left as an exercise.

(c) ⇒ (e) (then (b) ⇒ (d) follows similarly): With loss of generality assume that f is a nonnegative, lower semicontinuous function. For $k \in \mathbb{N}$, define

$$f_k = \sum_{\ell=0}^{\infty} \frac{\ell \wedge 4^k}{2^k} 1\!\!1_{I_{\ell,k}} \circ f = \frac{1}{2^k} \sum_{\ell=0}^{4^k} 1\!\!1_{J_{\ell,k}} \circ f ,$$

where

$$I_{\ell,k} + \Big(\frac{\ell}{2^k}, \frac{\ell+1}{2^k}\Big] \text{ and } J_{\ell,k} = \Big(\frac{\ell}{2^k}, \infty\Big) .$$

Clearly, $0 \leq f_k \nearrow f$ and therefore $\langle f_k, \mu \rangle \to \langle f, \mu \rangle$ as $k \to \infty$. We thus apply (c) to the open sets $\{f \in J_{\ell,k}\}$ and use lower semicontinuity to get

$$\langle f_k, \mu \rangle \leq \liminf_{n\to\infty} \langle f_k, \mu_n \rangle \leq \liminf_{n\to\infty} \langle f, \mu_n \rangle$$

for each $k \in \mathbb{N}$; and so, after letting $k \to \infty$, we have shown (c) $\Rightarrow$ (e).

(d) & (e) $\Rightarrow$ (f): Suppose $f \in \mathcal{B}(E;\mathbb{R})$ is continuous at μ-almost every $x \in E$, and define

$$\underline{f}(x) = \liminf_{y\to x} f(y) \quad \text{and} \quad \overline{f}(x) = \limsup_{y\to x} f(y) \quad \text{for } x \in E\,.$$

Then $\underline{f} \le f \le \overline{f}$ everywhere and equality holds μ almost surely. Furthermore, $\underline{f}$ is lower semicontinuous, $\overline{f}$ is upper semicontinuous, and both are bounded. Hence, by (d) and (e),

$$\begin{aligned}\limsup_{n\to\infty}\langle f, \mu_n\rangle \le \limsup_{n\to\infty}\langle \overline{f}, \mu_n\rangle \le \langle \overline{f}, \mu\rangle = \langle f, \mu\rangle = \langle \underline{f}, \mu\rangle \\ \le \liminf_{n\to\infty}\langle \underline{f}, \mu_n\rangle \le \liminf_{n\to\infty}\langle f, \mu_n\rangle\,.\end{aligned}$$

(c) & (c) $\Rightarrow$ (g): For $A \in \mathcal{E}$ denote $\text{int}(A)$ the interior of A and $\overline{A}$ the closure of A. The, (b) and (c) together imply

$$\begin{aligned}\mu(\overline{A}) \ge \limsup_{n\to\infty}\mu_n(\overline{A}) \ge \limsup_{n\to\infty}\mu_n(A) \ge \liminf_{n\to\infty}\mu_n(A) \\ \ge \liminf_{n\to\infty}\mu_n(\text{int}(A)) \ge \mu(\text{int}(A))\,.\end{aligned}$$

If A is a μ-continuity set, we conclude with (g).

(g) $\Rightarrow$ (a): With loss of generality we may assume that $f \in \mathcal{C}_\text{b}(E;\mathbb{R})$ satisfies $0 < f < 1$. Then

$$\langle f, \mu\rangle = \int_0^\infty \mu(f > t)\,\text{d}t = \int_0^1 \mu(f > t)\,\text{d}t\,,$$

and similarly for μ_n. As f is continuous, $\partial\{f > t\} \subset \{f = t\}$and hence $\{f > t\}$ is a μ-continuity set except for countably many t. Thus by (g) and bounded convergence,

$$\langle f, \mu_n\rangle = \int_0^1 \mu_n(f > t)\,\text{d}t \to \int_0^1 \mu(f > t)\,\text{d}t = \langle f, \mu\rangle \quad \text{as} \quad n \to \infty\,.$$

□

Theorem 2.41 *Let (E, d) be a Polish space and $\mu \in \mathcal{M}_1(E)$. Then for every $\varepsilon > 0$, there exists a compact set $K \subset E$ such that $\mu(K) \ge 1 - \varepsilon$.*

Proof. Let $\{p_k\}_{k\in\mathbb{N}}$ be a countable dense subset of E. Given $\mu \in \mathcal{M}!(E)$ and $\varepsilon > 0$, we can choose, for each $n \in \mathbb{N}$, an $\ell_n \in \mathbb{N}$ so that

$$\mu\Big(\bigcup_{k=1}^{\ell_n} B_E(p_k, 1/n)\Big) \ge 1 - \frac{\varepsilon}{2^n}\,.$$

Set

$$C_n := \bigcup_{k=1}^{\ell_n} \overline{B_E(p_k, 1/n)} \qquad \text{and} \quad K := \bigcap_{n=1}^{\infty} C_n\,.$$

By construction, we see that K is closed, and furthermore

$$K \subset \bigcup_{k=1}^{\ell_n} B_E(p_k, 2/n) \qquad \text{for all } n \in \mathbb{N}\,.$$

Thus K is totally bounded, and therefore K is compact. □

As Theorem 2.41 makes clear, probability measures on a Polish space like to be *nearly concentrated on a compact set*. The following fundamental theorem demonstrates the connection between tightness and relative compactness. We are not proving this theorem as the proof is relatively long and technical, the interested reader can find a proof in [Kal02, Bil99].

Theorem 2.42 *Let (E, d) be a separable metric space and $M \subset \mathcal{M}_1(E)$. Then $\overline{M}$ is compact if M is tight. Conversely, when E is Polish, M is tight if $\overline{M}$ is compact.*

We have now learned that $\mathcal{M}_1(E)$ inherits properties from E. We want to show that $\mathcal{M}_1(E)$ is Polish if E is a Polish space. For that we need the following lemma which is of considerable importance in its own right.

Lemma 2.43 *Let (E, d) be a Polish space, $(\mu_n)_{n\in\mathbb{N}}$ sequence of probability measures $\mu_n, \mu \in \mathcal{M}_1(E)$, and $\Phi \subset \mathcal{C}_\mathrm{b}(E;\mathbb{R})$ bounded subset which is equicontinuous at each $x \in E$. If $\mu_n \Rightarrow \mu$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \sup_{f\in\Phi} \Big|\langle f, \mu_n\rangle - \langle f, \mu\rangle\Big| = 0\,.$$

Definition 2.44 (Lévy's metric) Let (E, d) be a Polish space. The mapping $L\colon \mathcal{M}_1(E) \times \mathcal{M}_1(E) \to \mathbb{R}_+$ defined by

$$L(\mu,\nu) := \inf\Big\{\delta > 0\colon \mu(F) \le \nu(F^{(\delta)}) + \delta \text{ and } \nu(F) \le \mu(F^{(\delta)}) + \delta \text{ ,}\forall \text{ closed } F \subset E\Big\}\,,$$

is called *Lévy's metric*. Here, $F^{(\delta)}$ denotes the set of $x \in E$ which lie at distance less than δ from F.

Theorem 2.45 *Let (E, d) be a Polish space. Then L defined in Definition 2.44 is a complete metric, and therefore $(\mathcal{M}_1(E), L)$ is Polish and the metric is compatible with the weak convergence of probability measures.*

Proof. It is easy to realise that $L(\mu,\nu) = 0$ if and only if $\mu = \nu$, that L is symmetric and that L satisfies the triangle inequality. To show that the metric L is compatible with the weak convergence of probability measures we are left to show that

$$L(\mu_n, \mu) \to 0 \;\text{ as }\; n \to \infty \;\Leftrightarrow\; \mu_n \underset{n\to\infty}{\Rightarrow} \mu\,.$$

Suppose $L(\mu_n, \mu) \to 0$ as $n \to \infty$. Then, for every closed F,

$$\mu(F^{(\delta)}) + \delta \ge \limsup_{n\to\infty} \mu_n(F)\,,$$

for all $\delta > 0$; and therefore, $\mu(F) \geq \limsup_{n\to\infty} \mu_n(F)$. Hence, by Proposition 2.39, $\mu_n \underset{n\to\infty}{\Rightarrow} \mu$.

Now suppose $\mu_n \underset{n\to\infty}{\Rightarrow} \mu$ and let $\delta > 0$ be given as as a closed $F \subset E$. Define the function

$$\psi_F(x) := \frac{\mathrm{d}(x, (F^{(\delta)})^{\mathrm{c}})}{\mathrm{d}(x, (F^{(\delta)})^{\mathrm{c}}) + \mathrm{d}(x, F)}, \quad x \in E\,.$$

Then we can easily see that

$$\mathbb{1}_F \leq \psi_F \leq \mathbb{1}_{F^{(\delta)}} \text{ and } |\psi_F(x) - \psi_F(y)| \leq \frac{\mathrm{d}(x,y)}{\delta}\,.$$

By Lemma 2.43, we can choose $m \in \mathbb{N}$ so that

$$\sup_{n\geq m} \sup \Big\{ |\langle \psi_F, \mu_n\rangle - \langle \psi_F, \mu\rangle| \colon F \text{ closed in } E \Big\} < \delta\,,$$

which implies that for all $n \geq m$,

$$\mu(F) \leq \mu_n(f^{(\delta)}) + \delta \text{ and } \mu_n(F) \leq \mu(F^{(\delta)}) + \delta\,.$$

Hence $\sup_{n\geq m} L(\mu_n, \mu) \leq \delta$, and we have shown $L(\mu_n, \mu) \to 0$ as $n \to \infty$. Finally, we need to show that L is a complete metric, that is, we must show that if $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{M}_1(E)$ is L-Cauchy convergent, then it is a tight sequence. Pick $\varepsilon > 0$ and choose, for each $\ell \in \mathbb{N}$, an $m_\ell \in \mathbb{N}$ and a compact $K_\ell \subset E$ so that

$$\sup_{n\geq m_\ell} L(\mu_n, \mu) \leq \frac{\varepsilon}{2^{\ell+1}} \text{ and } \max_{1\leq n\leq m_\ell} \mu_n(K_\ell^{\mathrm{c}}) \leq \frac{\varepsilon}{2^{\ell+1}}\,.$$

Setting $\varepsilon_\ell = \varepsilon/2^\ell$, we get that

$$\sup_{n\in\mathbb{N}} \mu_n\big((K^{(\varepsilon_\ell)})^{\mathrm{c}}\big) \leq \varepsilon_\ell \qquad \text{for each } \ell \in \mathbb{N}\,.$$

In particular, if we define

$$K := \bigcap_{\ell+1}^{\infty} \overline{K_\ell^{(\varepsilon_\ell)}}\,,$$

then $\mu_n(K) \geq 1 - \varepsilon$ for all $n \in \mathbb{N}$.As each K_ℓ is compact, it is easy to see that K is compact. □

3 Random walks and their scaling limit

3.1 The simple random walk on $\mathbb{Z}^d$

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent identically distributed $\mathbb{Z}^d$-valued random variables on some probability space $(\Omega, \mathcal{F}, P)$ so that

$$P(X_j = e) = \begin{cases} \frac{1}{2d} & \text{if } |e| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that each $x \in \mathbb{Z}^d$ has exactly $2d$ nearest neighbours $x \pm \mathrm{e}_j, j = 1, \dots, d$. The *simple random walk (SRW) on $\mathbb{Z}^d$ with start at $x \in \mathbb{Z}^d$* is the sequence of $\mathbb{Z}^d$-valued random variables $(S_n)_{n\in\mathbb{N}_0}$ with

$$S_n = x + X_1 + \cdots + X_n \,.$$

We denote P_x the probability that the random walk starts at x (and wrote $\mathbb{E}_x$ for its expectation). The *n-step transition probability* is denoted

$$p_n(x, y) := P_x(S_n = y) \qquad x, y \in \mathbb{Z}^d, n \in \mathbb{N}_0 \,.$$

When the index x is missing in our expression we assume that $S_0 = 0$ and we write $p_n(x) = p(0, x)$.

Proposition 3.1 (Proporties) *For all $x, y \in \mathbb{Z}^d$ and $n \in \mathbb{N}_0$ the following properties hold.*

(a) $p_n(x, y) = p_n(y, x)$.

(b) $p_n(x, y) = p_n(y - x) = p_n(0, y - x)$.

(c) $p_n(y) = p_n(-y)$.

(d) $p_0(x, y) = \delta(x, y)$.

(e) $p_n(x, y) = P_x(S_n = y) = P(S_{n+m} = y | S_m = x)$ *for all* $m \in \mathbb{N}_0$.

Proof. The easy proofs are left as an exercise.

□

For $m \in \mathbb{N}$, define

$$\widetilde{S}_n := S_{n+m} - S_m = X_{m+1} + \cdots + X_{m+n} \,.$$

Then $(\widetilde{S}_n)_{n\in\mathbb{N}_0}$ is a simple random walk starting at 0 and independent of $\{X_1, \dots, X_m\}$. Furthermore, it is easy to see that

$$P(S_{n+1} = x_{n+1} | S_0 = x_0, \dots, S_n = x_n) = p_1(x_n, x_{n+1}) =: p(x_n, x_{n+1}) \,. \tag{3.1}$$

Equation (3.1) says that the probability to reach a site x_{n+1} in the next 'time step' only depends on the present state $S_n = x_n$ and not on the past of the random walk up to time n. This is called the *Markov property*.

Proposition 3.2 *For all $m, n \in \mathbb{N}_0$, and $x, y \in \mathbb{Z}^d$,*

$$p_{m+n}(x, y) = \sum_{z\in\mathbb{Z}^d} p_m(x, z) p_n(z, y) \,.$$

Proof. Using the Markov property we get

$$\begin{aligned} p_{m+n}(x, y) &= \sum_{y_i\in\mathbb{Z}^d, i=0,\dots,m+n; y_0=x, y_{m+n}=y} p(x, y_1)p(y_1, y_2)\cdots p(y_{m+n-1}, y_{m+n}) \\ &= \sum_{y_m\in\mathbb{Z}^d, y_{m+n}=y} p_m(x, y_m) p_n(y_m, y_{m+n}) \,. \end{aligned}$$

Note that the sum is non-zero only if the y_i's form a nearest neighbour path in that $|y_{i+1} - y_i| = 1, i = 0, \dots, m+n-1$. $\square$

Definition 3.3 (Markov Chain/Process) Let $E \neq \varnothing$ be at most countable. $\Pi = (\Pi(x,y))_{x,y\in E}$ is called a *stochastic matrix* if $\Pi(x,y) \in [0,1], x, y \in E$ and if for all $x \in E$, $\sum_{y\in E} \Pi(x,y) = 1$.

Let Π be a stochastic matrix. A sequence $(X_i)_{i\in\mathbb{N}_0}$ of e-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$ is called a *Markov chain or Markov process with state space E and transitions matrix* Π, if

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \dots, X_n = x_n) = \Pi(x_n, x_{n+1})$$

for every $n \in \mathbb{N}$ and every $x_0, \dots, x_{n+1} \in E$ with $P(X_0 = x_0, \dots, x_n = x_n) > 0$. The distribution $\alpha = P \circ X_0^{-1}$ of X_0 is called the *initial distribution of the Markov chain.*

The transition matrix of the simple random walk (SRW) on $\mathbb{Z}^d$ is

$$P(x,y) = \begin{cases} \frac{1}{2d} & \text{if } |y-x| = 1, \\ 0 & \text{otherwise}. \end{cases} \quad \text{for } x, y \in \mathbb{Z}^d. \tag{3.2}$$

3.2 The Local central Limit Theorem (LCLT)

We briefly discuss the asymptotic for $p_n(x)$ for large n. Let $S_0 \equiv 0$ in the following. Note that the position S_n of the SRW at time n is the sum of independent, identically distributed random variables each with mean 0 and variance $\frac{1}{d}\mathbb{1}$, where $\mathbb{1}$ is the identity matrix in d dimensions. The CLT, Theorem 2.20, says that $S_n/\sqrt{n}$ converges in distribution to a normal random variable in $\mathbb{R}^d$ with mean 0 and variance $\frac{1}{d}\mathbb{1}$, i.e., if $A \subset \mathbb{R}^d$ is an open ball,

$$\lim_{n\to\infty} P\Big(\frac{S_n}{\sqrt{n}} \in A\Big) = \int_A \Big(\frac{d}{2\pi}\Big) \mathrm{e}^{-\frac{d|x|^2}{2}} \, \mathrm{d}x_1 \cdots \mathrm{d}x_d .$$

Note that

- S_n takes value sin $\mathbb{Z}^d$.
- n even, then S_n has even parity (sum of its components is even), S_n has odd parity when n odd.
- A ball $A \subset \mathbb{R}^d$ contains about $n^{d/2}|A|$ points in the lattice $n^{-1/2}\mathbb{Z}^d$, where $|A|$ is the Lebesgue volume.
- About half of these points will have even parity.

Suppose now that n is even, then we expect that

$$P\Big(\frac{S_n}{\sqrt{n}} = \frac{x}{\sqrt{n}}\Big) \approx \frac{2}{n^{d/2}} \Big(\frac{d}{2\pi}\Big)^{d/2} \mathrm{e}^{-d|x|^2/2n} . \tag{3.3}$$

The Local Central Limit Theorem (LCLT) below makes this approximation statement precise. Before we can state that theorem, we shall make a few notations and collect a couple of facts. Define

$$\overline{p}_0(x) = \delta(x,0) \qquad \text{and } \overline{p}_n(x) = \overline{p}(n,x) = 2\Big(\frac{d}{2\pi n}\Big)^{d/2} \mathrm{e}^{-\frac{d|x|^2}{2n}} .$$

We write $n \leftrightarrow x$ if n and x have the same parity (i.e., if $n + x_1 + \cdots + x_d$ is even). Similarly we write $x \leftrightarrow y$ and $n \leftrightarrow m$. The error is defined as

$$E(n,x) := \begin{cases} p(n,x) - \overline{p}(n,x) & \text{if } n \leftrightarrow x, \\ 0 & \text{if } n \not\leftrightarrow x. \end{cases} \tag{3.4}$$

In the analysis for the LCLT we need to compare functions defined on the lattice $\mathbb{Z}^d$ with functions defined on $\mathbb{R}^d$ along with their derivatives. If $f\colon \mathbb{Z}^d \to \mathbb{R}$ and $y \in \mathbb{Z}^d$, define the discrete derivatives in direction y,

$$\begin{aligned} \nabla_y f(x) &:= f(x_y) - f(x)\,, \\ \nabla_y^2 f(x) &:= f(x+y) + f(x-y) - 2f(x)\,. \end{aligned}$$

If $f\colon \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^3$, $x, y \in \mathbb{Z}^d, y = |y|u$, then Taylor's theorem with remainder gives

$$\begin{aligned} |\nabla_y f(x) - |y| D_u f(x)| &\le \frac{1}{2}|y|^2 \sup_{0\le a\le 1} \{|D_{uu} f(x+ay)|\} \\ |\nabla_y^2 f(x) - |y|^2 D_{uu} f(x)| &\le \frac{1}{3}|y|^3 \sup_{0\le a\le 1} \{|D_{uuu} f(x+ay)|\}\,. \end{aligned}$$

Theorem 3.4 (Local Central Limit Theorem (LCLT)) *For the SRW on $\mathbb{Z}^d$ the following estimates hold for the error defined in* (3.4).

$$\begin{aligned} |E(n,x)| &\le O(n^{-(d+2)/2})\,, \\ |E(n,x)| &\le |x|^{-2} O(n^{-d/2})\,. \end{aligned}$$

If $y \leftrightarrow 0$ there exists a $c_y \in (0,\infty)$ such that

$$\begin{aligned} |\nabla_y E(n,x)| &\le c_y O(n^{-(d+3)/2}) \\ |\nabla_y E(n,x)| &\le c_y |x|^{-2} O(n^{-(d+1)/2}) \\ |\nabla_y^2 E(n,x)| &\le c_y O(n^{-(d+4)/2}) \\ |\nabla_y^2 E(n,x)| &\le c_y |x|^{-2} O(n^{-(d+2)/2})\,. \end{aligned}$$

Proof. A detailed proof can be found in [LL10] or [Law96]. We give some hints and leave details for the interested reader. The key point is to use characteristic functions (Fourier transformation) and the independence of the X_i's in the definition of the SRW. The latter fact leads to just considering the power of the single characteristic function. The whole proof is rather technical as it requires expansion of characteristic function and careful estimates of various integrals. All techniques and methods are standard in analysis, and the patient reader can work out all separate steps.

Characteristic function for lattice functions: If Y is any random variable taking value sin $\mathbb{Z}^d$, the characteristic function $\varphi_Y \equiv \varphi$, given by

$$\varphi(k) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle Y,k\rangle}] = \sum_{x\in\mathbb{Z}^d} P(Y=x)\,\mathrm{e}^{\mathrm{i}\langle x,k\rangle}\,, \tag{3.5}$$

has period 2π in each component. We can therefore think of φ as a function on $[-\pi,\pi]^d$ with periodic boundary conditions. The cube $\mathsf{BZ} := [-\pi,\pi]^d$ is called the *Brillouin zone*. The inversion formula for the characteristic function is

$$P(Y=y) = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \mathrm{e}^{-\mathrm{i}\langle y,k\rangle}\varphi(k)\,\mathrm{d}k\,.$$

It is easy to see that

$$\varphi_{X_1}(k) \equiv \varphi(k) = \frac{1}{d}\sum_{j=1}^d \cos(k_j)\,,$$

and hence the characteristic function for the SRW at time n is

$$\varphi_{S_n} \equiv \varphi^n(k) = \Big(\frac{1}{d}\sum_{j=1}^d \cos(k_j)\Big)\,.$$

We may assume that $n \leftrightarrow x$, then

$$p_n(x) = \frac{1}{(2\pi)^d}\int_{\mathsf{BZ}} \mathrm{e}^{-\mathrm{i}\langle x,k\rangle}\,\mathrm{d}k\,.$$

Since $n \leftrightarrow x$, we can replace k by $k + (\pi,\dots,\pi)$, and have

$$p_n(x) = 2\frac{1}{(2\pi)^d}\int_A \mathrm{e}^{-\mathrm{i}\langle x,k\rangle}\,\mathrm{d}k\,.$$

with $A = [-\pi/2,\pi/2]\times[-\pi,\pi]^{d-1}$. Expansion of φ around the origin gives $\varphi(k) = 1 - \frac{1}{2}|k|^2 + O(|k|^4)$. We can find $r \in (0,\pi/2)$ such that

$$\varphi(k) \le 1 - \frac{1}{4d}|k|^2 \quad \text{for } |k| \le r\,,$$

and there is thus a $\varrho < 1$ depending on the r such that $|\varphi(k)| \le \varrho$ for $|k| \ge r, k \in A$. Hence $p(n,x) = I(n,x) + J(n,x)$ with

$$I(n,x) = 2(2\pi)^{-d}\int_{|k|\le r} \mathrm{e}^{-\mathrm{i}\langle x,k\rangle}\,\mathrm{d}k\,,$$

and $|J(n,x)| \le \varrho^n$. The rest of the proof now concerns the integral $I(n,x)$ which needs to be further split in different parts, see details in [LL10]. □

We will demonstrate the usefulness of the LCLT in Section 3.3. Before we suggest the following insightful exercises.

Exercise 3.5 Show the following. Suppose $x \leftrightarrow y$. Then

$$\lim_{n\to\infty} \sum_{z\in\mathbb{Z}^d} |p_n(x,z) - p_n(y,z)| = 0\,.$$

Hint: *Use the central limit theorem for all* $|z| \geq n^\gamma$ *with* $\gamma > \frac{1}{2}$. *Then use the LCLT for* $|n| \leq n^\gamma$ *for* $\gamma < \frac{1}{2} + \frac{1}{d}$. ☕☕☕

Exercise 3.6 Prove for every $m \leftrightarrow 0$,

$$\lim_{n\to\infty} \sum_{z\in\mathbb{Z}^d} |p_n(z) - p_{n+m}(z)| = 0\,.$$

☕☕

3.3 Strong Markov property and basic potential theory

A *random time* is a random variable $\tau\colon \Omega \to \mathbb{N}_0 \cup \{\infty\}$. A *stopping time for the SRW is any random time* τ *which depends only on the past and present, i.e., for time* n *on* $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$. The future is the σalgebra $\mathcal{H}_n = \sigma(X_{n+1},\dots)$ and is independent of $\mathcal{F}_n$, i.e., $\mathcal{F}_n \perp \mathcal{H}_n$. A sequence $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$ of nested σ-sub-algebras $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots$ is a *filtration* for the SRW if $\mathcal{F}_n \subset \mathcal{G}_n$ and $\mathcal{G}_n \perp \mathcal{H}_n$. A random time τ is a stopping time with respect to the filtration $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$ if for each $n \in \mathbb{N}_0$, $\{\tau = n\} \in \mathcal{G}_n$.

Exercise 3.7 (a) Let $A \subset \mathbb{Z}^d$ and $k \in \mathbb{N}$. Show that $\tau = \inf\{n \geq k\colon S_n \in A\}$ is a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$.

(b) If τ_1 and τ_2 are stopping times then so are $\tau_1 \wedge \tau_2$ and $\tau_2 \vee \tau_2$.

(c) Let $(Y_i)_{i\in\mathbb{N}_0}$ be sequence of independent rand identically distributed random variables with $(Y_i)_{i\in\mathbb{N}_0} \perp (X_i)_{i\in\mathbb{N}_0}$ and $P(Y_i = 1) = 1 - P(Y_i = 0) = \lambda$. Define $T = \inf\{j \in \mathbb{N}_0\colon Y_j = 1\}$ and show that T is stopping with respect to the filtration $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$ where $\mathcal{G}_n(\sigma(X_1,\dots,X_n,Y_0,Y_1,\dots,Y_n)$. When $Y_i = 1$ occurs, the random walk will be stopped (killed). The so-called 'killing time' of the random walk has geometric distribution, i.e., $P(T = j) = (1-\lambda)^j\lambda$.

☕☕

When τ is a stopping time with respect to the filtration $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$, then we write

$$\mathcal{G}_\tau := \{A \in \mathcal{F}\colon \text{ for each } n \in \mathbb{N}_0\colon A \subset \{\tau \leq n\} \in \mathcal{G}_n\}\,.$$

Theorem 3.8 (Strong Markov property) *Let* τ *be a stopping time with respect to the filtration* $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$. *Them on* $\{\tau < \infty\}$ *the process* $(\widetilde{S}_n)_{n\in\mathbb{N}_0}$, *defined by*

$$\widetilde{S}_n := S_{n+\tau} - S_\tau$$

is a SRW starting at the origin and independent of $\mathcal{G}_\tau$.

Proof. Let $x_0, \ldots, x_n \in \mathbb{Z}^d$; $A \in \mathcal{G}_\tau$.

$$
\begin{aligned}
&P(\widetilde{S}_0 = x_0, \ldots, \widetilde{S}_n = x_n \cap A \cap \{\tau < \infty\}) \\
&= \sum_{j=0}^{\infty} P(\widetilde{S}_0 = x_0, \ldots, \widetilde{S}_n = x_n \cap A \cap \{\tau = j\}) \\
&= \sum_{j=0}^{\infty} P(S_j - S_j = x_0, \ldots, S_{j+n} - S_j = x_n) \cap A \cap \{\tau = j\}) \\
&= \sum_{j=0}^{\infty} P(S_0 = x_0, \ldots, S_n = x_n) P(A \cap \{\tau = j\}) \\
&= P(S_0 = x_0, \ldots, S_n = x_n) P(A \cap \{\tau < \tau\}) .
\end{aligned}
$$

□

We now come to the promised application of the LCLT and the strong Markov property. Let R_n denote the number of visits of the SRW to the origin 0 up through time n,

$$
R_n := \sum_{j=0}^{\infty} 1\!\!1\{S_j = 0\} , \quad R := R_\infty .
$$

By the LCLT, Theorem 3.4,

$$
\begin{aligned}
\mathbb{E}[R_n] &= \sum_{j=0}^{\infty} p_j(0) + \sum_{j \leq ,j \text{ even}} \Big(2\Big(\frac{d}{2\pi j}\Big)^{d/2} + O\Big(j^{-(d+2)/2}\Big)\Big) \\
&\sim \begin{cases} \sqrt{2/\pi} n^{1/2} + O(1) & \text{if } d = 1 , \\ \frac{1}{\pi} \log n + O(1) & \text{if } d = 2 , \\ c + O(n^{(2-d)/2}) & \text{of } d \geq 3 . \end{cases}
\end{aligned}
$$

From this we get that $\mathbb{E}[R] = \infty$ for $d = 1, 2$. Define $\tau := \inf\{j \in \mathbb{N} \colon S_j = 0\}$ and note that then $R = 1 + \sum_{j=\tau}^{\infty} 1\!\!1\{S_j = 0\}$. Thus, by the strong Markov property, Theorem 3.8,

$$
\mathbb{E}[R] = 1 + P(\tau < \infty) , \quad \text{ or } \; P(\tau = \infty) = \frac{1}{\mathbb{E}[R]} \begin{cases} = 0 & \text{if } d = 1, 2 , \\ > 0 & \text{if } d \geq 3 . \end{cases}
$$

Another application of Theorem 3.8 shows that if $d \geq 3$,

$$
P(R = j) = p(1 - p)^{j-1} , \quad \text{ with } p = P(\tau = \infty) .
$$

We summarise our findings in the following theorem where we define *transience* and *recurrence*.

Theorem 3.9 *If $d = 1, 2$, the simple random walk is* recurrent *, i.e.,*

$$
P(S_n = 0 \textit{ infinitely often }) = 1 .
$$

If $d \geq 3$, the simple random walk is transient *, i.e.,*

$$
P(S_n = 0 \textit{ infinitely often }) = 0 .
$$

Let $f\colon \mathbb{Z}^d \to \mathbb{R}$, then the *discrete Laplacian of f at $x \in \mathbb{Z}^d$* is defined as

$$\Delta f(x) = \Big(\frac{1}{2d}\sum_{|e|=1} f(x+e)\Big) - f(x) = \frac{1}{2d}\sum_{|e|=1} \nabla_e f(x)$$
$$= \frac{1}{2d}\sum_{j=1}^{d} \nabla^2_{\mathrm{e}_j} f(x) = \frac{1}{2d}\sum_{j=1}^{d} \big(-\nabla^*_{\mathrm{e}_j}\nabla_{\mathrm{e}_j} f(x)\big)\,,$$

where we define the adjoint discrete derivative in direction e_j as

$$\nabla^*_{\mathrm{e}_j} f(x) := f(x-\mathrm{e}_j) - f(x)\,.$$

The discrete Laplacian of a given function is also given in terms of the SRW on $\mathbb{Z}^d$, namely note that

$$\Delta f(x) = \mathbb{E}_x[f(S_1) - f(S_0)]\,.$$

Definition 3.10 (Martingale) A sequence $(M_n)_{n\in\mathbb{N}_0}$ of random variables on some probability space $(\Omega, \mathcal{F}, P)$ with $\mathbb{E}[|M_i|] < \infty$ for all $\in\in \mathbb{N}_0$ *is a martingale with respect to the filtration* $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ if each M_n is $\mathcal{F}_n$ measurable and

$$\mathbb{E}[M_m|\mathcal{F}_n] = M_n \;\text{ almost surely for all } n \le m\,, \tag{3.6}$$

i.e.,

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n\,. \tag{3.7}$$

Taking expectation value in (3.7) yields $\mathbb{E}[M_{n+1}] = \mathbb{E}[M_n]$ for all $n \in \mathbb{N}_0$ and thus

$$\mathbb{E}[M_n] = \mathbb{E}[M_0] \qquad \text{for all } n \in \mathbb{N}_0\,. \tag{3.8}$$

In order to verify (3.6) it suffices to prove (3.7), since if this holds, we obtain

$$\mathbb{E}[M_{n+2}|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[M_{n+2}|\mathcal{F}_{n+1}|\mathcal{F}_n]] = \mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n\,,$$

and so on.

Example 3.11 (a) Fair coin, $P(\xi_i = 1) = P(\xi_i = -1) = \frac{1}{2}$ and $M_n := \xi_1 + \cdots + \xi_n$. Let $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$, $X_0 = 0$, $\mathcal{F}_0 = \{\varnothing, \Omega\}$. Then $(X_n)_{n\in\mathbb{N}_0}$ is a martingale

(b) Suppose $X \in L^1$ and let $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ be a filtration. Then by the chain rule of conditional expectations we obtain that $M_n := \mathbb{E}[M_m|\mathcal{F}_n]$ is a martingale.

(c) Suppose $(X_i)_{i\in\mathbb{N}_0}$ is a sequence of independent, identically distributed random variables with $\mathbb{E}[X_i] = \mu$ for all $i \in \mathbb{N}_)$ and define $S_n := S_0 + X_1 + \cdots + X_n$. Then $M_n := S_n - n\mu$ is martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$. To see this note that $M_{n+1} - M_n = X_{n+1} - \mu \perp X_0, \ldots, X_n$ and thus

$$\mathbb{E}[M_{n+1} - M_n|\mathcal{F}_n] = \mathbb{E}[X_{n+1} - \mu] = 0\,.$$

♣

The optional sampling (or optional stopping) theorem states that (under certain conditions) if $(M_n)_{n\in\mathbb{N}_0}$ is a martingale and τ be a stopping time then

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0]\,. \tag{3.9}$$

so (3.9) is just the generalisation of (3.8) to random stopping times.

Proposition 3.12 *Suppose that $(M_n)_{n\in\mathbb{N}_0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and suppose that τ is a stopping time and that τ is bounded, $\tau \le K < \infty$. Then*

$$\mathbb{E}[M_\tau|\mathcal{F}_0] = M_0\,.$$

in particular, $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.

Proof. First note that the event $\{\tau > n\}$ is measurable with respect to $\mathcal{F}_n$. Furthermore,

$$M_\tau = \sum_{j=0}^{K} M_j 1\!\!1\{\tau = j\}\,.$$

We now take the conditional expectation with respect to $\mathcal{F}_{K-1}$,

$$\mathbb{E}[M_\tau|\mathcal{F}_{K-1}] = \mathbb{E}[M_k 1\!\!1\{\tau = K\}|\mathcal{F}_{K-1}] + \sum_{j=0}^{K-1}\mathbb{E}[M_j 1\!\!1\{\tau = j\}|\mathcal{F}_{K-1}]\,.$$

For $j \le K-1$, $M_j 1\!\!1\{\tau = j\}$ is $\mathcal{F}_{K-1}$-measurable; hence

$$\mathbb{E}[M_j 1\!\!1\{\tau = j\}|\mathcal{F}_{K-1}] = M_j\{\tau = j\}\,.$$

The event $\{\tau = K\}$ is the same as the event $\{\tau > K-1\}$. Hence,

$$\mathbb{E}[M_K 1\!\!1\{\tau = K\}|\mathcal{F}_{K-1}] = 1\!\!1\{\tau > K-1\}\mathbb{E}[M_K|\mathcal{F}_{K-1}] = 1\!\!1\{\tau > K-1\}M_{K-1}\,.$$

Therefore,

$$\begin{aligned}\mathbb{E}[M_\tau|\mathcal{F}_{K-1}] &= 1\!\!1\{\tau > K-1\}M_{K-1} + \sum_{j=0}^{K-1} M_j 1\!\!1\{\tau = j\}\\ &= 1\!\!1\{\tau > K-2\}M_{K-2} + \sum_{j=0}^{K-2} M_j 1\!\!1\{\tau = j\}\,.\end{aligned}$$

Repeating our argument again, this time conditioning with respect to $\mathcal{F}_{K-2}$, we get

$$\mathbb{E}[M_\tau|\mathcal{F}_{K-2}] = 1\!\!1\{\tau > K-3\}M_{K-2} + \sum_{j=0}^{K-3} M_j 1\!\!1\{\tau = j\}\,.$$

We continue this process until we get $\mathbb{E}[M_\tau|\mathcal{F}_0] = M_0$. □

Many examples of interest have stopping times which are not necessarily bounded. Suppose τ is a stopping time with $P(\tau < \infty) = 1$. Can we then conclude that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$? To answer this question we consider the following stopping times

$$\tau_n = \min\{\tau, n\} = n \wedge \tau\,.$$

Then

$$M_\tau = M_{\tau_n} + M_\tau 1\!\!1\{\tau > n\} - M_n 1\!\!1\{\tau > n\}\,. \tag{3.10}$$

Since T_n is a bounded stopping time, it follows that $\mathbb{E}[M_{\tau_n}] = \mathbb{E}[M_0]$. we would like to show that the other two terms on the right hand side of (3.10) do not contribute as $n \to \infty$. The second term is actually not much of a problem as the probability of the event $\{\tau > n\}$goes to 0 as $n \to \infty$, one can show (exercise) that if $\mathbb{E}[|M_\tau|] < \infty$ then $\mathbb{E}|M_\tau| 1\!\!1\{\tau > n\}] \to$ as $n \to \infty$. The third term is more troublesome. Here, we need to add the additional assumption that $\mathbb{E}[|M_n| 1\!\!1\{\tau > n\}] \to 0$ as $n \to \infty$.

Theorem 3.13 (Optional Sampling Theorem) *Suppose $(M_n)_{n\in\mathbb{N}_0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and τ is a stopping time satisfying $P(\tau < \infty) = 1$, $\mathbb{E}[|M_\tau|] < \infty$, and*

$$\lim_{n\to\infty} \mathbb{E}[|M_n| 1\!\!1\{\tau > n\}] = 0\,. \tag{3.11}$$

Then, $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.

Condition (3.11) is often hard to verify. Suppose X is a random variable $\mathbb{E}[X] < \infty$ and assume in addition that X has a Radon-Nokodym density with respect to the Lebesgue measure, i.e., $F_X(t) = \int_{-\infty}^t f(x)\,\mathrm{d}x$. Then it follows that

$$\lim_{M\to\infty} \mathbb{E}[|X| 1\!\!1\{|X| > M\}] = \lim_{M\to\infty} \int_M^\infty x f(x)\,\mathrm{d}x = 0\,.$$

Now suppose that we have a sequence $(X_n)_{n\in\mathbb{N}_0}$ of random variables. The sequence $(X_n)_{n\in\mathbb{N}_0}$ is *uniformly integrable* if for every $\varepsilon > 0$ there exists a $M > 0$ such that for each n,

$$\mathbb{E}[|X_n| 1\!\!1\{|X_n| > M\}] < \varepsilon\,. \tag{3.12}$$

Lemma 3.14 *Let $(X_n)_{n\in\mathbb{N}_0}$ be uniformly integrable. Then, for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $P(A) < \delta$, then*

$$\mathbb{E}[|X_n| 1\!\!1_A] < \varepsilon\,, \quad \textit{for each } n\,. \tag{3.13}$$

Proof. Pick $\varepsilon > 0$ and choose $M > 0$ sufficiently large so that $\mathbb{E}[|X_n| 1\!\!1\{|X_n| > M\}] < \varepsilon/2$ for all n. For $\delta = \varepsilon/(2M)$ we get for every A with $P(A) < \delta$,

$$\mathbb{E}[|X_n| 1\!\!1_A] \le \mathbb{E}[|X_n| 1\!\!1_A; |X_n| \le M] + \mathbb{E}[|X_n|; |X_n| > M] < MP(A) + \varepsilon/2 < \varepsilon\,.$$

□

Now suppose that $(M_n)_{n\in\mathbb{N}_0}$ is a uniformly integrable martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and T is a stopping time with $P(T < \infty) = 1$. Then

$$\lim_{n\to\infty} P(T > n) = 0\,,$$

and hence by uniformly integrability,

$$\lim_{n\to\infty} \mathbb{E}[|M_n|1\!\!1\{T > n\}] = 0\,,$$

that is, condition (3.11) holds. With this we have proven the following version of the optional sampling theorem.

Theorem 3.15 (Optional Sampling Theorem - 2nd version) *Suppose that $(M_n)_{n\in\mathbb{N}_0}$ is a uniformly integrable martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and τ is a stopping time satisfying $P(\tau < \infty) = 1$ and $\mathbb{E}[|M_\tau|] < \infty$. Then, $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.*

Example 3.16 Consider the SRW on $\{0, 1, \ldots, N\}$, $N \in \mathbb{N}$, with absorbing boundaries (SRW stops when it reaches either 0 or N). Set $S_0 = x \in \{0, 1, \ldots, N\}$ and $T = \inf\{j \in \mathbb{N}_0 \colon S_j = 0, N\}$. Then

$$P_x(T < \infty) = P(\bigcup_{n\in\mathbb{N}} \{T < n\}) = 1 - \limsup_{n\to\infty} P_x(T \geq n) = 1\,.$$

according to Lemma 3.17 below. Furthermore, $M_n := S_{n\wedge T}$ is a bounded martingale and Theorem 3.13 states that for $x \in \{0, 1, \ldots, N\}$,

$$x = \mathbb{E}_x[M_0] = \mathbb{E}_x[M_T] = NP_x(S_T = N)\,.$$

Therefore,

$$P_x(S_T = N) = \frac{x}{N}\,.$$

Define the function $F\colon \{0, 1, \ldots, N\}[0, 1], F(x) = P(S_T = N|S_0 = x)$. This function clearly satisfies the following iteration and boundary conditions,

$$\begin{aligned} &F(x) = \frac{1}{2}F(x+1) + \frac{1}{2}F(x-1)\,, \quad x \in \{1, \ldots, N-1\}\,, \\ &F(0) = 0, F(N) = 1\,. \end{aligned} \tag{3.14}$$

Then the only function F solving (3.14) with the boundary condition $F(0) = a, F(N) = b, a, b \in \mathbb{R}$ instead of $F(0) = 0, F(N) = 1$ is the linear function

$$F(x) = a + \frac{x(b-a)}{N}\,.$$

♣

Lemma 3.17 *Let $\Lambda \subset \mathbb{Z}^d$ be a finite set and $\tau := \inf\{j \in \mathbb{N} \colon S_j \notin \Lambda\}$. Then there exist $C \in (0, \infty$ and $\varrho \in (0, 1)$ (depending on Λ) such that for each $x \in \Lambda$,*

$$P_x(\tau \geq n) \leq C\varrho^n\,, \quad n \in \mathbb{N}\,.$$

Proof. Let $R = \sup\{|x| \colon x \in \Lambda\}$. Then for each $x \in \Lambda$, there is a path of length $R + 1$ starting at x and ending outside of Λ, hence

$$P_x(\tau \leq R + 1) \geq \Big(\frac{1}{2d}\Big)^{R+1}\,.$$

By the Markov property,

$$\begin{aligned} P_x(\tau > k(R+1)) &= P_x(\tau > (k-1)(R+1))P_x(\tau > k(R+1)|\tau > (k-1)(R+1)) \\ &\le P_x(\tau > (k-1)(R+1))(1-(2d)^{-(R+1)}) \,, \end{aligned}$$

and hence

$$P_x(\tau > k(R+1)) \le \varrho^{k(R+1)} \,,$$

where $\varrho = (1-(2d)^{-(R+1)})^{1/(R+1)}$. Write $n = k(R+1)+j$ with $j \in \{1,\dots,R+1\}$. Then

$$P_x(\tau \ge n) \le P_x(\tau > k(R+1)) \le \varrho^{k(R+1)} \le \varrho^{-(R+1)}\varrho^n \,,$$

and conclude with $C = C(R) = \varrho^{-(R+1)}$. □

Proposition 3.18 *Suppose f is a bounded function, harmonic on $\Lambda \subset \mathbb{Z}^d$, and $\tau = \inf\{j \in \mathbb{N}_0 \colon S_j \notin \Lambda\}$. Then $M_n := f(S_{n\wedge\tau})$ is a martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$.*

Proof. Assume that $S_0 = x$. By the Markov property,

$$\mathbb{E}_x[f(S_{n+1})|\mathcal{F}_n] = \mathbb{E}_{S_n}[f(S_1)] = f(S_n) + \Delta f(S_n) \,.$$

Let $B_n := \{\tau > n\}$, then $M_{n+1} = M_n$ on B_n^{c}, and

$$\begin{aligned} \mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[M_{n+1}1\!\!1_{B_n}|\mathcal{F}_n] + \mathbb{E}[M_{n+1}1\!\!1_{B_n^{\mathrm{c}}}|\mathcal{F}_n] \\ &= \mathbb{E}[f(S_{n+1})1\!\!1_{B_n}|\mathcal{F}_n] + \mathbb{E}[M_n 1\!\!1_{B_n^{\mathrm{c}}}|\mathcal{F}_n] \\ &= 1\!\!1_{B_n}\mathbb{E}[f(S_{n+1})|\mathcal{F}_n] + M_n 1_{B_n^{\mathrm{c}}} \\ &= 1\!\!1_{B_n}(f(S_n) + \Delta f(S_n)) + M_n 1\!\!1_{B_n^{\mathrm{c}}} \\ &= 1\!\!1_{B_n} f(S_n) + M_n 1\!\!1_{B_n^{\mathrm{c}}} = M_n \,, \end{aligned}$$

where we used that $\Delta f(S_n) = 0$ on B_n in the last equation. □

For $\Lambda \subset \mathbb{Z}^d$ denotes its boundary $\partial\Lambda := \{x \in \Lambda^{\mathrm{c}} \colon |x-y| = 1 \text{ for some } y \in \Lambda\}$, and its closure by $\overline{\Lambda} = \Lambda \cup \partial\Lambda$.

Theorem 3.19 (Discrete Dirichlet boundary value problem) *Let $\Lambda \subset \mathbb{Z}^d$ be finite and $F \colon \partial\Lambda \to \mathbb{R}$ be given. The unique function $f \colon \overline{\Lambda} \to \mathbb{R}$ satisfying*

(a) $\Delta f(x) = 0$ for all $x \in \Lambda$,

(b) $f(x) = F(x)$ for $x \in \partial\Lambda$,

is given as

$$f(x) := \mathbb{E}_x[F(S_\tau)] \,, \quad \tau := \inf\{j \in \mathbb{N}_0 \colon S_j \notin \Lambda\} \,. \tag{3.15}$$

Proof. It is straightforward to see that f defined by (3.15) satisfies (a) and (b). To show uniqueness assume that f satisfies (a) and (b) and let $x \in \Lambda$. Then $M_n := f(S_{n\wedge\tau})$ is a bounded martingale and by the optional stopping theorem, Theorem 3.13,

$$f(x) = \mathbb{E}_x[M_0] = \mathbb{E}_x[M_\tau] = \mathbb{E}_x[F(S_\tau)] \,.$$

□

The *Green function of the SRW* is defined is the number of visits up to time n to $y \in \mathbb{Z}^d$ when starting in $x \in \mathbb{Z}^d$,

$$G_n(x,y) := \mathbb{E}_x\Big[\sum_{j=0}^n 1\!\!1\{S_j = y\}\Big] = \sum_{j=0}^n p_j(x,y) = \sum_{j=0}^n p_j(y-x)\,, \tag{3.16}$$

or, for all dimensions $d \geq 3$, one can obtain the limit $n \to \infty$ (total number of visits to y),

$$G(x,y) = G_\infty(x,y) = \sum_{j=0}^\infty p_j(x,y)\,. \tag{3.17}$$

Let $G(x) := G(0,x)$ and note

$$\Delta G(x) = \mathbb{E}_x\Big[\sum_{j=1}^\infty 1\!\!1\{S_j = y\}\Big] - \mathbb{E}_x\Big[\sum_{j=0}^\infty 1\!\!1\{S_j = y\}\Big] = \mathbb{E}[-1\!\!1\{S_0 = x\}] = -\delta(x,0)\,. \tag{3.18}$$

With the help of the LCLT, Theorem 3.4, some computation and estimates, one may derive that for $d \geq 3$, as $|x| \to \infty$,

$$G(x) \sim a_d\,|x|^{2-d}\,, \quad a_d = \frac{d}{2}\Gamma(d/2-1)\pi^{-d/2}\,. \tag{3.19}$$

3.4 Discrete Heat Equation and its scaling limit

Suppose $t_i \in \mathbb{N}_0$, $i \in \mathbb{N}_0$, are ordered times of the SRW, i.e., $t_i \leq t_{i+1}$. For any $x_0, x_1 \in \mathbb{Z}^d$ denote $P(x_1,t_1;x_0,t_0)$ the (transition) probability that the SRW is at x_1 at time t_1 when he was at x_0 at time $t_0 \leq t_1$. In the following we write $x \sim y$ when $|x-y| = 1$, i.e., when $x.y \in \mathbb{Z}^d$ are nearest-neighbours. The transition probabilities depend only on the differences $t_1 - t_0$ and $X_1 - x_0$ and satisfy the following properties:

$$P(x_1,t_0;x_0,t_0) = \delta(x_1,x_0)\,. \tag{3.20}$$

$$\sum_{x_1\in\mathbb{Z}^d} P(x_1,t_1;x_0,t_0) = 1\,, \quad \text{for any } t_1 \geq t_0\,. \tag{3.21}$$

$$P(x,t+1;x_0,t_0) = \frac{1}{2d}\sum_{x'\in\mathbb{Z}^d\colon x'\sim x} P(x',t;x_0,t_0)\,, \quad \text{for } t \geq t_0, t \in \mathbb{N}_0 \tag{3.22}$$

or rewrite using the definition of the discrete Laplacian to obtain the following discrete heat equation (DHE) *discrete heat equation*,

$$P(x,t+1;x_0,t_0) - P(x,t;x_0,t_0) = \Delta P(x,t;x_0,t_0)\,. \tag{3.23}$$

To solve the discrete heat equation (3.23) we shall use Fourier transform, that is, characteristic functions. Let us briefly recall that for any function $f\colon \mathbb{Z}^d \to \mathbb{R}$, the Fourier transfrom reads as

$$\widehat{f}(k) = \sum_{y\in\mathbb{Z}^d} f(y)\,\mathrm{e}^{-\mathrm{i}\langle k,y\rangle}\,, \quad k \in \mathsf{BZ} := [-\pi,\pi]^d\,. \tag{3.24}$$

Here BZ denotes the Brillouin Zone, and the function defined in (3.24) is 2π - periodic. Furthermore, the Fourier transformation is defined only for functions which are $\ell_1(\mathbb{Z}^d)$ respectively $\ell_2(\mathbb{Z}^d)$, for details see [Spi01]. Using

$$\frac{1}{(2\pi)^d}\int_{\mathsf{BZ}} \mathrm{e}^{\mathrm{i}\langle k,x-y\rangle}\,\mathrm{d}k = \delta(x,y)\,, \tag{3.25}$$

it is easy to see that

$$f(x) = \frac{1}{(2\pi)^d}\int_{\mathsf{BZ}} \widehat{f}(k)\,\mathrm{e}^{\mathrm{i}\langle k,x\rangle}\,\mathrm{d}k\,. \tag{3.26}$$

Returning to our problem solving the discrete heat equation we see that

$$P(x,t;x_0,t_0) = \frac{1}{(2\pi)^d}\int_{\mathsf{BZ}} \widehat{P}(k,t)\,\mathrm{e}^{\mathrm{i}\langle k,x\rangle}\,\mathrm{d}k\,, \tag{3.27}$$

where $\widehat{P}(k,t)$ is the Fourier transform of $P(x,t;x_0,t_0)$.

From (3.23) we deduce that

$$\widehat{P}(k,t+1) = \frac{1}{d}\sum_{j=1}^{d}\cos(k_j)\widehat{P}(k,t) \tag{3.28}$$

with

$$\widehat{P}(k,t_0) = \mathrm{e}^{-\mathrm{i}\langle k,x_0\rangle}$$

as follows from (3.20). Using (3.27) and (3.28) one obtains the solution

$$P(x_1,t_1;x_0,t_0) = \frac{1}{(2\pi)^d}\int_{\mathsf{BZ}} \mathrm{e}^{\mathrm{i}\langle k,x_1-x_0\rangle}\left(\frac{1}{d}\sum_{j=1}^{d}\cos(k_j)\right)^{t_1-t_0}. \tag{3.29}$$

After solving the discrete heat equation we now aim to analyse the scaling limit when we scale both the lattice by $\varepsilon > 0$ and the time by $\tau > 0$, that is, we replace the lattice $\mathbb{Z}^d$ by $\varepsilon\mathbb{Z}^d$ and the unit time step by τ. Note that the left hand side of (3.29) is only a function of $x_1 - x_0$ and $t_1 - t_0$. Perfomrming the substitutions $t \to t/\tau$ and $x \to x/\varepsilon$ and $k \to \varepsilon k$, equation (3.29) becomes

$$P(x_1,t_1;x_0,t_0) = \frac{\varepsilon^d}{(2\pi)^d}\int_{\mathsf{BZ}_\varepsilon} \mathrm{e}^{\mathrm{i}\langle k,x_1-x_0\rangle}\left(\frac{1}{d}\sum_{j=1}^{d}\cos(\varepsilon k_j)\right)^{(t_1-t_0)/\tau}, \tag{3.30}$$

where $\mathsf{BZ}_\varepsilon = [-\pi/\varepsilon,\pi/\varepsilon]^d$. we now take the limit as ε and τ go to zero, keeping distances and time intervals fixed. We consider a volume Δx around x which is large with respect to the elementary lattice volume ε^d, but which is also sufficiently small to ensure that the transition probability remains nearly constant within Δx; this last requirement is also fulfilled if $(t_1 - t_0)/\tau$ is also large. This finally permits a so-called transition probability density $p = P/\varepsilon^d$ to be defined as

$$p(x_1,t_1;x_0,t_0)\Delta x_1 = \sum_{x_1'\in\Delta x_1} P(x_1',t_1;x_0,t_0) \approx \frac{\Delta x_1}{\varepsilon^d}P(x_1,t_1;x_0,t_0)\,, \tag{3.31}$$

and thus

$$p(x_1, t_1; x_0, t_0) = \lim_{\varepsilon,\tau\to 0} \frac{1}{(2\pi)^d} \int_{\mathsf{BZ}_\varepsilon} \mathrm{e}^{\mathrm{i}\langle k, x_1 - x_0\rangle} \Big(\frac{1}{d}\sum_{j=1}^{d} \cos(\varepsilon k_j)\Big)^{(t_1-t_0)/\tau} . \tag{3.32}$$

This limit is nontrivial only when ε and τ vanish in such a way that the ratio τ/ε^2 is kept fixed. One can show this by the expansion of the cosine

$$\Big(\frac{1}{d}\sum_{j=1}^{d} \cos(\varepsilon k_j)\Big)^{(t_1-t_0)/\tau} = \Big(1 - \frac{\varepsilon^2}{2d}k^2 + \cdots\Big)^{(t_1-t_0)/\tau} \to \mathrm{e}^{-(t_1-t_0)k^2}$$

in which the time scale has been fixed using

$$\tau = \frac{1}{2d}\varepsilon^2 .$$

Hence

$$\begin{aligned} p(x_1, t_1; x_0, t_0) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp\big(-(t_1-t_0)k^2 + \mathrm{i}\langle x_1 - x_0, k\rangle\big)\, \mathrm{d}k \\ &= \frac{1}{(4\pi(t_1-t_0)^{d/2}} \exp\Big(-\frac{(x_1-x_0)^2}{4(t_1-t_0)}\Big) . \end{aligned} \tag{3.33}$$

This is the well-known kernel of the diffusion equation in continuous space $\mathbb{R}^d$. It is positive, symmetric, and satisfies

$$\int_{\mathbb{R}^d} p(x, t; x_0, t_0)\, \mathrm{d}x = 1 . \tag{3.34}$$

$$\lim_{t_1\to t_0} p(x, t_1; x_0, t_0) = \delta(x_1 - x_0) . \tag{3.35}$$

$$\Big(\frac{\partial}{\partial t} - \Delta\Big) p(x, t; x_0, t_0) = 0 . \tag{3.36}$$

$$\int_{\mathbb{R}^d} p(x_2, t_2; x_1, t_1) p(x_1, t_1; x_0, t_0)\, \mathrm{d}x_1 = p(x_2, t_2; x_0, t_0) . \tag{3.37}$$

Condition (3.34) is the conservation law for the probabilities and is the continuous counterpart of (3.21), while (3.35) describes the initial condition. Equation (3.37) is the *diffusion equation* or *heat equation* . Finally, (3.37) expresses the obvious fact that the walker was certainly somewhere at an intermediate time t_1. This last relation, characteristic of a Markov process, is compatible with the convolution properties of Gaussian integrals.

Remark 3.20 (a) Note that we have considered above only the so-called forward derivative in time. We can repeat all our arguments for the backward differences to obtain convergence to the continuous equation (3.37). The continuous diffusion law/heat equation (3.37) is isotropic and translationally invariant (in $\mathbb{R}^d$), whereas the discrete version only presents the cubic symmetries.

(b) In our limiting procedure above we have fixed the time intervals $t_1 - t_0$ and the spatial distances $x_1 - x_0$ and considered the case where the spatial and the time steps ε and τ vanished, with the ratio τ/ε^2 kept fixed. To be precise, note that $t_1 - t_0$ and the components of $x_1 - x_0$ should be multiples of τ and ε respectively.

(c) Alternatively, we could have taken $x_1 - x_0$ and $t_1 - t_0$ large with respects to the spacings. Our approach here is comparable to the so-called ultraviolet limit in field theory, where a cutoff parameter (here $1/\varepsilon$ being a natural scale for the momenta k) tends to infinity, whereas physical (measurable) quantities are kept fixed. The cutoff for frequencies is τ^{-1} with $\tau^{-1} \sim \varepsilon^{-2}$. When a (Brownian) curve is followed at constant speed ε/τ, the typical distance behaves as

$$|x_1 - x_0| \sim |t_1 - t_0|^{\frac{1}{2}} .$$

(d) Defining an *characteristic exponent or critical exponent ν of the end-to-start distance* by

$$|x_1 - x_0| \sim |t_1 - t_0|^{\nu} ,$$

we see that Brownian motion (to be discussed in detail in Section 3.5 below) has *critical exponent $\nu = \frac{1}{2}$*.

(e) A *bond* in $\mathbb{Z}^d$ is an unordered pair $\{x, y\}$ with $x \neq y \in \mathbb{Z}^d$. A nearest-neighbour bond is a bond $\{x, y\}$ with $|x - y| = 1$. An oriented pair (x, y) is called a *step of the walk* with initial site x and terminal site y. A walk (path) ω in the lattice $\mathbb{Z}^d$ is a sequence of sites $\omega(0), \omega(1), \ldots, \omega(N)$, we call $|\omega| := N$ the *length* of the walk.

$$\begin{aligned} P(x_1, t_1; x_0, t_0) &= \frac{\#\{\text{paths joining } x_0 \text{ to } x_1 \text{ with } t_1 - t_0 \text{ steps}\}}{\#\{\text{paths originating from } x_0 \text{ with } t_1 - t_0 \text{ steps}\}} \\ &= \sum_{\substack{\omega, \partial\omega = \{x_0, x_1\} \\ |\omega| = t_1 - t_0}} \Big(\frac{1}{2d}\Big)^{|\omega|} , \end{aligned}$$

where $\partial\omega$ denotes the boundary of the path ω, i.e., $\partial\omega = \{x_0, x_1\}$.

$\diamond$

3.5 Brownian motion

In the last section we have seen that

$$p(x, y; t) = \frac{1}{(4\pi t)^{d/2}} \exp\Big(- \frac{|y - x|}{4t}\Big)$$

solves the heat equation

$$\frac{\partial u(x, t)}{\partial t} = \Delta u(x, t) , \quad t \geq 0, x \in \mathbb{R}^d .$$

In probability theory we typically write

$$\frac{\partial u(x, t)}{\partial t} = \frac{1}{2} \Delta u(x, t) , \quad t \geq 0, x \in \mathbb{R}^d , \tag{3.38}$$

and we thus re-define

$$p(x,y;t)=\frac{1}{(2\pi t)^{d/2}}\exp\Big(-\frac{|y-x|}{2t}\Big)\,. \tag{3.39}$$

The ultimate aim to to introduce Brownian motion as the scaling limit of the SRW as indicated in the previous section. First of all, a stochastic process (for continuous time) is a family $(X_t)_{t\ge 0}$ of random variables X_t in some state space. There are actually various ways how to proceed when we consider the state space $\mathbb{R}^d$ and our findings. One way is to characterise the transition probabilities via the transition probability density $p(x,y;t), x,y\in\mathbb{R}^d, t\ge 0$, that is,

(i) $p(x,\cdot;t)$ is a probability density function for all $t\ge 0, x\in\mathbb{R}^d$.

(ii) $p(x,B;t):=\int_B p(x,y;y)\,\mathrm{d}y$ is the transition probability that $X_t\in B$ when $x_0=x$, $B\in\mathcal{B}(\mathbb{R}^d)$.

(iii) For all $t,s\ge 0, B\in\mathcal{B}(\mathbb{R}^d)$,

$$p(x,B;t+s)=\int_{\mathbb{R}^d} p(x,\mathrm{d}y;t)p(y,B;s)\,.$$

The idea is then to define a semigroup of transition probabilities, i.e., for any bounded measurable function f,

$$P_tf(x):=\int_{\mathbb{R}^d} p(x,\mathrm{d}y;t)\,f(y)\,,x\in\mathbb{R}^2\,,$$

such that the process is fully characterised by this semigroup so that

$$P_x(X_t\in B):=p(x,B;t)\qquad\text{for } B\in\mathcal{B}(\mathbb{R}^d)\,.$$

This is the general theory of Markov processes and an interested reader might consult [Kal02] or reference in there.

An alternative way is to define all finite-dimensional distributions of the process $(X_t)_{t\ge 0}$. First, suppose that $x_0=x\in\mathbb{R}^d$, that is , the initial site of the process is fixed, one easily generalises this to any initial probability distribution $\mu\in\mathcal{M}_1(\mathbb{R}^d)$.

A finite-dimensional distribution is defined for any finite time vector $\underline{t}=(t_1,\dots,t_n)\in\mathbb{R}_+^n$ with $0=t_0\le t_1\le t_2\le\cdots\le t_n, B\in\mathcal{B}(\mathbb{R}^{dn})$. The *finite-dimensional distributions of the process* $(X_t)_{t\ge 0}$ are the probabilities of the vector $(X_{t_1},\dots,X_{t_n})$ of the process evaluated at the given times, i.e., for all $B\in\mathcal{B}(\mathbb{R}^{dn})$,

$$P_x(\{(X_{t_1},\dots,X_{t_n})\in B\}):=\int_B\prod_{i=1}^n p(y_{i-1},y_i;t_i-t_{i-1})\,\mathrm{d}y_1\cdots\mathrm{d}y_n\,,\quad y_0:=x\,. \tag{3.40}$$

Suppose we have defined all finite-dimensional distributions as of (3.40), the question is then whether that determines uniquely the process $(X_t)_{t\ge 0}$. We review standard measure theory material where this is the case for certain classes of finite-dimensional distributions. Our process$(X_t)_{t\geq 0}$ takes values in the *path space*

$$\Omega := \mathbb{R}^{[0,\infty)} = \{\omega\colon [0,\infty) \to \mathbb{R}^d\}\,,$$

and we shall define a probability measure on Ω which is compatible with the finite-dimensional distributions. Measurable sets of the form

$$\begin{aligned} C =&\{\omega \in \Omega\colon (\omega(t_1),\dots,\omega(t_n)\in A\}\,,\\ &n\in\mathbb{N}, A\in\mathcal{B}(\mathbb{R}^n), t_i\in[0,\infty), i=1,\dots,n\,, \end{aligned} \tag{3.41}$$

are called *cylinder sets or cylinder events* . We denote by $\mathcal{C}$ the collection of all cylinder sets and by $\mathcal{F} = \sigma(\mathcal{C})$ the smallest σ-algebra on Ω containing all cylinder events.

Definition 3.21 Denote T the set of all finite sequences $\underline{t} = (t_1,\dots,t_n), t_i \geq 0$, and write $|\underline{t}| = n$ for its length. Then $(Q_{\underline{t}})_{\underline{t}\in\mathsf{T}}$ is a family of finite-dimensional distributions if $Q_{\underline{t}} \in \mathcal{M}_1(\mathbb{R}^{d|\underline{t}|})$. The set $(Q_{\underline{t}})_{\underline{t}\in\mathsf{T}}$ is said to be a *consistent family of finite-dimensional distributions* if

(i) If $\underline{s} = (t_{i_1},\dots,t_{i_n}) \in \mathsf{T}$ is a permutation of $\underline{t}\in\mathsf{T}$, then for every $A_i \in \mathcal{B}(\mathbb{R}^d), i = 1,\dots,|\underline{t}| = n$,

$$Q_{\underline{t}}(A_1\times\cdots\times A_n) = Q_{\underline{s}}(A_{i_1}\times\cdots\times A_{i_n})\,.$$

(ii) For every $\underline{t} = (t_1,\dots,t_n)\in\mathsf{T}$ and $\underline{s} = (t_1,\dots,t_{n-1})\in\mathsf{T}$, $n\in\mathbb{N}$,

$$Q_{\underline{t}}(A\times\mathbb{R}) = Q_{\underline{s}}(A) \quad A\in\mathcal{B}(\mathbb{R}^{n-1})\,.$$

Now it is readily clear that given a probability measure $P \in \mathcal{M}_1(\Omega,\mathcal{F})$ on the path space, then one obtains a consistent family of finite-dimensional distributions via

$$Q_{\underline{t}}(A) := P((\omega(t_1),\dots,\omega(t_n))\in A)\,,\quad ,\underline{t}\in\mathsf{T}, |\underline{t}| = n, A\in\mathcal{B}(\mathbb{R}^{dn})\,. \tag{3.42}$$

Theorem 3.22 (Daniell 1018; Kolmogorov 1933) *Let be a consistent family of finite-dimensional distributions. Then there exists a probability measure* $\in \mathcal{M}_1(\Omega,\mathcal{F})$ *such that* (3.42) *holds.*

Proof. See standard measure theory and probability books, [BB01, Kal02, Dur19]. □

The idea is then to define a measure via (3.40) and identify the process $(X_t)_{t\geq 0}$ we called earlier Brownian motion with the corresponding path measure. Before we do that, let us check the fundamental properties of the process obtain from our probability transition density.

Exercise 3.23 Using (3.40), show that the increments $\{X_{t_i} - X_{t_{i-1}}; 1\leq i\leq n\}, n\in\mathbb{N}$, are independent and normally distributed, i.e., increment $X_{t_i} - X_{i_{i-1}}$ has law

$$p(0,\cdot;t_i - t_{i-1}) = \mathsf{N}(0, t_i - t_{i-1})\,.$$

☕☕

Definition 3.24 (Brownian motion) A process $(B_t)_{t\ge 0}$ of $\mathbb{R}$-valued random variables is called *standard (or one-dimensional) Brownian motion with start at* $x \in \mathbb{R}$ if the following holds.

(a) $B_0 = x$.

(b) For all $0 \le t_1 \le \cdots \le t_n, n \in \mathbb{N}$, the increments $B_{t_n} - B_{t_{n-1}}, \dots, B_{t_2} - B_{t_1}$ are independent $\mathbb{R}$ -valued random variables.

(c) For $t \ge 0$ for all $h > 0$, the increments $B_{t+h} - B_t$ are $\mathsf{N}(0, h)$ distributed.

(d) Almost surely, the function $t \mapsto B_t$ is continuous.

A process $(B_t)_{t\ge 0}$ of $\mathbb{R}^d$-valued random variables is called *d-dimensional Brownian motion with start in* $x \in \mathbb{R}^d$ if all coordinate processes $(B_t^{(i)})_{t\ge 0}$ are standard Brownian motion with start in $x_i \in \mathbb{R}$, $i = 1, \dots, d$.

Note that the σ-algebra $\mathcal{F}$ generated by all cylinder doe snot include events that the whole Brownian motion path is continuous. Thus, in order to proceed in line with our definition of Brownian motion, we need to obtain continuous version of the process defined by the probability measure on the path space. This can be down but we skip these details here and refer the interested reader to the book [KS98]. Furthermore, there is an elegant construction of Brownian motions which immediately provides continuity of the random paths, the so-called Lévy construction, see for example [MP10]. We finally briefly discuss another way to construct Brownian motion, namely by directly studying the scaling limit of the corresponding random walk using our techniques and methods from the CLT, Theorem 2.20. The key idea is to scale a discrete time random walk with linear extrapolation to continuous time and where the scale refers to the number terms in the sum. This way one obtain a sequence $(X^{(n)})_{\nu\mathbb{N}}$ of processes $X^{(n)} = (X_t^{(n)})_{t\ge 0}$, and one can make use of the following fundamental result. A continuous process is a process with almost surely continuous paths.

Theorem 3.25 *Let $(X^{(n)})_{n\in\mathbb{N}}$ be a tight sequence of continuous processes $X^{(n)} = (X_t^{(n)})_{t\ge 0}$ with the following property that, whenever $0 \le t_1 \le \cdots \le t_m < \infty, m \in \mathbb{N}$, then*

$$(X_{t_1}^{(n)}, \dots, X_{t_m}^{(n)}) \underset{n\to\infty}{\overset{\mathrm{d}}{\longrightarrow}} (W_{t_1}, \dots, W_{t_m}) \,, \tag{3.43}$$

for some process $(W_t)_{t\ge 0}$. Let $P_n \in \mathcal{M}_1(\mathcal{C}([0,\infty;\mathbb{R}), \mathcal{F})$ be induced by the process $X^{(n)}$. Then

$$P_n \underset{n\to\infty}{\overset{\mathrm{P}}{\longrightarrow}} P \,,$$

under which the coordinate mapping process $W_t(\omega) = \omega(t)$ satisfies (3.43).

Remark 3.26 The crucial step is to show tightness for the sequence of processes. For continuous processes one needs a version of the Arzela-Ascoli theorem to construct compact sets in the space of all continuous paths. Details can be found in [KS98].

◇

We define now the sequence of scaled $\mathbb{R}$-valued processes. Suppose that $(\xi_j)j \in \mathbb{N}$ is a sequence of independent, identically distributed $\mathbb{R}$-valued random variables with mean zero and variance $\sigma^2 \in (0, \infty)$. Define the discrete time random walk $(S_k)_{k \in \mathbb{N}_0}$ by

$$S_0 := 0 \qquad S_k := \sum_{j=1}^{k} \xi_j \,, \quad k \in \mathbb{N} \,.$$

A continuous time process $Y = (Y_t)_{t \geq 0}$ can be obtained from $(S_k)_{k \in \mathbb{N}_0}$ by linear interpolation,

$$Y_t = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\xi_{\lfloor t \rfloor + 1} \,, \quad t \geq 0 \,,$$

where $\lfloor t \rfloor$ is the greatest integer $\leq t$. We scale time by n and space by $\sqrt{n}$ and obtain a sequence $(X^{(n)})_{n \in \mathbb{N}}$ of piecewise continuous processes $(X_t^{(n)})_{t \geq 0}$,

$$X_t^{(n)} := \frac{1}{\sigma \sqrt{n}} Y_{nt} \,, \quad t \geq 0 \,. \tag{3.44}$$

Pick $s = \frac{k}{n}$ and $t = \frac{k+1}{n}$, $k \in \mathbb{N}$, then

$$x_t^{(n)} - X_s^{(s)} = \frac{1}{\sigma \sqrt{n}} \xi_{k+1}$$

is independent of $\sigma(\xi_1, \ldots, \xi_k)$. Furthermore, $X_t^{(n)} - X_s^{(n)}$ has zero mean and variance $t - s$. Thus $X^{(n)}$ is approximately Brownian motion.

Theorem 3.27 *For all* $0 \leq t_1 < t_2 < \cdots < t_m, m \in \mathbb{N}$,

$$(X_{t_1}^{(n)}, \ldots, X_{t_m}^{(n)}) \xrightarrow[n \to \infty]{\mathrm{d}} (B_{t_1}, \ldots, B_{t_m}) \,,$$

where $(B_t)_{t \geq 0}$ *is standard Brownian motion.*

Proof. The proof is a straightforward application of the CLT, and is therefore left as an exercise. It suffices to show the proof for $m = 2$. The shortest route is by using characteristic functions exploiting the independence of the given sequence.

□

Theorem 3.28 (The Invariance Principle of Donsker (1951)) *Suppose that* $(\xi_j)j \in \mathbb{N}$ *is a sequence of independent, identically distributed* $\mathbb{R}$*-valued random variables with mean zero and variance* $\sigma^2 \in (0, \infty)$*. Define the continuous process* $X^{(n)}$ *as in* (3.44) *above, and let* $P_n \in \mathcal{M}_1(\mathcal{C}([0, \infty))$ *be induced by* $X^{(n)}$*. Then* P_N *converges weakly (as probability measures) to a measure* $P^* \in \mathcal{M}_1(\mathcal{C}([0, \infty))$ *under which the coordinate mapping process*

$$W_t \colon \mathcal{C}([0, \infty)) \to \mathbb{R}, (\omega) \mapsto W_t(\omega) = \omega(t)$$

on $\mathcal{C}([0, \infty))$ *is a standard, one-dimensional Brownian motion, i.e.,* $W_t = B_t$*, and* $B_0 = 0$.

To prove this theorem we need a couple of notations and results. The proof is below.

We need a characterisation of tightness in the space $\mathcal{C}([0,\infty);\mathbb{R})$ of continuous paths with horizon $[0,\infty)$ and state space $\mathbb{R}$. For any path $\omega \in \mathcal{C}([0,\infty);\mathbb{R})$, $T > 0$, and $\delta > 0$, the *modulus of continuity* on $[0,T]$ is

$$m^T(\omega,\delta) := \max_{|s-t|\le\delta, s,t\in[0,T]} \{|\omega(s)-\omega(t)|\} .$$

We need the following version of the Arzelà-Ascoli theorem adapted to $\mathcal{C}([0,\infty);\mathbb{R})$.

Proposition 3.29 *$A \subset \mathcal{C}([0,\infty);\mathbb{R})$ has a compact closure (i.e., $\overline{A}$ is compact) if and only if*

$$\text{(i)} \sup_{\omega\in A}\{|\omega(0)|\} < \infty \quad \textit{and} \quad \text{(ii)} \lim_{\delta\downarrow 0}\sup_{\omega\in A}\{m^T(\omega,\delta)\} = 0 , \quad \forall T > 0 .$$

Proof. This technical result is proved in [KS98, MP10, Kal02]. □

A sequence $(X^{(n)})_{n\in\mathbb{N}}$ of continuous processes $X^{(n)} = (X_t^{(n)})_{t\ge 0}$ is tight when the sequence of probability measures $P_n \in \mathcal{M}_1(\mathcal{C}([0,\infty);\mathbb{R}))$ is tight, where P_n is induced by $X^{(n)}$.

Theorem 3.30 *A sequence $(P_n)_{\in\mathbb{N}}$ of probability measures $P_n \in \mathcal{M}_1(\mathcal{C}([0,\infty))$ is tight if and only if the following two conditions hold.*

$$\lim_{\lambda\to\infty}\sup_{n\in\mathbb{N}} P_n(|\omega(0)| > \lambda) = 0 . \tag{3.45}$$

$$\lim_{\delta\downarrow 0}\sup_{n\in\mathbb{N}} P_n(m^T(\omega,\delta) > \varepsilon) = 0 \quad \forall T > 0, \forall\varepsilon > 0 . \tag{3.46}$$

Proof. Suppose $(P_n)_{n\in\mathbb{N}}$ is tight. Pick $\eta > 0$. Then there exists a compact set K with $P_n(K) \ge 1-\eta$ for all $n \in \mathbb{N}$, and thus $P_n(K^c) \le \eta$. For $\lambda > 0$ sufficiently large it follows that $|\omega(0)| \le \lambda$ for $\omega \in K$, see Proposition 3.29. For given $T > 0$ and $\varepsilon > 0$, there exists a $\delta_0 > 0$ such that $m^T(\omega,\delta) \le \varepsilon$ for all $0 < \delta < \delta_0$ and for every $\omega \in K$. Thus one obtains (3.45) and (3.46).
Now suppose that (3.45) and (3.46) hold. Given $T \in \mathbb{N}$ and $\eta > 0$, we choose $\lambda > 0$ in such a way that

$$\sup_{n\in\mathbb{N}} P_n(|\omega(0)| > \lambda) \le \frac{\eta}{2^{T+1}} .$$

Furthermore, we choose $\delta_k > 0$ such that

$$\sup_{n\in\mathbb{N}} P_n\Big(m^T(\omega,\delta_k) > \frac{1}{k}\Big) \le \frac{\eta}{2^{T+k+1}} .$$

$$A_T := \{\omega \in \mathcal{C}([0,\infty))\colon |\omega(0)| \le \lambda; m^T(\omega,\delta_k) \le 1/k, k \in \mathbb{N}\} ,$$

and

$$A := \bigcap_{T=1}^{\infty} A_T .$$

Then, according to Proposition 3.29, A is compact and

$$P_n(A_T) \geq 1 - \sum_{k=0}^{\infty} \frac{\eta}{2^{T+k+1}} = 1 - \frac{\eta}{2^T} \quad \text{and } P_n(A) \geq 1 - \eta\,.$$

Thus $(P_n)_{n\in\mathbb{N}}$ is tight. □

Proof of Theorem 3.28. In light of Theorem 3.25 and Theorem 3.27 we only need to show that $(X^{(n)})_{n\in\mathbb{N}}$ is tight. We shall use Theorem 3.30 adapted to the specific setting of our scaled processes. Namely, one can show (see [KS98]) via CLT estimates and tail normal estimate, that

$$\forall \varepsilon > 0,\ \lim_{\delta\downarrow 0} \limsup_{n\to\infty} \frac{1}{\delta} P\Big(\max_{1\leq j\leq \lfloor n\delta\rfloor+1} \{|S_j|\} > \varepsilon\sigma\sqrt{n}\Big) = 0\,,$$

$$\forall T > 0,\ \lim_{\delta\downarrow 0} \limsup_{n\to\infty} P\Big(\max_{\substack{1\leq j\leq \lfloor n\delta\rfloor+1 \\ 0\leq k\leq \lfloor nT\rfloor+1}} \{|S_{j+k} - S_k|\} > \varepsilon\sigma\sqrt{n}\Big) = 0\,.$$

Then it follows that these two estimates imply (3.45) and (3.46) in Theorem 3.30. □

Definition 3.31 The (unique) measure $P^* \in \mathcal{M}_1(\mathcal{C}([0,\infty))$ in Theorem 3.28, under which the coordinate mapping process is a standard, one-dimensional Brownian motion, is called *Wiener measure*. We denote P_x the Wiener measure for one-dimensional Brownian motion with deterministic start in $x \in \mathbb{R}$ (i.e., $B_0 = x$), and denote P_μ the Wiener measure for one-dimensional Brownian motion with initial probability distribution $\mu \in \mathcal{M}_1(\mathbb{R})$.

4 Large Deviation Theory

4.1 Introduction and Definition

We start with an easy example before motivating the theory and coming up with definitions.

Example 4.1 Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$. Denote $S_n = \sum_{i=1}^n X_i$ the number of successes (e.g., coin tossing), then for every $a > 1$, we have that

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq an) = -I(a)\,,$$

where

$$I(z) = \begin{cases} \log 2 + z\log z + (1-z)\log(1-z) & \text{if } z \in [0,1], \\ \infty & \text{otherwise} \end{cases}.$$

This we justify as follows. Claim for $a > 1$ is trivial. For $a \in (\frac{1}{2}, 1]$ we observe that $P(S_n \geq an) = 2^{-n}\sum_{k\geq an}\binom{n}{k}$, which yields the estimate

$$2^{-n}Q_n(a) \leq P(S_n \geq an) \leq (n+1)2^{-n}Q_n(a),$$

where $Q_n(a) = \max_{k \geq an} \binom{n}{k}$. Maximum is attained at $k = [an]$, the smallest integer $\geq an$. Stirling's formula gives therefore

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n(a) = -a \log a - (1-a)\log(1-a).$$

Upper and lower bound merge on an exponential scale as $n \to \infty$, and henceforth we arrive at the desired statement. Since $\mathbb{E}(X_1) = \frac{1}{2}$ and $a > \frac{1}{2}$ the statement deals with large deviations in the upward direction. From the symmetry it is clear that the same holds for $P(S_n \leq an)$ with $a < \frac{1}{2}$. This is seen by $I(1-z) = I(z)$.

♣

The zero of the function I in Example 4.1 corresponds the SLLN, Theorem 2.10, as it implies that

$$\sum_{n\in\mathbb{N}} P\Big(\Big|\frac{1}{n}S_n - 1/2\Big| > \delta\Big) < \infty$$

for every $\delta > 0$. Furthermore, $I'(1/2) = 0$ and $I''(1/2) = 4 = 1/\sigma^2$ with $\sigma^2 = \text{Var}(X_1) = \frac{1}{4}$. Recall that an application of Chebychev's inequality gives an estimate

$$P(|\frac{1}{n}S_n - \frac{1}{2}| > \delta) \leq \frac{1}{\delta^2 n}\,,$$

but this estimate is of order $\frac{1}{n}$ and therefore not summable for an application of Borel-Cantelli Lemma. It is therefore desirable to find out exactly how fast the large deviation probabilities $P(|1/nS_n - 1/2| > \delta)$ decay. In this chapter we are studying deviations of the order n, so well beyond what is described by the CLT. Derivations of this size are called '*large*'. Suppose $(X_i)_{i\in\mathbb{N}}$ are i.i.d. random variables with mean μ. A large deviation event $\{S_n = \sum_{i=1}^n X_i \geq (\mu + a)n\}, a > 0$, (or, $\{S_n = \sum_{i=1}^n X_i \leq (\mu + a)n\}, a < 0$) has a probability which goes to zero as $n \to \infty$. Under certain conditions of the tail of the distribution of X_1, the decay is exponential in n as we have seen in Example 4.1 above:

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq (\mu + a)n) = -I(a) < \infty\,, \quad I \geq 0\,.$$

The *large deviation principle (LDP)* which we define below characterises the limiting behaviour, as $n \to \infty$, of a family of probability measures $(\mu_n)_{n\in\mathbb{N}}$ on some measurable space $(E, \mathcal{B})$ in terms of a rate function. This characterisation is via asymptotic upper and lower exponential bounds on the value that μ_n assigns to measurable subsets of E. Throughout, E is a topological space so that open and closed subsets of E are well-defined. The simplest situation is when $\mathcal{B}$ is the Borel-σ-algebra $\mathcal{B}(E)$.

Definition 4.2 (Rate function) A *rate function* I is a lower semicontinuous mapping $I\colon E \to [0, \infty]$, that is, for all $\alpha \in [0, \infty)$, the *level set* $\mathcal{L}_I(\alpha) := \{x \in E\colon I(x) \leq \alpha\}$ is a closed subset of E. A *good rate function* I is a rate function for which all the level sets $\mathcal{L}_I(\alpha)$ are compact subsets of E. The *domain of* I is $\mathcal{D}(I) = \{x \in E\colon I(x) < \infty\}$.

Definition 4.3 (Large deviation principle) (a) A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures $\mu_n \in \mathcal{M}_1(E, \mathcal{B})$ satisfies *the large deviation principle (LDP) with rate (speed)* n *and rate function* I if, for all $M \in \mathcal{B}$,

$$-\inf_{x\in \mathrm{int}(M)} I(x) \le \liminf_{n\to\infty} \frac{1}{n}\log\mu_n(M) \le \limsup_{n\to\infty}\frac{1}{n}\log\mu_n(M) \le -\inf_{x\in\overline{M}} I(x)\,. \tag{4.1}$$

(b) When $\mathcal{B}(E)\subset\mathcal{B}$, the LDP is equivalent to the following bounds:

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(K) &\le -\inf_{x\in F} I(x)\,, \qquad \text{for all closed } F\subset E\,,\\ \liminf_{n\to\infty}\frac{1}{n}\log\mu_n(G) &\ge -\inf_{x\in G} I(x)\,, \qquad \text{for all open } G\subset E\,.\end{aligned} \tag{4.2}$$

(c) A set $M\in\mathcal{B}$ is called *I-continuity set* if

$$\inf_{x\in\mathrm{int}(M)} I(x) = \inf_{x\in\overline{M}} I(x) =: I_M\,,$$

in which case

$$\lim_{n\to\infty}\frac{1}{n}\log\mu_n(M) = -I_M\,.$$

Remark 4.4 (a) If we are dealing with non-atomic measures we have that $\mu_n(\{x\})=0$ for every $x\in E$. Thus, if the lower bound in (4.1) was to hold with the infimum over M instead of its interior $\mathrm{int}(M)$, we would conclude that $I\equiv\infty$, contradiction the upper bound of (4.1) because $\mu_n(E)=1$ for all n.

(b) Since $\mu_n(E)=1$ for all 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, this means that there exists at least one point x for which $I(x)=0$. Furthermore, the upper bound trivially holds whenever $\inf_{x\in\overline{M}} I(x)=0$, while the lower bound trivially holds whenever

$$\inf_{x\in\mathrm{int}(M)} I(x)=\infty\,.$$

(c) Suppose that I is a rate function. Then (4.1) is equivalent to the following two bounds:

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

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(M)\le-\alpha\,. \tag{4.3}$$

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

$$\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(M)\ge -I(x)\,. \tag{4.4}$$

(d) The rate (speed or scale) of a large deviation principle can be any sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n \to \infty$ as $n \to \infty$, and the obtain our upper and lower bounds replacing $\frac{1}{n}$ by $\frac{1}{a_n}$.

(e) The CLT tells us by how much the partial sum *normally* exceeds its average, namely by an order of $\sqrt{n}$. More precisely,

$$P(S_n - n\mu \geq \sqrt{n}x) \to 1 - \Phi(x/\sigma)\,, \text{ as } n \to \infty\,,$$

where Φ is the distribution function of the standard normal law. Thus, for any sequence $(a_n)_{n\in\mathbb{N}}$ with $\sqrt{n} \ll a_n \ll n$, we still have

$$P(S_n - \mu n \geq a_n) \to 0\,, \text{ as } n \to \infty\,,$$

and neither the CLT nor the large deviation principle tell us how fast this convergence is. This question is in the remit of the *moderate deviation principle*.

◇

The following lemma states roughly that the rate of growth for a finite sum of sequences equals the maximal rate of growth of the summands.

Lemma 4.5 (Laplace Principle) *For a sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n \to \infty$ as $n \to \infty$ and a finite number N of nonnegative sequences $(b_n^{(1)})_{n\in\mathbb{N}}, \dots, (b_n^{(N)})_{n\in\mathbb{N}}$, the following holds.*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \sum_{i=1}^N b_n^{(i)} = \max_{1\leq i\leq N} \limsup_{n\to\infty} \frac{1}{a_n} \log b_n^{(i)}\,.$$

Proof. Observe that

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

and conclude with the statement dividing by a_n and taking the limes superior. Suppose $b_n^{(1)} = \max_{1\leq i\leq N} \log b_n^{(i)}$, then

$$\log\Big(\sum_{i=1}^N b_n^{(i)}\Big) = \log b_n^{(1)} + \log\Big(1 + \sum_{i=2}^N \frac{b_n^{(i)}}{b_n^{(1)}}\Big)\,.$$

□

Definition 4.6 (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 (4.3) holds for every α and all compact subsets of $\mathcal{L}_I(\alpha)^c$, and the lower bound (4.4) holds for all measurable subsets.

4.2 Combinatorial Techniques for finite sample spaces

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 $S = k_{\rm B} \log W$ 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 $k_{\rm B}$ (the Boltzmann constant $k_{\rm B} = 1.3806 \times 10^{-23}{\rm m}^2{\rm kgs}^{-2}{\rm K}^{-1}$) 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., μ 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 μ corresponds to the degree of uncertainty about the actual micro stateω when only μ is known, and can thus be measured by $\log|\mathsf{T}_n(\mu)|$*, the logarithmic number of micro states leading to μ.*

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

$$\mathsf{T}_n(\nu) := \{\omega \in E^n \colon L_n^\omega = \nu\} \tag{4.5}$$

is the set of all $\omega \in E^n$ of *type* μ.

Definition 4.7 Denote $\mathcal{L}_n$ the set of all possible types of sequences of length in E, i.e.,

$$\mathcal{L}_n := \{\nu \in \mathcal{M}_1(E) \colon \nu = L_n^\omega \text{ for some } \omega \in E^n\}\,.$$

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

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

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

Proposition 4.8 (Entropy as degree of ignorance) *Let $\mu_n, \mu \in \mathcal{M}_1(E)$ be probability measures such that $\mu_n \to \mu$ as $n \to \infty$ and $n\mu(x) \in \mathbb{N}_0$ for all $x \in E$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log|\mathsf{T}_n(\mu_n)| = -\sum_{x\in E} \mu(x) \log \mu(x)\,. \tag{4.6}$$

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]. □

Definition 4.9 (Shannon Entropy) Suppose E is finite and $\mu \in \mathcal{M}_1(E)$. The *(Shannon) entropy of* μ is defined as

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

Definition 4.10 (Relative entropy)
Suppose E is finite and $\mu, \nu \in \mathcal{M}_1(E)$. For $\mu \in \mathcal{M}_1(E)$ denote

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

its support. The *relative entropy of* ν *with respect to* μ is

$$\mathsf{H}(\nu|\mu) := \begin{cases} \sum_{x\in E}\nu(x)\log\frac{\nu(x)}{\mu(x)} & \text{if } E_\nu \subset E_\mu\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{4.7}$$

Exercise 4.11 (Properties of relative entropy) Show that $\mathsf{H}(\cdot|\mu)$ is (i) nonnegative and convex, (ii) $\mathsf{H}(\cdot|\mu)$ is finite on $\{\nu \in \mathcal{M}_1(E) \colon E_\nu \subset E_\mu\}$, (iii) $\mathsf{H}(\cdot|\mu)$ 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 $\mathcal{M}_1(E)$. 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

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

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

$$\mathrm{d}(\mu,\nu) := \frac{1}{2}\sum_{x\in E}|\mu(x)-\nu(x)|\,, \tag{4.8}$$

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

Exercise 4.12 Show that, according to the SLLN, Theorem 2.10,

$$\mathrm{d}(L_n,\mu) \underset{n\to\infty}{\longrightarrow} 0 \ \text{ a.s.}\,.$$

☕☕

In the following theorem we derive the large deviation statement for L_n away from μ.

Theorem 4.13 (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 μ_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*

$$I_\mu(\nu) = \mathsf{H}(\nu|\mu)\,.$$

For the proof we shall need the following two lemmas.

Lemma 4.14 *If $x \in \mathsf{T}_n(\nu), \nu \in \mathcal{L}_n$, then*

$$P((X_1,\dots,X_n)=x) = \exp\big(-n(\mathsf{H}(\nu)+\mathsf{H}(\nu|\mu))\big)\,. \tag{4.9}$$

Proof.

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

Then, using independence, for $x=(x_1,\dots,x_n)\in\mathsf{T}_n(\nu)\subset E^n$,

$$P((X_1,\dots,X_n)=x) = \prod_{i=1}^n \mu(x_i) = \prod_{y\in E}\mu(y)^{n\nu(y)} = \exp\big(n\sum_{y\in E}\nu(x)\log\mu(x)\big)\,.$$

□

Lemma 4.15 *(a)* $|\mathcal{L}_n| \le (n+1)^{|E|}$.

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

$$\frac{1}{p_1(n)}\,\mathrm{e}^{n\mathsf{H}(\nu)} \le |\mathsf{T}_n(\nu)| \le p_2(n)\,\mathrm{e}^{n\mathsf{H}(\nu)}\,.$$

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) $\mathsf{T}_n(\nu)$ 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 $|\mathsf{T}_n(\nu)|$ is multinomial,

$$|\mathsf{T}_n(\nu)| = \frac{n!}{\prod_{x\in E}(n\nu(x))!}\,.$$

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

$$n\log\frac{n}{\mathrm{e}} \le \log n! \le n\log\frac{n}{\mathrm{e}} + c_1\log n + c_2\,.$$

Now,

$$\begin{aligned}\log|\mathsf{T}_n(\nu)| &\le \log n! - \sum_{x\in E}\log(n\nu(x))! \le n\log\frac{n}{\mathrm{e}} - \sum_{x\in E} n\nu(x)\log\frac{n\nu(x)}{\mathrm{e}} + c_1\log n + c_2\\ &= n\mathsf{H}(\nu) + c_1\log n + c_2\,,\end{aligned}$$

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 4.13. Pick a Borel set $A \subset \mathcal{M}_1(E)$. Then, using the upper bound in Lemma 4.15,

$$
\begin{aligned}
P(L_n \in A) &= \sum_{\nu \in \mathcal{L}_n \cap A} P(L_n = \nu) = \sum_{\nu \in \mathcal{L}_n \cap A} \sum_{x \in \mathsf{T}_n(\nu)} P(X = (X_1, \dots, X_n) = x) \\
&\le \sum_{\nu \in \mathcal{L}_n \cap A} p_2(n) \mathrm{e}^{n \mathsf{H}(\nu)} \, \mathrm{e}^{-n(\mathsf{H}(\nu) + \mathsf{H}(\nu|\mu))} \\
&\le (n+1)^{|E|} p_2(n) \, \mathrm{e}^{-n \inf_{\nu \in A \cap \mathcal{L}_n} \mathsf{H}(\nu|\mu)} \,.
\end{aligned}
$$

The lower bound reads

$$
\begin{aligned}
P(L_n \in A) &= \sum_{\nu \in \mathcal{L}_n \cap A} P(L_n = \nu) \ge \sum_{\nu \in \mathcal{L}_n \cap A} \frac{1}{p_1(n)} \mathrm{e}^{n \mathsf{H}(\nu|\mu)} \\
&\ge \frac{1}{p_1(n)} \, \mathrm{e}^{-n \inf_{\nu \in A \cap \mathcal{L}_n} \mathsf{H}(\nu|\mu)} \,.
\end{aligned}
$$

Since

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

we obtain

$$
\begin{aligned}
\limsup_{n \to \infty} \frac{1}{n} \log P(L_n \in A) &= - \liminf_{n \to \infty} \{ \inf_{\nu \in A \cap \mathcal{L}_n} \mathsf{H}(\nu|\mu) \} \\
\liminf_{n \to \infty} \frac{1}{n} \log P(L_n \in A) &= - \limsup_{n \to \infty} \{ \inf_{\nu \in A \cap \mathcal{L}_n} \mathsf{H}(\nu|\mu) \} \,.
\end{aligned}
$$

The desired upper bound of the large deviation principle in Theorem 4.13 follows, since $A \cap \mathcal{L}_n \subset A$ for all n.

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

$$
\{ \nu' \in \mathcal{M}_(E) \colon \mathrm{d}(\nu', \nu) < \delta \}
$$

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 $x \in E$: $|\nu(x) - \nu'(x)| \le C/n$ for some $C > 0$. Thus there exist a sequence $\nu_n \in A \cap \mathcal{L}_n$ such that $\nu_n \to \nu$ as $n \to \infty$. Moreover, without loss of generality, we may assume that $E_{\nu_n} \subset E_\mu$, and hence

$$
- \limsup_{n \to \infty} \{ \inf_{\nu' \in A \cap \mathcal{L}_n} \mathsf{H}(\nu'|\mu) \} \ge - \lim_{n \to \infty} \mathsf{H}(\nu_n|\mu) = -\mathsf{H}(\nu|\mu) \,.
$$

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

$$
- \limsup_{n \to \infty} \{ \inf_{\nu' \in A \cap \mathcal{L}_n} \mathsf{H}(\nu'|\mu) \} \ge - \inf_{\nu \in \mathrm{int} A} \mathsf{H}(\nu|\mu) \,.
$$

□

Exercise 4.16 Prove that for every open set $A \subset \mathcal{M}_1(E)$,

$$-\lim_{n\to\infty}\{\inf_{\nu\in A\cap\mathcal{L}_n} \mathsf{H}(\nu|\mu)\} = \lim_{n\to\infty}\frac{1}{n}\log P(L_n \in A) = -\inf_{\nu\in A}\mathsf{H}(\nu|\mu)\,.$$

4.3 Cramér Theorem, Varadhan Lemma, and basic principles

We now let $E = \mathbb{R}^d$ and let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent, identically distributed $\mathbb{R}^d$-valued random variables with law $\mu \in \mathcal{M}_1(\mathbb{R}^d)$. Recall the partial sum $S_n = \sum_{i=1}^n X_i$. The *empirical mean* is

$$\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i\,. \tag{4.10}$$

Definition 4.17 (Logarithmic moment generating function) Let $\mu \in \mathcal{M}_1(\mathbb{R}^d)$. The *logarithmic moment generating function* associated with μ is defined as

$$\Lambda(\lambda) := \log\mathbb{E}\Big[\mathrm{e}^{\langle\lambda,X_1\rangle}\Big]\,,\quad \lambda\in\mathbb{R}^d\,, \tag{4.11}$$

where the expectation is with respect to μ. Sometimes Λ is also called the *cumulant generating function*.

Note that $\Lambda(0) = 0$, and while $\Lambda(\lambda) > -\infty$ for all λ, it is possible to have $\Lambda(\lambda) = \infty$. We denote μ_n the law of the empirical mean $\widehat{S}_n$ under $\mu^{\otimes n}$. From the WLLN, Theorem2.2, we know that for $m := \mathbb{E}[X_1] = \int_{\mathbb{R}^d} x\,\mu(\mathrm{d}x)$,

$$\widehat{S}_n \xrightarrow[n\to\infty]{\mathrm{P}} m\,.$$

Hence, $\mu_n(F) \underset{n\to\infty}{\to} 0$ for any closed set F such that $m \notin F$. The logarithmic rate of this convergence is given by the following function.

Definition 4.18 (Fenchel-Legendre transform) The *Fenchel-Legendre transform* of Λ is

$$\Lambda^*(x) := \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda,x\rangle - \Lambda(\lambda)\}\,,\quad x\in\mathbb{R}^d\,. \tag{4.12}$$

Cramér's theorem characterises the logarithmic rate of the above convergence with rate function Λ^*. To ease notation and understanding we first study the case $d = 1$, and provide later arguments for the general case $d \geq 1$.

Theorem 4.19 (Cramér Theorem in $\mathbb{R}$) *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent, identically distributed $\mathbb{R}$-valued random variables with law $\mu \in \mathcal{M}_1(\mathbb{R})$ and denote μ_n the* *law of the empirical mean $\widehat{S}_n$ under $\mu^{\otimes n}$. Then $(\mu_n)_{n\in\mathbb{N}}$ satisfies the LDP on $\mathbb{R}$ with rate n and rate function Λ^*,i.e.,*

$$
\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) &\le -\inf_{x\in F} \Lambda^*(x)\,, \quad \textit{for all closed } F\subset\mathbb{R}\,,\\
\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) &\ge -\inf_{x\in G} \Lambda^*(x)\,, \quad \textit{for all open} G\subset\mathbb{R}\,.
\end{aligned}
\tag{4.13}
$$

The following lemma states the properties of Λ and Λ^* that are needed for proving Theorem 4.19.

Lemma 4.20 *Let $\mu\in\mathcal{M}_1(\mathbb{R})$ and $m := \int_{\mathbb{R}} x\,\mu(\mathrm{d}x)$.*

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

(b) If $\mathcal{D}(\Lambda)=\{0\}$, then $\Lambda^*\equiv 0$. If $\Lambda(\lambda)<\infty$ for some $\lambda>0$, then $m<\infty$ (possibly $m=-\infty$), and

$$
\Lambda^*(x) = \sup_{\lambda\ge 0}\{\lambda x - \Lambda(x)\}\,, \quad \textit{for all } x\ge m\,, \tag{4.14}
$$

is, for all $x>m$, a nondecreasing function. Likewise, if $\Lambda(\lambda)<\infty$ for some $\lambda<0$, then $m>-\infty$ (possibly $m=\infty$), and

$$
\Lambda^*(x) = \sup_{\lambda\le 0}\{\lambda x - \Lambda(x)\}\,, \quad \textit{for all } x\le m\,, \tag{4.15}
$$

is, for all $x<m$, a nondecreasing function. When $m\in\mathbb{R}$, then $\Lambda^*(m)=0$, and always,

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

(c) Λ is differentiable in $\operatorname{int}(\mathcal{D}(\Lambda))$ with

$$
\Lambda'(\eta) = \frac{1}{\mathbb{E}[\mathrm{e}^{\lambda X_1}]}\,\mathbb{E}[X_1\mathrm{e}^{\eta X_1}]\,, \tag{4.16}
$$

and

$$
\Lambda'(\eta) = y \Rightarrow \Lambda^*(y) = \eta y - \Lambda(\eta)\,.
$$

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

$$
\begin{aligned}
\Lambda(\alpha\lambda_1+(1-\alpha)\lambda_2) &= \log\mathbb{E}[(\mathrm{e}^{\lambda_1X_1})^{\alpha}(\mathrm{e}^{\lambda_2X_1})^{1-\alpha}] \le \log\Big(\mathbb{E}[\mathrm{e}^{\lambda_1X_1}]^{\alpha}\mathbb{E}[\mathrm{e}^{\lambda_2X_1}]^{1-\alpha}\Big)\\
&= \alpha\Lambda(\lambda_1)+(1-\alpha)\Lambda(\lambda_2)\,,
\end{aligned}
$$

implying convexity for Λ.

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

Furthermore, $\Lambda(0) = 0$, and so $\Lambda^*(x) \geq 0x - \Lambda(0) = 0$. Suppose that $x_n \to x$ as $n \to \infty$. Then, lower semicontinuity of Λ^* follows since

$$\liminf_{n\to\infty} \Lambda^*(x_n) \geq \liminf_{n\to\infty} (\lambda x_n - \Lambda(\lambda)) = \lambda x - \Lambda(\lambda)\,.$$

Hence, Λ^* 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,

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

and thus if $\Lambda(\lambda) < \infty$ we get that $m < \infty$. If $m = -\infty$, then $\Lambda(\lambda) = \infty$ for λ negative, and (4.14) trivially holds. In case $m \in R$, we obtain with the precious 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 (4.14) follows. The monotonicity of Λ^* on $[m, \infty)$ (nondecreasing) follows from (4.14), since for every $\lambda \geq 0$, the function $\lambda a - \Lambda(\lambda)$ 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 (4.14),

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

Hnece,

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

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 (4.16) follows by interchanging the order of differentiation and integration which we justify by the dominated convergence theorem as follows:

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

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

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

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 $g(\eta) = \sup_{\lambda\in\mathbb{R}} g(\lambda) = \Lambda^*(y)$.

□

Proof of Theorem 4.19. Proof of the upper bound in (4.13): Let $\varnothing \neq F \subset \mathbb{R}$ 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 4.20 it follows that m exists (possibly as extended real number). For all x and $\lambda \geq 0$, an application of the (exponential with function $\mathrm{e}^{\lambda x}$, $\lambda \geq 0$) Chebychev inequality yields

$$\mu_n([x,\infty)) = P(\widehat{S}_n \geq x) \leq \mathbb{E}[\mathrm{e}^{n(\widehat{S}_n - x)}] = \mathrm{e}^{-n\lambda x} \prod_{i=1}^m \mathbb{E}[\mathrm{e}^{\lambda X_i}] = \mathrm{e}^{-n(\lambda x - \Lambda(\lambda))}\,.$$

Now, if the mean $m < \infty$, then by (4.14) in Lemma 4.20, for every $x > m$, we obtain an upper by optimising over all $\lambda \in \mathbb{R}$, i.e.,

$$\mu_n([x,\infty)) \leq \mathrm{e}^{-n\Lambda^*(x)} \qquad \text{for every } x > m\,. \tag{4.17}$$

This follows from the proof of (4.14). Equivalently, if $m > -\infty$ and $x < m$, we can use an estimate via the exponential Chebychev inequality for $\lambda > 0$,

$$P(-\widehat{S}_n \geq -x) \leq \mathbb{E}[\exp(-n(\lambda(-\widehat{S}_n) - \widetilde{\Lambda}(\lambda)))]\,,$$

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

$$P(-\widehat{S}_n \geq -x) \leq \exp\big(-n \sup_{\lambda\leq 0}\{\lambda x - \Lambda(\lambda)\}\big) = \exp\big(-n\Lambda^*(x)\big)\,,$$

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 λ is not changing the supremum over $\lambda \leq 0$ as long as $x < m$. Therefore,

$$\mu_n((-\infty,x]) \leq \mathrm{e}^{-n\Lambda^*(x)}\,, \qquad \text{for every } x < m\,. \tag{4.18}$$

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 4.20, $\Lambda^*(m) = 0$, 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 (4.17) for $x = x_+$ and (4.18) for $x = x_-$ such that the union of events bounds ensures that

$$\mu_n(F) \leq \mu_((-\infty, x_-]) + \mu_n([x_=,\infty)) \leq 2\mathrm{e}^{-nI_F}\,,$$

and the upper bound in (4.13) follows as $n \to \infty$.

(ii) Suppose now $m = -\infty$. As Λ^* is nondecreasing, it follows from $\inf_{x\in\mathbb{R}} \Lambda^*(x) = 0$ that $\lim_{x\to-\infty} \Lambda^*(x) = 0$, and hence $x_* = \inf\{x \in \mathbb{R} \colon x \in F\}$ 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 (4.17) for $x = x_*$, we obtain the large deviations upper bound in (4.13). The third case (iii) $m = +\infty$ follows analogously to the second case.

Proof of the lower bound in (4.13): The key idea is to prove that for every $\delta > 0$ and every probability measure $\mu \in \mathcal{M}_1(\mathbb{R})$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n((-\delta,\delta)) \geq \inf_{\lambda\in\mathbb{R}} \{\Lambda(\lambda)\} = -\Lambda^*(0)\,, \tag{4.19}$$

where μ_n is the law of $\widehat{S}_n$ under $\mu^{\otimes n}$. The proof of (4.19) will keep us busy below, it is actually the major part of the work. Suppose now that (4.19) holds. We can then quickly see that the lower bound in (4.13) holds. First recall that we write Λ 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 Λ_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 (4.19) that for every $x \in \mathbb{R}$ and every $\delta > 0$,

$$\liminf_{n\to\infty} \mu_n((x-\delta, x+\delta)) \geq -\Lambda^*(x)\,. \tag{4.20}$$

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

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

and we can optimise the right hand site of (4.20) over all $x' \in G$ to obtain the large deviation lower bound in (4.13).

Proof of (4.19)**:** 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 $\operatorname{supp}(\mu) \subset \mathbb{R}$ is a bounded subset. These assumptions ensure that $\Lambda(\lambda) \to \infty$ when $|\lambda| \to \infty$ and that Λ is finite everywhere, i.e., $\mathcal{D}(\Lambda) = \mathbb{R}$. Then, according to part (c) of Lemma 4.20, Λ 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 $\widetilde{\mu} \in \mathcal{M}_1(\mathbb{R})$ by tilting the measure μ, that is, we define the Radon-Nikodym density to be

$$\frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\mu}(x) = \mathrm{e}^{\eta x - \Lambda(\eta)}\,, \tag{4.21}$$

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

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

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

We now denote $\widetilde{\mu}_n$ the law of $\widehat{S}_n$ under $\widetilde{\mu}^{\otimes n}$, and we observe that for every $\varepsilon > 0$ we obtain the estimate

$$\begin{aligned}\mu_n((-\varepsilon,\varepsilon)) &= \int_{\{x\in\mathbb{R}^n:\, |\sum_{i=1}^n x_i|<n\varepsilon\}} \mu(\mathrm{d}x_1)\cdots\mu(\mathrm{d}x_n) \\ &\geq \mathrm{e}^{-n\varepsilon|\eta|}\int_{\{x\in\mathbb{R}^n:\, |\sum_{i=1}^n x_i|<n\varepsilon\}} \exp\Big(\eta\sum_{i=1}^n x_i\Big)\,\mu(\mathrm{d}x_1)\cdots\mu(\mathrm{d}x_n) \\ &= \mathrm{e}^{-n\varepsilon|\eta|}\,\mathrm{e}^{n\Lambda(\lambda)}\,\widetilde{\mu}_n((-\varepsilon,\varepsilon))\,.\end{aligned}$$

By (4.16) and our choice of η,

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

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

$$\lim_{n\to\infty}\widetilde{\mu}((-\varepsilon,\varepsilon)) = 1\,. \tag{4.22}$$

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

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

and (4.19) follows by taking the limit $\varepsilon\to 0$ and using

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

2.) Suppose that $\mathrm{supp}(\mu)$ 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

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

Denote ν the law of X_1 conditioned on the event $\{|X_1|\leq M\}$, and let ν_n the law of $\widehat{S}_n$ conditioned on $\{|X_i|\leq M; i=1,\ldots,n\}$. 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 (4.19) holds for ν_n. The logarithmic moment generating function for ν is

$$\Lambda_\nu(\lambda) = \log\Big(\frac{\mathbb{E}[\mathrm{e}^{\lambda X_1}\mathbb{1}\{|X_1|\leq M\}]}{\mu([-m,M])}\Big) = \Lambda_M(\lambda)-\log\mu([-M,M])\,,$$

Thus

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

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

$$\liminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\delta,\delta)) \ge -I^*\,, \tag{4.23}$$

and we shall show that $\inf_{\lambda\in\mathbb{R}}\{\Lambda(\lambda)\} \le -I^*$ to conclude with (4.19). Note that Λ_M and thus $-I_M$ is denote decreasing in M, and

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

which shows that $-I^* \le 0$. 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 λ_0 in their intersection. By Lebesgue's monotone convergence theorem,

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

and thus our bound (4.23) yields (4.19).

3.) Suppose now that either $\mu((-\infty,0)) = 0$ or $\mu((0,\infty)) = 0$, then Λ is a monotone function with $\inf_{\lambda\in\mathbb{R}}\{\Lambda(\lambda)\} = \log\mu(\{0\})$. Hence, in this case, (4.19) follows from

$$\mu_n((-\delta,\delta)) \ge \mu_n(\{0\}) = \mu(\{0\})^n\,.$$

□

Remark 4.21 (a) The pivotal step in proving the large deviation upper bound is to optimise over exponential Chebychev inequalities for $\lambda \ge 0$ considering the function $e^{\lambda x}$. Then consideration of the mean m and the argument x of Λ^* 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.

◇

Exercise 4.22 Prove by an application of Fatou's lemma that Λ is lower semicontinuous.

Exercise 4.23 Compute Λ^* for the following distrbutions:

(a) $X \sim \mathsf{Poi}(\lambda)$, Poisson distribution with parameter $\lambda > 0$.

(b) $X \sim \mathsf{Ber}(p), p \in [0,1]$, Bernoulli distributed with success probability p.

(c) $X \sim \mathsf{Exp}(\lambda)$, exponentially distributed with parameter $\lambda > 0$.

(d) $X \sim \mathsf{N}(\mu,\sigma^2)$.

Exercise 4.24 Prove that Λ is $\mathcal{C}^\infty$ in the interior $\mathrm{int}(\mathcal{D}_\Lambda)$ and that Λ^* is strictly convex, and $\mathcal{C}^\infty$ in the interior of the set $\mathsf{F} := \{\Lambda'(\lambda) \colon \lambda \in \mathrm{int}(\mathcal{D}_\Lambda)\}$

☕☕☕

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 $\mu \in \mathcal{M}_1(\mathbb{R}^d)$.

Theorem 4.25 (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 μ_n the law of the empirical mean $\widehat{S}_n$ 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 Λ^*.*

Before we are discussing the proof of the $\mathbb{R}^d$ version of Cramér's Theorem, we show that actually Sanov's theorem, Theorem 4.13, can be deduced as a consequence of Cramér's theorem in $\mathbb{R}^d$. Note that the empirical mean of the random vectors

$X_i := \left(1\!\!1_{a_1}(Y_i), \dots, 1\!\!1_{a_{|E|}}(Y_i)\right), \quad i = 1, \dots, n, Y_i \in E$ independent, identically distributed

with law $\mu \in \mathcal{M}_1(E)$ equal L_n^Y, i.e.,

$$\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i = L_n^Y, \quad Y = (Y_1, \dots, Y_n).$$

Moreover, as the X_i are bounded, we have $\mathcal{D}(\Lambda) = \mathbb{R}^{|E|}$, and thus the following corollary of Cramér's theorem is obtained.

Corollary 4.26 *For any set $\Gamma \subset \mathcal{M}_1(\mathbb{R}^{|E|})$,*

$$\begin{aligned} -\inf_{\nu\in\mathrm{int}(\Gamma)}\{\Lambda^*(\nu)\} &\le \liminf_{n\to\infty}\frac{1}{n}\log P(L_n^Y \in \Gamma) \\ &\le \limsup_{n\to\infty}\frac{1}{n}\log P(L_n^Y \in \Gamma) \le -\inf_{\nu\in\overline{\Gamma}}\{\Lambda^*(\nu)\}, \end{aligned}$$

where Λ^* is the Legendre-Fenchel transform of the logarithmic moment generating function

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

with $\lambda = (\lambda_1, \dots, \lambda_{|E|}) \in \mathbb{R}^{|E|}$.

Proof. The large deviation bounds follow from the ones in Theorem 4.25 and from the fact that $P(L_N^Y \in \Gamma) = P(\widehat{S}_N \in \Gamma)$. □

Exercise 4.27 Show that in the setting of Corollary 4.26,

$$\Lambda^*(x) = \mathsf{H}(x|\mu).$$

☕☕

Solution. As

$$\Lambda^*(\nu) = \sup_{\lambda \in \mathbb{R}^{|E|}} \{\langle \lambda, \nu \rangle - \Lambda(\lambda)\},$$

we obtain maximiser for

$$\nu_x = \frac{e^{\lambda_x}\mu_x}{\sum_{y\in E} e^{\lambda_y}\mu_y},$$

and thus we see that this holds only if $\sum_{x\in E}\nu_x = 1$, that is $\mathcal{D}(\Lambda^*) = \mathsf{Sim}_E = \mathcal{M}_1(E)$. Now Jensen's inequality implies that

$$\Lambda(\lambda) \geq \sum_{x\in E} \nu(x) \log \frac{\mu(x)}{\nu(x)} e^{\lambda_x} = \langle \lambda, \nu \rangle - \mathsf{H}(\nu|\mu),$$

and thus

$$\mathsf{H}(\nu|\mu) \leq \Lambda^*(\nu).$$

By choosing $\lambda_x = \log(\nu(x)/\mu(x))$ for any $x \in E$ with $\nu(x) > 0$ and $\lambda_x \to \infty$ for any $x \in E$ with $\nu(x) = 0$, we obtain equality. We also obtain that

$$\mathsf{H}(\nu|\mu) = \Lambda^*(\nu) = +\infty$$

whenever there are $x \in E$ with $\nu(x) > 0$ but $\mu(x) = 0$. ☺

We now want to compare Cramér's Theorem for finite sets E with Sanov's Theorem, Theorem 4.13 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 $\widehat{S}_n := \frac{1}{n}\sum_{i=1}^n X_i$, where $X_i = f(Y_i)$ for some function $f\colon E \to \mathbb{R}$.Without loss of generality, we assume further that $E_\mu = E$ and that $f(a_1) < f(a_2) < \cdots < f(a_{|E|})$. Then $\widehat{S}_n \in [f(a_1), f(a_{|E|})] =: K$, and writing $Y = (Y_1, \ldots, Y_n)$ and $F := (f(a_1), \ldots, f(a_{|E|})) \in \mathbb{R}^{|E|}$, we see that

$$\widehat{S}_n = \sum_{i=1}^{|E|} f(a_i) L_n^Y(a_i) =: \langle f, L_n^Y \rangle,$$

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 $n \in \mathbb{N}$,

$$\widehat{S}_n \in A \iff L_n^Y \in \{\nu \in \mathcal{M}_1(E)\colon \langle f, \nu \rangle \in A\} =: \Gamma. \tag{4.24}$$

Theorem 4.28 (Cramér's theorem for subsets of $\mathbb{R}$) *For any $A \subset \mathbb{R}$,*

$$\begin{aligned} -\inf_{x\in \operatorname{int}(A)}\{I(x)\} &\leq \liminf_{n\to\infty} \frac{1}{n} \log P_\mu(\widehat{S}_n \in A) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log P_\mu(\widehat{S}_n \in A) \leq -\inf_{x\in A}\{I(x)\}, \end{aligned}$$

where

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

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

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}\,, \tag{4.25}$$

where

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

Proof. Suppose that $f\colon E \to \mathbb{R}$ is constant, i.e., $f(x) = c \in \mathbb{R}$ for all $x \in E$. Then $X_i = c, \widehat{S}_n = c$, and hence $\Gamma = \mathcal{M}_1(E)$ in (4.24). Note that when $x \neq c$ there is no $\nu \in \mathcal{M}_1(E)$ with $\langle f, \nu \rangle = c$, and thus the infimum in the definition of I is over an empty set and therefore infinity. Hence,

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

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

$$\lim_{n\infty} \frac{1}{n} \log \mathbb{E}[\mathrm{e}^{n\lambda \widehat{S}_n}] = \Lambda(\lambda) = \log \mathrm{e}^{\lambda c} = \lambda c\,,$$

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 (4.24). Then the lower and upper bounds follow from Sanov's theorem, Theorem 4.13. Furthermore, due to (4.24),

$$\inf_{\nu \in \mathrm{int}(\Gamma)} \{\mathsf{H}(\nu|\mu)\} = \inf_{x \in \mathrm{int}(A)} \Big\{ \inf_{\nu\colon \langle f,\nu\rangle = x} \{\mathsf{H}(\nu|\mu)\} \Big\}\,.$$

Jensen's inequality yields

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

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

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

Thus

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

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

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

and therefore (4.25) holds for all $x \in \{\Lambda'(\lambda)\colon \lambda \in \mathbb{R}\}$. An easy computation shows that

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

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

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

Hence, (4.25) holds for all $x \in \mathrm{int}(K)$. 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

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

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

□

We now turn to showing how one can prove Cramér's theorem in $\mathbb{R}^d$.

Proof of Cramér's Theorem in $\mathbb{R}^d$, Theorem 4.25. A detailed proof can be found in [DZ98]. It combines Lemma 4.29 with elements of the proof of Theorem 4.43 below. It is actually a special case of Theorem 4.43 as the space $\mathbb{R}^d$ is a self dual vector space. □

The following lemma summarises the properties of Λ and Λ^* needed to prove Theorem 4.25. This is almost like Lemma 4.20 but without the monotonicity statement.

Lemma 4.29 *(a) Λ is convex and differentiable everywhere, and Λ^* is a good convex rate function.*

(b)

$$y = \nabla\Lambda(\eta) \implies \Lambda^*(y) = \langle \eta, y\rangle - \Lambda(\eta)\,.$$

Proof. Exercise. □

We will later prove a more sophisticated version of the theorem, here let us just mention that we shall obtain the upper bound first for compact sets which can be suitably covered by balls. A large deviation principle is called weak when the upper bound holds only for compact sets. hence we need to know how to left the upper bound for compact sets to general closed sets. Recall the Definition 4.6. To strengthen the weal LDP to a full LDP requires a way of showing that most of the probability mass (at least on an exponential scale) is concentrated on compact sets. Here, we assume that E is a topological Hausdorff space.

Definition 4.30 (Exponential tightness) Suppose that all compact subsets of E belong to the σ-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

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(K_\alpha^{\mathrm{c}}) < -\alpha\,.$$

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

Proposition 4.31 (Exponential tightness) *Let $(\mu_n)_{n\in\mathbb{N}}$ be exponentially tight.*

(a) If the upper bound (4.3) *holds for some $\alpha < \infty$ and all compact subsets of the complement $\mathcal{L}_I(\alpha)^{\rm c}$, then it holds for all measurable sets M with $\overline{M} \subset \mathcal{L}_I(\alpha)^{\rm c}$. If $\mathcal{B}(E) \subset \mathcal{B}$ and the upper bound* (4.3) *holds for all compact sets, then it also holds for all closed sets.*

(b) If the lower bound (4.4) *holds (the lower bound in* (4.2) *when $\mathcal{B}(E) \subset \mathcal{B}$) holds 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 $\overline{M} \subset \mathcal{L}_I(\alpha)^{\rm c}$, and let K_α be the compact set in the definition for exponential tightness. Then $\overline{M} \cap K_\alpha \in \mathcal{B}$ and $K_\alpha^{\rm c} \in \mathcal{B}$.

$$\mu_n(M) \le \mu_n(\overline{M} \cap K_\alpha) + \mu_n(K_\alpha^{\rm c}) \,. \tag{4.26}$$

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

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

Thus

$$\limsup_{n\to\infty} \frac{1}{n} \log \text{R.H.S. of (4.26)} = \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(\overline{M} \cap K_\alpha) \wedge \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(K_\alpha^{\rm c}) \,,$$

and therefore

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(M) \le -\alpha \,.$$

(b) We apply the lower bound (4.4) to the open set $K_\alpha^{\rm c}$, and obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(K_\alpha^{\rm c}) \ge - \inf_{x \in K_\alpha^{\rm c}} \{I(x)\} \,,$$

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

If a set E is given the coarse topology $\{\varnothing, E\}$, the only information implied by the LDP is that $\inf_{x\in E} I(x) = 0$, and our rate functions satisfy this requirement. We must therefore put some constraint on the topology of the set E. Recall that a topological space E is Hausdorff if, for every pair of distinct points x and y, there are exist disjoint neighbourhoods of x and y. We often need a further requirement, called regular.

Definition 4.32 (a) A function $f\colon E \to \mathbb{R}$, E Hausdorff space, is called *lower semicontinuous (l.s.c.) (upper semicontinuous (u.s.c.))* if its level sets $\mathcal{L}_f(\alpha) = \{x \in E\colon f(x) \le \alpha\}$ are closed (respectively $\{x \in E\colon f(x) \ge \alpha\}$ are closed).

(b) A topological Hausdorff space E is called *regular* if, for any closed set $F \subset E$ and any point $x \notin F$, there exist disjoint open sets G_1 and G_1 such that $F \subset G_1$ and $x \in G_2$.

(c) A topological Hausdorff space E is called *completely regular topological space* if E is a Hausdorff space such that for any closed set $F \subset E$ and any point $x \notin F$, there exists a continuous function $f\colon E \to \mathbb{R}$ such that $f(x) = 1$ and $f(y) = 0$ for all $y \in F$. Such a space

Remark 4.33 (Regular spaces) (a) Note that $f\colon E \to \mathbb{R}$ is continuous if and only if f is lower semicontinuous and upper semicontinuous. The indicator/characteristic function $1\!\!1_A$ is lower semicontinuous for every open set A, and $1\!\!1_F$ is upper semicontinuous for any closed F.

(b) For any neighbourhood $G \ni x$, $x \in E$, there exists a neighbourhood $A \ni x$ such that $\overline{A} \subset G$.

(c) Every metric space is regular. If a topological vector space is Hausdorff, then it is regular.

(d) A lower semicontinuous function f satisfies, at every point x,

$$f(x) = \sup_{\substack{G \ni x \\ \text{neighbourhood}}} \inf_{y \in G} \{f(y)\}\,, \quad x \in E\,. \tag{4.27}$$

(e) Because of (4.27), for any $x \in E$ and any $\delta > 0$, there is a neighbourhood $G = G(x, \delta) \ni x$, such that

$$\inf_{y \in G} \{f(y)\} \ge (f(x) - \delta) \wedge 1/\delta\,.$$

Let $A = A(x, \delta)$ be a neighbourhood of x such that $\overline{A} \subset G$. Then

$$\inf_{y \in \overline{A}} \{f(y)\} \ge \inf_{y \in G} \{f(y)\} \ge (f(x) - \delta) \wedge 1/\delta\,. \tag{4.28}$$

$\diamond$

Proposition 4.34 (Uniqueness of the rate function) *A sequence $(\mu_n)_{n \in \mathbb{N}}$ of probability measures $\mu_n \in \mathcal{M}_1(E)$ on a regular space E can have at most one rate function associated with its LDP.*

Proof. Suppose there are two rate functions I and J for the LDP associated with $(\mu_n)_{n\in\mathbb{N}}$. Without loss of generality, as $I \not\equiv J$, assume that for $x_0 \in E$, $I(x_0) > J(x_0)$. Fix $\delta > 0$ and consider the open set $A \ni x_0$ with

$$\inf_{y\in\overline{A}}\{I(y)\} \geq (I(x_0)-\delta)\wedge 1/\delta\,.$$

Such as an open set exists due to (4.28). By the LDP it follows that

$$-\inf_{y\in\overline{A}}\{I(y)\} \geq \limsup_{n\to\infty}\frac{1}{n}\log\mu_n(A) \geq \liminf_{n\to\infty}\frac{1}{n}\log\mu_n(A) \geq -\inf_{y\in A}\{J(y)\}\,.$$

Therefore,

$$J(x_0) \geq \inf_{y\in A}\{J(y)\} \geq \inf_{y\in\overline{A}}\{I(x)\} \geq (I(x_0)-\delta)\wedge 1/\delta\,.$$

Since $\delta > 0$ is arbitrary, this contradicts the assumption that $I(x_0) > J(x_0)$. □

Theorem 4.35 (Contraction principle) *Let E and Y be Hausdorff spaces and $f\colon E \to Y$ be continuous. Suppose $I\colon E \to [0,\infty)$ is a good rate function.*

(a) For each $y \in Y$, define

$$J(y) := \inf_{x\in E\colon f(x)=y}\{I(x)\}\,. \tag{4.29}$$

Then J is a good rate function on Y, where as usual the infimum over the empty set is taken as ∞.

(b) If $(\mu_n)_{n\in\mathbb{N}}$, $\mu_n \in \mathcal{M}_1(E)$, satisfies the LDP on E with rate n and rate function I, then $(\nu_n)_{n\in\mathbb{N}}$ with $\nu_n := \mu_n \circ f^{-1} \in \mathcal{M}_1(Y)$ satisfies the LDP on Y with rate n and rate function J.

Proof. (a) J is nonnegative by definition. Since I is a good rate function, for all $y \in f(E)$ the infimum in the definition of J is obtained at some point of E (lower semicontinuous functions attain their minimum on compact sets). Thus, we obtained for the level set $\mathcal{L}_J(\alpha)$,

$$\mathcal{L}_J(\alpha) \subset \{f(x)\colon I(x) \leq \alpha\} = f(\mathcal{L}_I(\alpha))\,.$$

As the level sets $\mathcal{L}_I(\alpha)$ are compact, so are the level sets $\mathcal{L}_J(\alpha) \subset Y$.

(b) For every $A \subset Y$,

$$\inf_{y\in A}\{J(y)\} = \inf_{y\in A}\inf_{x\in E\colon f(x)=y}\{I(x)\} = \inf_{x\in f^{-1}(A)}\{I(x)\}\,. \tag{4.30}$$

Since f is continuous, the set $f^{-1}(A)$ is an open (closed) subset of E for any open (closed) $A \subset E$. Therefore, the LDP for $(\nu_n)_{n\in\mathbb{N}}$ follows as a consequence of the LDP for $(\mu_n)_{n\in\mathbb{N}}$ and (4.30). □

We assume that E is a regular topological Hausdorff space.

Theorem 4.36 (Varadhan's Lemma) *Suppose that* $(\mu_n)_{n\in\mathbb{N}}$ *satisfies the LDP with a good rate function* $I\colon E\to[0,\infty]$*, and let* $H\colon E\to\mathbb{R}$ *be a continuous function. Assume that either the tail-condition*

$$\lim_{M\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}_{\mu_n}[\mathrm{e}^{nH}1\!\!1\{H\ge M\}]=-\infty\,, \tag{4.31}$$

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

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{\gamma nH}]<\infty\,, \tag{4.32}$$

hold. Then

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_{\mu_n}[\mathrm{e}^{nH}]=\sup_{x\in E}\{H(x)-I(x)\}\,.$$

Remark 4.37 (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 (4.31). The moment condition (4.32) implies the tail condition (4.31) as we see using Hölder's inequality,

$$\begin{aligned}\int_{\{H\ge M\}}\mathrm{e}^{nH(x)}\,\mu_n(\mathrm{d}x)&\le\Big(\int\mathrm{e}^{\gamma nH(x)}\,\mu_n(\mathrm{d}x)\Big)^{1/\gamma}(\mu_n(H\ge M))^{1-\frac{1}{\gamma}}\\&\le\Big(\int\mathrm{e}^{\gamma nH(x)}\,\mu_n(\mathrm{d}x)\Big)^{1/\gamma}\Big(\mathrm{e}^{-\gamma Mn}\int\mathrm{e}^{\gamma nH(x)}\,\mu_n(\mathrm{d}x)\Big)^{1-\frac{1}{\gamma}}\\&=\exp\big((1-\gamma)Mn\big)\Big(\int\mathrm{e}^{\gamma nH(x)}\,\mu_n(dx)\Big)\,.\end{aligned}$$

◇

Proof of Theorem 4.36. The proof is an immediate consequences of the following two lemmas and Remark 4.37. □

Lemma 4.38 *If* $H\colon E\to\mathbb{R}$ *is lower semicontinuous and the large deviation lower bound holds with* $I\colon E\to[0,\infty]$*, then*

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}]\ge\sup_{x\in E}\{H(x)-I(x)\}\,.$$

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

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}]&\ge\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}1\!\!1_G]\ge\inf_{y\in G}\{H(y)\}+\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(G)\\&\ge\inf_{y\in G}\{H(y)\}-\inf_{y\in G}\{I(y)\}\ge H(x)-I(x)-\delta\,.\end{aligned}$$

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

Lemma 4.39 *If $H\colon E \to \mathbb{R}$ is an upper semicontinuous for which the tail condition* (4.31) *holds, and if the large deviation upper bound holds with the good rate function $I\colon E \to [0,\infty]$, then*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}[\mathrm{e}^{nH}] \le \sup_{x\in E}\{H(x) - I(x)\}\,.$$

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

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

Clearly, this function satisfies the tail condition (4.31). 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

$$\inf_{y\in\overline{A_x}}\{I(y)\} \ge I(x) - \delta\,, \qquad \sup_{y\in\overline{A_x}}\{H(y)\} \le H(x) + \delta\,,$$

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 $\mathcal{L}_I(\alpha) \subset \bigcup_{i+1}^N A_{x_i}$. Therefore,

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{nH}] &\le \sum_{i=1}^N \mathbb{E}[\mathrm{e}^{nH}1\!\!1_{A_{x_i}}] + \mathrm{e}^{nM}\mu_n\big(\big(\bigcup_{i=1}^N A_{x_i}\big)^{\mathrm{c}}\big) \\
&\le \sum_{i=1}^N \mathrm{e}^{n(H(x_i)+\delta)}\mu_n(\overline{A_{x_i}}) + \mathrm{e}^{nM}\mu_n\big(\big(\bigcup_{i=1}^N A_{x_i}\big)^{\mathrm{c}}\big)\,.
\end{aligned}$$

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

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}] \\
&\qquad \le \max\Big\{\max_{1\le i\le N}\{H(x_i) + \delta - \inf_{y\in\overline{A_{x_i}}}\{I(y)\}\}, M - \inf_{y\in\big(\bigcup_{i=1}^N A_{x_i}\big)^{\mathrm{c}}}\{I(y)\}\Big\} \\
&\qquad \le \max\Big\{\max_{1\le i\le N}\{H(x_i) - I(x_i) + 2\delta\}, M - \alpha\Big\} \\
&\qquad \le \max\Big\{\sup_{x\in E}\{H(x) - I(x)\}, M - \alpha\Big\} + 2\delta\,.
\end{aligned}$$

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

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}] \\
&\qquad \le \sup_{x\in E}\{H(x) - I(x)\} \vee \limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}[\mathrm{e}^{nH}1\!\!1\{H \ge M\}]\,.
\end{aligned}$$

Now the tail condition (4.31) completes the proof by taking the limit $M \to \infty$. $\square$

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

Theorem 4.40 (Tilted LDP) *Let (E, d) be a Polish space. Suppose that $(\mu_n)_{n\in\mathbb{N}}$ satisfies the LDP with a good rate function $I\colon E \to [0,\infty]$, and let $H\colon E \to \mathbb{R}$ be a continuous function that is bounded from above. Then define*

$$Z_n(H) := \int_E \mathrm{e}^{nH(x)}\, \mu_n(\mathrm{d}x)\,,$$

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

$$\frac{\mathrm{d}\mu_n^H}{\mathrm{d}\mu_n}(x) = \frac{\mathrm{e}^{nH(x)}}{Z_n(H)}\,, \quad x \in E\,.$$

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

$$I^H(x) = I(x) - H(x) + \sup_{y\in E}\{H(y) - I(y)\}\,, \quad x \in E\,. \tag{4.33}$$

Proof. From Theorem 4.36 we know that

$$\lim_{n\to\infty} \frac{1}{n} \log Z_n(H) = \sup_{y\in E}\{H(y) - I(y)\}\,.$$

Then we obtain the large deviation bounds by simply repeating the above arguments in the proof of Theorem 4.36, For example, let $K \subset E$ be closed, then

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mu_n^H(K) &= \limsup_{n\to\infty} \frac{1}{n} \log \int_K \mathrm{e}^{nH(x)}\, \mu_n(\mathrm{d}x) - \limsup_{n\to\infty} \frac{1}{n} \log Z_n(H) \\ &\le \sup_{y\in K}\{H(x) - I(x)\} - \sup_{y\in E}\{H(y) - I(y)\} = -\inf_{y\in K}\{I^H(y)\}\,,\end{aligned}$$

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 similalry. $\square$

From our study of Cramér's theorem in both $\mathbb{R}$ and $\mathbb{R}^d$, Theorem 4.19 and Theorem 4.25, we have seen that when the space E is a vector space then the logarithmic moment generating function plays a vital role, in summary, exponential moments of linear function suffice. As we just learned from Varadhan's lemma, Theorem 4.36, the function H can nonlinear. we now briefly study the possibility to invert Varadhan's Lemma for nonlinear functions.

Definition 4.41 Let E be a completely regular topological space. For each Borel-measurable function $f\colon E \to \mathbb{R}$ and sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures $\mu_n \in \mathcal{M}_1(E)$, define

$$\Lambda_f := \lim_{n\to\infty} \frac{1}{n} \log\Big(\int_E \mathrm{e}^{nf(x)}\, \mu_n(\mathrm{d}x)\Big)\,, \tag{4.34}$$

provided the limit exists.

Remark 4.42 When E is a vector space, then the functionals Λ_f for continuous linear functions $f \in E^*$ (elements of the dual E^*, e.g., in $\mathbb{R}^d$, $f(x) = \langle \lambda, x\rangle, \lambda \in \mathbb{R}^d$) are just the values of the logarithmic moment generating function. $\diamond$

Theorem 4.43 (Bryc) *Let E be a completely regular topological space. Suppose that the sequence $(\mu_n)_{n\in\mathbb{N}}$, $\mu_n \in \mathcal{M}_1(E)$, is exponentially tight and that the limit Λ_f in (4.34) exists for every $f \in \mathcal{C}_\mathrm{b}(E)$. Then $(\mu_n)_{n\in\mathbb{N}}$ satisfies the LDP with good rate function*

$$I(x) = \sup_{f\in\mathcal{C}_\mathrm{b}(E)} \{f(x) - \Lambda_f\}\,. \tag{4.35}$$

Furthermore, for every $f \in \mathcal{C}_\mathrm{b}(E)$,

$$\Lambda_f = \sup_{x\in E}\{f(x) - I(x)\}\,. \tag{4.36}$$

Proof of Bryc's Theorem. If $f \equiv 0$ then $\Lambda_0 = 0$ and $I \geq 0$. The function I as a supremum of continuous functions is lower semicontinuous, and have that I is a rate function. It suffices therefore to show the weal LDP as the sequence $(\mu_n)_{n\in\mathbb{N}}$ is exponentially tight. By Varadhan's Lemma 4.36, we see that

$$\Lambda_f = \sup_{x\in E}\{f(x) - I(x)\}\,.$$

The statement follows by showing the lower and upper bound in Lemma 4.44 and Lemma 4.45. $\square$

Lemma 4.44 (Lower bound) *If Λ_f exists for each $f \in \mathcal{C}_\mathrm{b}(E)$, then, for every open $G \subset E$ and $x \in G$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq -I(x)\,.$$

Proof. Pick $x \in E$ and a neighbourhood $G \ni x$. Since E is completely regular, there exists a continuous function $f\colon E \to [0,1]$ such that $f(x) = 1$ and $f(y) = 0$ for all $y \in G^\mathrm{c}$. Define $f_m := m(f-1), m \in \mathbb{N}$. Then $f_m \in \mathcal{C}_\mathrm{b}(E)$. Thus

$$\int_E \mathrm{e}^{nf_m(x)}\, \mu_n(\mathrm{d}x) \leq \mathrm{e}^{-nm}\mu_n(G^\mathrm{c}) + \mu_n(G) \leq \mathrm{e}^{-nm} + \mu_n(G)$$

and

$$\begin{aligned}\max\{\liminf_{n\to\infty} \frac{1}{n}\log\mu_n(G); -m\} \geq \liminf_{n\to\infty}\frac{1}{n}\log\int_E \mathrm{e}^{nf_m(x)}\,\mu_n(\mathrm{d}x) = \Lambda_{f_m} = \\ - = (f_m(x) - \Lambda_{f_m} \geq - \sup_{f\in\mathcal{C}_\mathrm{b}(E)}\{f(x) - \Lambda_f\} = -I(x)\,.\end{aligned}$$

$\square$

Lemma 4.45 (Upper bound) *f* Λ_f *exists for each* $f \in \mathcal{C}_{\rm b}(E)$*, then, for every compact set* $K \subset E$*,*

$$\limsup_{n\to\infty} \frac{1}{n}\log\mu_n(K) \le -\inf_{x\in K}\{I(x)\}\,.$$

Proof. Fix a compact set $K \subset E$ and $\delta > 0$, and define $I^\delta = \min\{I(x)-\delta; 1/\delta\}$. Then, for any $x \in K$, there exists $g = g^x \in \mathcal{C}_{\rm b}(E)$ such that

$$g(x) - \Lambda_g \ge I^\delta(x)\,. \tag{4.37}$$

There exists furthermore a neighbourhood $A_x \ni x$ such that

$$\inf_{y\in A_x}\{g(y)-g(x)\} \ge -\delta\,.$$

We use Chebyshev's inequality for the function $\psi(y) = \exp(ng(y) - ng(x)$ with

$$\inf_{y\in A_x}\{\psi(y)\} = \exp(\inf_{y\in A_x}\{ng(y)-ng(x)\})$$

to arrive at

$$\mu_n(A_x) \le E_{\mu_n}\Big[{\rm e}^{ng-ng(x)}\Big]\exp(-\inf_{y\in A_x}\{ng(y)-ng(x)\})\,,$$

yielding

$$\frac{1}{n}\log\mu_n(A_x) \le \delta - \Big(g(x) - \frac{1}{n}\log\int_E {\rm e}^{ng(y)}\,\mu_n({\rm d}y)\Big)\,.$$

We can now extract a finite cover $\bigcup_{i=1}^N A_{x_i} \supset K$ from the open cover of the compact set K, and by the union of events bound,

$$\frac{1}{n}\log\mu_n(K) \le \frac{1}{n}\log N + \delta - \min_{1\le i\le N}\Big\{g^i(x_i) - \frac{1}{n}\log\int_E {\rm e}^{ng^i(y)}\,\mu_n({\rm d}y)\Big\}\,,$$

where g^i is the function g as above for $x_i \in K$. Thus,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(K) \le \delta - \min_{1\le i\le N}\{g^9(x_i) - \Lambda_{g^i}\} \le \delta - \min_{1\le i\le N}\{I^\delta(x_i)\}\,,$$

and therefore

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(K) \le \delta - \inf_{x\in K}\{I^\delta(x)\}\,.$$

We conclude by noting that

$$\lim_{\delta\to0}\inf_{x\in K}\{I^\delta(x)\} = \inf_{x\in K}\{I(x)\}\,,$$

and taking the limit $\delta \to 0$. □

We finish our basic introduction to the theory of large deviations with considering Hausdorff topological vector space E, and recall that such spaces are regular. 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 ha slaw $\mu_n \in \mathcal{M}_1(E)$. We define the *logarithmic moment generating function for* μ_n as

$$\Lambda_{\mu_n}(\lambda) := \log \mathbb{E}[\mathrm{e}^{\langle \lambda, X_n \rangle}] = \log \int_E \mathrm{e}^{\lambda(x)} \, \mu_n(\mathrm{d}x) \,, \quad \lambda \in E^* \,, \tag{4.38}$$

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

$$\overline{\Lambda}(\lambda) := \limsup_{n\to\infty} \frac{1}{n} \log \Lambda_{\mu_n}(n\lambda) \,, \tag{4.39}$$

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

$$f^*(x) := \sup_{\lambda\in E^*} \{\langle \lambda, x \rangle - f(\lambda)\} \,, \quad x \in E \,. \tag{4.40}$$

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

Theorem 4.46 (A General Upper bound) *Let* $(\mu_n)_{n\in\mathbb{N}}$ *be a sequence of probability measures. Then the following holds.*

(a) $\overline{\Lambda}$ *of* (4.39) *is convex on* E^* *and* $\overline{\Lambda}^*$ *is a convex rate function.*

(b) For any compact set $K \subset E$*,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(K) \le -\inf_{x\in K} \{\overline{\Lambda}^*(x)\} \,. \tag{4.41}$$

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

$$\overline{\Lambda}(\cdot) := \limsup_{n\to\infty} \frac{1}{n} \log \Lambda_{\mu_n}(n\cdot)$$

is also convex function. As $\Lambda_{\mu_n}(0) = 0$ for all $n \in \mathbb{N}$, we have that $\overline{\Lambda}(0) = 0$ and thus $\overline{\Lambda}^* \ge 0$. Note that $g(\lambda) := \langle \lambda, x \rangle - \overline{\Lambda}(\lambda)$ is continuous for every $\lambda \in E^*$. Then the lower semicontinuity of $\overline{\Lambda}^*$ follows from the fact that the supremum over continuous functions is lower semicontinuous. The convexity is shown as in Lemma 4.20.
(b) The upper bound follows exactly the steps in the proof of the upper bound in Lemma 4.45 to prove Theorem 4.43. 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.
□

We conclude our introduction to basic large deviation theory by giving a few results concerning the case when the random variables involved are not necessarily independent our identically distributed. What we just learned is that the crucial steps for LDPs will be the proof of the lower bound. In Theorem 4.19, Theorem 4.13, and Theorem 4.25, the independence allows to use law of large numbers to tilt our measures towards a measure which turns a rare events into an event with probability almost one. We first give an abstract version of the corresponding theorems and will then sketch how the proof of the lower bound is performed in the case $E = \mathbb{R}^d$, a variant of Theorem 4.25, called the Gärtner-Ellis theorem.

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

$$\langle \lambda, x\rangle - \overline{\Lambda}^*(x) > \langle \lambda, z\rangle - \overline{\Lambda}^*(z)\,, \qquad \text{for all } z \neq x\,. \tag{4.42}$$

Theorem 4.47 (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 $F \subset E$,

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

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

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_{\mu_n}(n\lambda) \;\; \textit{exists and} \;\; \overline{\Lambda}(\gamma\lambda) < \infty \;\; \textit{for some} \;\; \gamma > 1\,.$$

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

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

(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 that a function $f\colon E^* \to \mathbb{R}$ 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 4.48 *Let $(\mu_n)_{n\in\mathbb{N}}$ be an exponentially tight sequence of probability measures on a Banach space E. Suppose that the function $\Lambda(\cdot) = \lim_{n\to\infty} \frac{1}{n} \log \Lambda_{\mu_n}(n\cdot)$ 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 Λ^*.*

Proof. The crucial point is to show that (c) in Theorem 4.47 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].

□

To demonstrate the role of the exposed points we show the lower bound for the Gärtner-Ellis Theorem in $\mathbb{R}^d$.

Theorem 4.49 (Gärtner-Ellis Theorem in $\mathbb{R}^d$) *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:*

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{n} \log \Lambda_{\mu_n}(n\lambda) \quad \text{exists as an extended real number for all } \lambda \in \mathbb{R}^d\,, \tag{4.43}$$

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

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

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

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

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

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

Remark 4.50 A convex function $\Lambda\colon \mathbb{R}^d \to (-\infty,\infty]$ is *essentially smooth* if

(a) $\operatorname{int}(\mathcal{D}(\Lambda)) \neq \varnothing$.

(b) Λ is differentiable in $\operatorname{int}(\mathcal{D}(\Lambda))$.

(c) Λ is steep, that is, $\lim_{n\to\infty}|\nabla\Lambda(\lambda_n)| = \infty$ whenever $(\lambda_n)_{n\in\mathbb{N}}$ sequence in $\operatorname{int}(\mathcal{D}(\Lambda))$ 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 Λ is essentially smooth and the LDP holds. ◇

Proof. The upper bound is proved similar to the upper bound in Theorem 4.43, for details see Chapter 2.3 in [DZ98], or better Chapter V in [dH00].
Lower bound (b): We need to show that for $y \in \mathcal{E}$,

$$\lim_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\delta(y)) \geq -\Lambda^*(y)\,. \tag{4.44}$$

Fix $y \in \mathcal{E}$ and let $\eta \in \text{int}(\mathcal{D}(\Lambda))$ 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 $\widetilde{\mu}_n$ via the density,

$$\frac{\mathrm{d}\widetilde{\mu}_n}{\mathrm{d}\mu_n}(z) = \exp\left(n\langle \eta, z\rangle - \Lambda_{\mu_n}(n\eta)\right). \tag{4.45}$$

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

$$\begin{aligned}\frac{1}{n}\log\mu_n(B_\delta(y)) &= \frac{1}{n}\Lambda_{\mu_n}(n\eta) - \langle\eta, y\rangle + \frac{1}{n}\int_{B_\delta(y)} \mathrm{e}^{n\langle \eta, y-z\rangle}\,\widetilde{\mu}_n(\mathrm{d}z)\\ &\geq \frac{1}{n}\Lambda_{\mu_n}(n\eta) - \langle \eta, y\rangle - |\eta|\delta + \frac{1}{n}\log\widetilde{\mu}_n(B_\delta(y)).\end{aligned}$$

Therefore,

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mu_n(B_\delta(y)) &\geq \Lambda(\eta) - \langle\eta, y\rangle + \liminf_{n\to\infty}\frac{1}{n}\log\widetilde{\mu}_n(B_\delta(y))\\ &\geq -\Lambda^*(y) + \liminf_{n\to\infty}\frac{1}{n}\log\widetilde{\mu}_n(B_\delta(y))\end{aligned}$$

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 $\widetilde{\mu}_n$. One can easily show that

$$\frac{1}{n}\widetilde{\Lambda}_{\widetilde{\mu}_n}(n\lambda) \underset{n\to\infty}{\longrightarrow} \widetilde{\Lambda}(\lambda) = \Lambda(\lambda+\eta) - \Lambda(\eta),$$

where the limiting moment generating function $\widetilde{\Lambda}$ satisfies assumption (4.43) as clearly $\widetilde{\Lambda}(0) = 0$ and $\widetilde{\lambda} < \infty$ for $|\lambda|$ small enough. Define

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

Since $(\widetilde{\mu}_n)_{n\in\mathbb{N}}$ satisfies the assumptions (4.43), we can apply Lemma 4.29 and part (a) above to show that $(\widetilde{\mu})_{n\in\mathbb{N}}$ satisfies a large deviation upper bound with the good rate function $\widetilde{\Lambda}^*$. Thus, for the closed set $B_\delta(y)^{\mathrm{c}}$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\widetilde{\mu}_n(B_\delta(y)^{\mathrm{c}}) \leq -\inf_{x\in B_\delta(y)^{\mathrm{c}}}\{\widetilde{\Lambda}^*(x)\} = \widetilde{\Lambda}^*(x_0)$$

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 $\widetilde{\Lambda}^*(x_0) > 0$. At this point we use the property that y is an exposed point for Λ^* with exposing hyperplane η. First,

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

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

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

Thus, for every $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \widetilde{\mu}_n(B_\delta(y)^{\mathrm{c}}) < 0\,.$$

This implies that $\widetilde{\mu}_n(B_\delta(y)^{\mathrm{c}}) \to 0$ as $n \to \infty$ and hence $\widetilde{\mu}_n(B_\delta(y)) \to 1$ as $n \to \infty$, and in particular,

$$\liminf_{n\to\infty} \frac{1}{n} \log \widetilde{\mu}_n(B_\delta(y)) = 0\,.$$

□

5 Random Fields

5.1 Setting and definitions

A random field $\varphi = (\varphi_x)_{x\in\mathbb{Z}^d}$ over $\mathbb{Z}^d$ is the family of random variables $\varphi_x \in \mathbb{R}$.

5.2 The discrete Gaussian Free Field (DGFF)

We first revised some basic facts on Gaussian random variables and measures. We say $\varphi_\Lambda = (\varphi_x)_{x\in\Lambda}, \Lambda \subset \mathbb{Z}^d$ finite, is a *Gaussian vector* or *Gaussian random variable* or simply *Gaussian* if, for all $t_\Lambda = (t_x)_{x\in\Lambda} \in \mathbb{R}^\Lambda$, the real-valued random variable

$$\langle t_\lambda, \varphi_\Lambda\rangle := \sum_{x\in\Lambda} t_x\varphi_x$$

is a Gaussian random variable or a normal random variable (possibly degenerate when the variance vanishes), that is, it is normally distributed. Recall that a real-valued random variable X is a Gaussian random variable or a normal random variable if

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}tX}] = \exp\Big(\mathrm{i}t\mathbb{E}[X] - \frac{1}{2}t^2\mathrm{Var}(X)\Big) \qquad \text{for all } t \in \mathbb{R}\,.$$

If Φ_Λ is a Gaussian random variable with law/distribution $\mu_\Lambda \in \mathcal{M}_1(\mathbb{R}^\Lambda)$, we call μ_Λ a *finite-volume Gaussian measure* or simply a *Gaussian measure*. Suppose that φ_Λ is a Gaussian random variable with law/distribution $\mu_\Lambda \in \mathcal{M}_1(\mathbb{R}^\Lambda)$. Then it is easy to show via direct computation that the following holds.

$$\begin{aligned}
\mathbb{E}_{\mu_\Lambda}[\langle t_\Lambda, \varphi_\Lambda\rangle] &= \langle t_\Lambda, \mu_\Lambda\rangle\,, \quad m_\Lambda = (m_x)_{x\in\Lambda}\,, t_x = \mathbb{E}_{\mu_\Lambda}[\varphi_x]\,,\\
\mathrm{Var}_{\mu_\Lambda}(\langle t_\Lambda, \varphi_\Lambda\rangle) &= \langle t_\Lambda, C_\Lambda t_\Lambda\rangle\,, \quad C_\Lambda = (C(x,y))_{x,y\in\Lambda}\,, C(x,y) = \mathrm{Cov}_{\mu_\Lambda}(\varphi_x,\varphi_y)\,,\\
\mathbb{E}[\mathrm{e}^{\mathrm{i}\langle t_\Lambda,\varphi_\Lambda\rangle}] &= \exp\Big(\mathrm{i}\langle t_\Lambda, m_\Lambda\rangle - \frac{1}{2}\langle t_\Lambda, C_\Lambda t_\Lambda\rangle\Big)\,.
\end{aligned}$$

We also write $\varphi_\Lambda \sim \mathsf{N}(m_\Lambda, C_\Lambda)$, and φ_Λ is centred if $m_\Lambda = 0$. The matrix C_Λ is symmetric and nonnegative definite. When C_Λ is positive definite, then the matrix is invertible and there exist a density with respect to the Lebesgue measure, that is, $\varphi_\Lambda \sim \mathsf{N}(m_\Lambda, C_\Lambda)$ with C_Λ positive definite, has law

$$\mu_\Lambda(\mathrm{d}\varphi_\Lambda) = \frac{1}{(2\pi)^{|\Lambda|/2}\sqrt{\det(C_\Lambda)}}\,\mathrm{e}^{-\frac{1}{2}\langle\varphi_\Lambda - m_\Lambda, A_\Lambda(\varphi_\Lambda - m_\Lambda)\rangle}\prod_{x\in}\mathrm{d}\varphi_x\,, \varphi_\Lambda \in \mathbb{R}^\Lambda\,,$$

and $A_\Lambda = C_\Lambda^{-1}$. The *Laplace transform* for $J \in \mathbb{C}^\Lambda$ is

$$\int_{\mathbb{R}^\Lambda} \mathrm{e}^{-1/2\langle\varphi, A_\Lambda\varphi\rangle}\prod_{x\in\Lambda}\mathrm{d}\varphi_x = \det(2\pi C_\Lambda)^{1/2}\,\mathrm{e}^{\frac{1}{2}\langle J, C_\Lambda J\rangle}\,, J \in \mathbb{C}^\Lambda\,, \tag{5.1}$$

which follows by direct calculation (completing the square in the exponent).

Definition 5.1 Let $\Lambda \subset \mathbb{Z}^d$ and $\Delta \subset \Lambda$ a finite subset and $\mu_\Lambda \in \mathcal{M}_1(\mathbb{R}^\Lambda), \mu_\Delta \in \mathcal{M}_1(\mathbb{R}^\Delta)$. Then the probability measure μ_Δ is said be *compatible with* μ_Λ if

$$\mu_\Delta(A) = \mu_\Lambda(A \times \mathbb{R}^{\Lambda\setminus\Delta}) \qquad \text{whenever } A \in \mathcal{B}(\mathbb{R}^\Delta)\,.$$

Proposition 5.2 *Let $\Lambda, \Delta \subset$ be finite and $\Delta \subset \Lambda$. Then the following holds.*

(a) If μ_Λ is a Gaussian measure and $\mu_\Delta \in \mathcal{M}_1(\mathbb{R}^\Delta)$ is compatible with μ_Λ then μ_Δ is a Gaussian measure.

(b) Suppose μ_Δ and μ_Λ are Gaussian measures. Then μ_Δ is compatible with μ_Λ if and only if C_Δ is a submatrix of C_Λ.

Proof. This follows directly from the characteristic function and the Laplace transform (5.1). The details are left as an exercise in Gaussian calculus.

□

Definition 5.3 A measure $\mu \in \mathcal{M}_1(\mathbb{R}^{\mathbb{Z}^d})$ is said to be a *Gaussian measure* or an *infinite-volume Gaussian measure* if the compatible measure on $\mathbb{R}^\Lambda$ is Gaussian for all finite $\Lambda \subset \mathbb{Z}^d$. The family $\varphi = (\varphi_x)_{x\in\mathbb{Z}^d}$ of real-valued random variables $\varphi_x \in \mathbb{R}$ is called a *Gaussian field* if $\varphi_\Lambda = (\varphi_x)_{x\in\Lambda}$ is a Gaussian vector for all finite $\Lambda \subset \mathbb{Z}^d$.

Definition 5.4 The matrix $C = (C(x,y))_{x,y\in\mathbb{Z}^d}$ is *positive definite* if the submatrix $C_\Lambda = (C(x,y))_{x,y\in\Lambda}$ is positive definite for all finite $\Lambda \subset \mathbb{Z}^d$.

Remark 5.5 Given a positive definite matrix $C = (C(x.y))_{x,y\in\mathbb{Z}^d}$, for each finite $\Lambda \subset \mathbb{Z}^d$ there exists a unique Gaussian measure μ_Λ with covariance C_Λ. Whenever $\Lambda \subset \Lambda'$, μ_Λ is compatible with $\mu_{\Lambda'}$. By Kolmogorov's theorem there exists a probability measure $\mu \in \mathcal{M}_1(\mathbb{R}^{\mathbb{Z}^d})$ such that μ_Λ is compatible with μ for every finite $\Lambda \subset \mathbb{Z}^d$. By our last definition μ is a Gaussian measure (Gaussian field) and it has covariance C. It is uniqiue if we choose the σalgebra generated by all cylinder events. ◇

Proposition 5.6 *Suppose $C_j, j = 1,\dots,n$, are positive definite $\Lambda \times \Lambda$ matrices. If $\varphi_j \sim \mathsf{N}(0, C_j)$ and the $(\varphi_j)_{j=1,\dots,n}$ are independent then the sum is Gaussian, i.e.,*

$$\sum_{j=1}^n \varphi_j \sim \mathsf{N}(0, \sum_{j=1}^n C_j)\,.$$

The proof is left as am exercise.

Exercise 5.7 Prove the statement in Proposition 5.6 using the Laplace transform respectively the characteristic function. ☕☕

Proposition 5.8 (Gaussian moments) *Let $\mu \in \mathcal{M}_1(\mathbb{R}^{\mathbb{Z}^d})$ be a Gaussian measure, then*

$$\mathbb{E}_\mu[\varphi_{x_1}\cdots\varphi_{x_{2n}}] = \sum_{\mathcal{P}} \sum_{\{x,y\}\in\mathcal{P}} \mathbb{E}_\mu[\varphi_x\varphi_y]\,, \tag{5.2}$$

where $\mathcal{P}$ is the set of all partitions of $\{1,2,\dots,2n\}$ into subsets which each have two elements.

Proof. We just give a rough sketch, details are left as an exercise to the reader. First pick a sufficiently large set $\Lambda \subset \mathbb{Z}^d$ and compute the logarithmic generating function

$$\Lambda(t_\Lambda) = \log \mathbb{E}_{\mu_\Lambda}\Big[\mathrm{e}^{\langle t_\Lambda, \varphi_\Lambda\rangle}\Big]\,, \quad t_\Lambda \in \mathbb{R}^\Lambda\,,$$

Then one obtains the moments by taking partial derivatives,

$$\frac{\partial^{2n}\Lambda(t_\Lambda)}{\partial t_{i_1}\cdots\partial t_{i_{2n}}}\Big|_{t_\Lambda=0}\,.$$

□

Remark 5.9 If P is a polynomial in φ_Λ, then one can show that

$$\int_{\mathbb{R}^\Lambda} P(\varphi_\lambda)\,\mu(\mathrm{d}\varphi) = \exp\Big(\frac{1}{2}\sum_{x,y\in\Lambda} C(x,y)\frac{\partial}{\partial\varphi(x)}\frac{\partial}{\partial\varphi(y)}\Big)\Big|_{\varphi_\Lambda=0}P\,,$$

where the exponential is defined by expanding it as a power series. ◇

After this general introduction to Gaussian measure, we are now considering the discrete Gaussian Free Field (GFF). We start defining finite-volume distributions. We denote the space of infinite-volume configurations by $\Omega := \mathbb{R}^{\mathbb{Z}^d}$, and for a finite $\Lambda \subset \mathbb{Z}^d$ we write $\Omega_\Lambda = \mathbb{R}^\Lambda$. We write $\varphi_\Lambda = (\varphi_x)_{x\in\Lambda}$ and also denote the projection $\Omega \to \Omega_\Lambda$ by φ_Λ.

The σ-algebra generated by all cylinder events is denoted $\mathcal{F}$. Recall that a cylinder event is any event of the form

$$\{\omega \in \Omega \colon \varphi_\Delta(\omega) \in A\}\,, \quad \text{for some finite } \Delta \subset \mathbb{Z}^d \text{ and } A \in \mathcal{B}(\mathbb{R}^\Delta)\,.$$

The random field is defined in terms of an energy function, called *Hamiltonian* or *Hamilton function* . This function allows to specify the finite-volume distributions in $\Lambda \subset \mathbb{Z}^d$ finite with arbitrary boundary conditions.

Definition 5.10 (Hamiltonian and finite-volume distributions (GFF)) Let $\Lambda \subset \mathbb{Z}^d$ be finite, $\beta > 0, m \geq 0$.

(a) The Hamiltonian in Λ with *inverse temperature* β and *mass* m is defined as

$$H_{\Lambda,m}(\varphi) := \frac{\beta}{4d} \sum_{\substack{\{x,y\}\cap\Lambda\neq\varnothing \\ |x-y|=1}} (\varphi_x - \varphi_y)^2 + \frac{m^2}{2} \sum_{x\in\Lambda} \varphi_x^2 \,, \tag{5.3}$$

and we call the Hamiltonian *massless* when $m = 0$ and *massive* when $m \neq 0$.

(b) Let $\eta \in \Omega$ be a configuration. The finite-volume distribution (also called *Gibbs distribution*) in Λ with boundary condition η is the probability measure $\gamma^\eta_{\Lambda,m} \in \mathcal{M}_1(\Omega, \mathcal{F})$ defined by

$$\gamma^\eta_{\Lambda,m}(A) = \frac{1}{Z_{\Lambda,m}(\eta)} \int \mathbb{1}_A(\varphi)\,\mathrm{e}^{-H_{\Lambda,m}(\varphi)} \prod_{x\in\Lambda} \mathrm{d}\varphi_x \prod_{x\in\Lambda^{\mathrm{c}}} \delta_{\eta_x}(\mathrm{d}\varphi_x) \tag{5.4}$$

with normalisation, called *partition function*,

$$Z_{\Lambda,m}(\eta) = \int \mathrm{e}^{-H_{\Lambda,m}(\varphi)} \prod_{x\in\Lambda} \mathrm{d}\varphi_x \prod_{x\in\Lambda^{\mathrm{c}}} \delta_{\eta_x}(\mathrm{d}\varphi_x)\,. \tag{5.5}$$

Remark 5.11 (a) The finite-volume distributions are also called Gibbs distributions. A Gibbs measure on Ω is a probability measure $\mu \in \mathcal{M}_1(\Omega, \mathcal{F})$ whose conditional expectations are given by the Gibbs distribution, that is, for every finite $\Lambda \subset \mathbb{Z}^d, \eta \in \Omega$, and events $A \in \mathcal{F}$,

$$\mu(A|\mathcal{F}_{\Lambda^{\mathrm{c}}})(\eta) = \gamma^\eta_{\Lambda,m}(A)\,,$$

where $\mathcal{F}_{\Lambda^{\mathrm{c}}}$ is the σ-algebra of events outside of Λ.

(b) Suppose we would take empty boundary conditions, formally $\eta = \varnothing$, that is, there is no boundary to Λ. Then $Z_{\Lambda,m}(\varnothing) = \infty$. To see that suppose for simplicity that $\mathbb{Z} \supset \Lambda = \{x, y, z\}$, $x \sim y \sim z$ (nearest neighbours), $d = 1$, and $\beta/4d = 1$. Then

$$\begin{aligned}
&\int \exp\big(- \varphi_x^2 - (\varphi_x - \varphi_y)^2 - (\varphi_z - \varphi_z)^2 - \varphi_z^2\big)\,\mathrm{d}\varphi_x \mathrm{d}\varphi_y \mathrm{d}\varphi_z \\
&= \sqrt{\pi/2} \int \exp\big(- \varphi_y^2/2 + 2\varphi_y\varphi_z - 2\varphi_z^2\big)\,\mathrm{d}\varphi_y \mathrm{d}\varphi_z \\
&= \pi/2 \int \exp(-\varphi_y^2/2 + 4\varphi_y^2/8)\,\mathrm{d}\varphi_y = \infty\,.
\end{aligned}$$

◇

The next step is to rewrite the Hamiltonian into a quadratic form to obtain the Gaussian structure and the corresponding covariance matrix. This is essentially the discrete version of integration by parts, i.e., summation by parts. This require a couple of new notations and definitions. First recall that

$$x \sim y \Leftrightarrow |x-y| = 1\,.$$

We denote the set nearest neighbour bonds touching a finite $\Lambda \subset \mathbb{Z}^d$ by

$$\mathcal{E}_\Lambda := \{\{x,y\}\colon x,y \in \mathbb{Z}^d, \{x,y\}\cap\Lambda \neq \varnothing, x\sim y\}\,.$$

For each bond $\{x,y\} \in \mathcal{E}_\Lambda$, define the discrete gradient across this bond as

$$(\nabla\varphi)_{xy} := \varphi(y) - \varphi(x)\,,$$

and the *graph Laplacian*

$$(\mathcal{L}\varphi)(x) := \sum_{y\sim x} (\nabla\varphi)_{xy} = \sum_{j=1}^d \big(\varphi(x\pm \mathrm{e}_j) - \varphi(x)\big) = 2d\Delta\varphi(x)\,.$$

We define on the whole lattice $\mathbb{Z}^d$ the matrix

$$\mathcal{L}(x,y) := \begin{cases} -2d & \text{if } x = y\,,\\ 1 & \text{if } x\sim y\,,\\ 0 & \text{otherwise}\,,\end{cases} \tag{5.6}$$

and write

$$(\mathcal{L}\varphi)(x) = \sum_{y\in\mathbb{Z}^d} \mathcal{L}(x,y)\varphi(y)\,.$$

We use $\varphi_x \equiv \varphi(x)$ whenever it is convenient.

Lemma 5.12 (Discrete Green identities) *Let $\Lambda \subset \mathbb{Z}^d$ be finite and $\varphi,\eta,\psi \in \Omega$. Then the following holds.*

(a)

$$\sum_{\{x,y\}\in\mathcal{E}_\Lambda} (\nabla\psi)_{xy}(\nabla\varphi)_{xy} = -\sum_{x\in\Lambda}\psi(x)(\mathcal{L}\varphi)(x) + \sum_{\substack{x\in\Lambda, y\in\Lambda^c\\ x\sim y}} \psi(y)(\nabla\varphi)_{xy}\,.$$

(b)

$$\sum_{x\in\Lambda}\Big(\varphi(x)(\mathcal{L}\psi)(x) - \psi(x)(\mathcal{L}\varphi)(x)\Big) = \sum_{\substack{x\in\Lambda, y\in\Lambda\\ x\sim y}} \Big(\varphi(y)(\nabla\psi)_{xy} - \psi(y)(\nabla\varphi)_{xy}\Big)\,.$$

Proof. (a) For bonds in Λ we get using the symmetry between x and y (in all sums below x and y are nearest neighbours)

$$\begin{aligned}\sum_{\{x,y\}\in\mathcal{E}_\Lambda\cap\Lambda} (\nabla\psi_{xy}(\nabla\varphi)_{xy} &= \sum_{\{x,y\}\in\mathcal{E}_\Lambda\cap\Lambda}\psi(y)(\varphi(y)-\varphi(x)) - \sum_{\{x,y\}\in\mathcal{E}_\Lambda\cap\Lambda}\psi(x)(\varphi(y)-\varphi(x))\\ &= -\sum_{x\in\Lambda}\psi(x)\sum_{y\in\Lambda\colon y\sim x}(\varphi(y)-\varphi(x))\\ &= -\sum_{x\in\Lambda}\psi(x)(\mathcal{L}\varphi)(x) + \sum_{x\in\Lambda}\psi(x)\sum_{y\in\Lambda^c\colon y\sim x}(\varphi(y)-\varphi(x))\end{aligned}$$

We now add the nearest neighbour bonds that touch both Λ and $\Lambda^{\rm c}$.

$$\begin{aligned}\sum_{\{x,y\}\in\mathcal{E}_\Lambda}(\nabla\psi)_{xy}(\nabla\varphi)_{xy} &= \sum_{\{x,y\}\in\mathcal{E}_\Lambda\cap\Lambda}(\nabla\psi)_{xy}(\nabla\varphi)_{xy} + \sum_{\substack{x\in\Lambda,y\in\Lambda^{\rm c}\\ x\sim y}}(\nabla\psi)_{xy}(\nabla\varphi)_{xy}\\ &= -\sum_{x\in\Lambda}\psi(x)(\mathcal{L}\varphi)(x) + \sum_{\substack{x\in\Lambda,y\in\Lambda^{\rm c}\\ x\sim y}}\psi(y)(\varphi(y)-\varphi(x))\,.\end{aligned}$$

(b) To prove (b) use (a) twice, interchanging the roles of φ and ψ.

□

For $\Lambda\subset\mathbb{Z}^d$ finite the restriction $\mathcal{L}_\Lambda$ of the graph Laplacian $\mathcal{L}$ is

$$\mathcal{L}_\Lambda = (\mathcal{L}(x,y))_{x,y\in\Lambda}\,. \tag{5.7}$$

Suppose that $x\in\Lambda$. It is important to note that $(\mathcal{L}\varphi)(x)$ depends on some variables $\varphi(y)$ located outside of Λ whereas

$$(\mathcal{L}_\Lambda\varphi)(x) = \sum_{y\in\Lambda}\mathcal{L}(x,y)\varphi(y)$$

involves only field variables $\varphi(y)$ inside Λ. Note that

$$\langle\varphi,\mathcal{L}_\Lambda\varphi\rangle := \sum_{x,y\in\Lambda}\varphi(x)\mathcal{L}(x,y)\varphi(y) \quad \text{and } \langle\varphi,\mathcal{L}_\Lambda\varphi\rangle = \langle\mathcal{L}_\Lambda\varphi,\varphi\rangle\,.$$

Our aim is now to rewrite the Hamiltonian (5.3) with $m=0$. Suppose $\eta\in\Omega$ fixed and $\varphi\equiv\eta$ of Λ, i.e., $\varphi(x)=\eta(x)$ for all $x\in\mathbb{Z}^d\setminus\Lambda$. Using Lemma 5.12 we get

$$\begin{aligned}\sum_{\{x,y\}\in\mathcal{E}_\Lambda}(\varphi(x)-\varphi(y))^2 &= -\sum_{x\in\Lambda}\varphi(x)(\mathcal{L}\varphi)(x) + \sum_{x\in\Lambda}\sum_{y\in\Lambda^{\rm c}\colon y\sim x}\eta(y)(\nabla\varphi)_{xy}\\ &= -\langle\varphi,\mathcal{L}_\Lambda\varphi\rangle - 2\sum_{x\in\Lambda}\sum_{y\in\Lambda^{\rm c}\colon y\sim x}\varphi(x)\eta(y) + B_\Lambda(\eta)\,,\end{aligned} \tag{5.8}$$

where

$$B_\Lambda(\eta) = \sum_{x\in\Lambda}\sum_{y\in\Lambda^{\rm c}\colon y\sim x}\eta(y)^2$$

is a boundary term depending solely on η. Can we write the right hand side of (5.8) as a quadratic form $-\langle\varphi-u,\mathcal{L}_\Lambda(\varphi-u)\rangle$ for some $u\in\Omega$? We have

$$\begin{aligned}\langle\varphi-u,\mathcal{L}_\Lambda(\varphi-u)\rangle &= \langle\varphi,\mathcal{L}_\Lambda\varphi\rangle - 2\sum_{x\in\Lambda}\varphi(x)(\mathcal{L}_\Lambda u)(x) + \langle u,\mathcal{L}_\Lambda u\rangle\\ &= \langle\varphi,\mathcal{L}_\Lambda\varphi\rangle - 2\sum_{x\in\Lambda}\varphi(x)(\mathcal{L}u)(x) + 2\sum_{x\in\Lambda}\sum_{y\in\Lambda^{\rm c}\colon y\sim x}\varphi(x)u(y) + \widetilde{B}_\Lambda(u)\,,\end{aligned}$$

where

$$\widetilde{B}_\Lambda(u) = \langle u,\mathcal{L}u\rangle\,.$$

Now comparing this last expression with (5.8) we obtain

$$\sum_{\{x,y\}\in\mathcal{E}_\Lambda} (\varphi(x)-\varphi(y))^2 = -\langle \varphi - u, \mathcal{L}_\Lambda(\varphi-u)\rangle - 2\sum_{x\in\Lambda}\sum_{y\in\Lambda^c\colon y\sim x}\varphi(x)(u(y)-\varphi(y)) + \widehat{B}_\Lambda\,, \tag{5.9}$$

where $\widehat{B}_\Lambda$ depends only on η and u outside of Λ. We have the desired quadratic form as soon as the following two conditions are met:

(i) u is harmonic in Λ, i.e., $(\mathcal{L}u)(x) = 0$ for all $x \in \Lambda$.

(ii) $u = \eta$ off Λ, i.e., $u(y) = \eta(y)$ for all $y \in \Lambda^c$.

Lemma 5.13 *Suppose $\varphi \equiv \eta$ off $\Lambda \subset \mathbb{Z}^d$ finite. Assume that $u \in \Omega$ solves the Dirichlet problem (DP) in Λ with boundary condition η,*

$$\begin{cases} u & \text{is harmonic in } \Lambda, \text{ i.e., } (\mathcal{L}u)(x) = 0, x\in\Lambda\,,\\ u(x)=\eta(x) & \text{for all } x\in\Lambda^c\,. \end{cases}$$

Then

$$\sum_{\{x,y\}\in\mathcal{E}_\Lambda} (\varphi(x)-\varphi(y))^2 = -\langle \varphi - u, \mathcal{L}_\Lambda(\varphi-u)\rangle + \widehat{B}_\Lambda\,.$$

The Hamiltonian is given as

$$H_\Lambda(\varphi) = \frac{1}{2}\langle \varphi-u, (-\frac{1}{2d}\mathcal{L}_\Lambda)(\varphi-u)\rangle\,,$$

and

$$-\frac{1}{2d}\mathcal{L}_\Lambda = -\Delta_\Lambda = \mathbb{1}_\Lambda - P_\Lambda$$

with matrix $\mathbb{P}_\Lambda$ defined as

$$P_\Lambda(x,y) = \begin{cases} \frac{1}{2d} & x\sim y, x,y\in\Lambda\,,\\ 0 & \text{otherwise}\,. \end{cases}$$

Thus P_Λ is just the restriction of the transition matrix P of the SRW on $\mathbb{Z}^d$, see (3.2). We have now the Hamiltonian as a quadratic form, and to obtain the covariance matrix we need to simply get the inverse matrix of $\mathbb{1}_\Lambda - P_\Lambda$. We can improved the estimate on the stopping time $\tau_{\Lambda^c} = \inf\{k\in\mathbb{N}\colon S_k\notin\Lambda\}$ in Lemma 3.17 such there is a constant $c = c(\Lambda) > 0$ such that

$$P_x(\tau_{\Lambda^c} > n) \le \mathrm{e}^{-cn}\,. \tag{5.10}$$

Using this, we can prove the following statement.

Proposition 5.14 *For $\Lambda\subset\mathbb{Z}^d$ finite the matrix $1_\Lambda - P_\Lambda$ is invertible and*

$$(\mathbb{1}_\Lambda - P_\Lambda)^{-1} =: G_\Lambda$$

is the Green function of the SRW in Λ with killing upon leaving Λ,

$$G_\Lambda(x,y) = \mathbb{E}_x\Big[\sum_{k=0}^{\tau_{\Lambda^c}-1}\mathbb{1}\{S_k = y\}\Big]\,. \tag{5.11}$$

Proof.

$$(\mathbb{1}_\Lambda - P_\Lambda)(\mathbb{1}_\Lambda + P_\Lambda + P|L^2 + \cdots + P_\Lambda^n) = (\mathbb{1}_\Lambda - P_\Lambda^{n+1}).$$

For each $n \in \mathbb{N}$,

$$\begin{aligned} P_\Lambda^n(x,y) &= \sum_{x_1,\dots,x_{n-1}\in\Lambda} P(x,x_1)P(x_1,x_2)\cdots P(x_{n-1},y) = P_x(S_n = y, \tau_{\Lambda^c} > n) \\ &\le P_x(\tau_{\Lambda^c} > n) \le \mathrm{e}^{-cn}, \end{aligned}$$

where we used (5.10). Thus the series $G_\Lambda = \mathbb{1}_\Lambda + P_\Lambda + P_\Lambda^2 + \cdots$ converges and

$$(\mathbb{1}_\Lambda - P_\Lambda)G_\Lambda = \mathbb{1}_\Lambda.$$

By symmetry, $G_\Lambda(\mathbb{1}_\Lambda - P_\Lambda) = \mathbb{1}_\Lambda$. Furthermore, as the random walk is killed upon leaving Λ, noting that

$$\mathbb{E}_x[\mathbb{1}\{S_k = y\}] = P_\Lambda^k(x,y),$$

$$G_\Lambda(x,y) = \mathbb{E}_x\Big[\sum_{k=0}^{\tau_{\Lambda^c}-1} \mathbb{1}\{S_k = y\}\Big].$$

□

Proposition 5.15 *Let $\Lambda \subset \mathbb{Z}^d$ be finite and $\eta \in \Omega$. The solution $u \in \Omega$ to the Dirchlet problem is given by*

$$u(x) := \mathbb{E}_x[\eta(S_{\tau_{\Lambda^c}})], \quad x \in \mathbb{Z}^d. \tag{5.12}$$

Proof. Suppose $y \in \Lambda^c$, then $P_y(\tau_{\Lambda^c} = 0) = 1$, and thus

$$u(y) = \mathbb{E}_y[\eta(S_0)] = \eta(y).$$

Now let $x \in \Lambda$. Using the Markov property of the SRW we obtain

$$\begin{aligned} u(x) = \mathbb{E}_x[\eta(S_{\tau_{\Lambda^c}})] &= \sum_{y\colon y\sim x} E_x[\eta(S_{\tau_{\Lambda^c}}), S_1 = y] = \sum_{y\colon y\sim x} P_x(S_1 = y)\mathbb{E}_x[\eta(S_{\tau_{\Lambda^c}})|S_1 = y] \\ &= \frac{1}{2d}\sum_{y\colon y\sim x} u(y), \end{aligned}$$

which implies that $(\Delta u)(x) = 0$ for all $x \in \Lambda$. Thus u is harmonic in Λ and $u \equiv \eta$ off Λ.

□

Theorem 5.16 *Under γ_Λ^η (note $m \equiv 0$) $(\varphi(x))_{x\in\Lambda}$ is a Gaussian random vector with mean $u_\Lambda = (u(x))_{x\in\Lambda}$,*

$$u(x) = \mathbb{E}_x[\eta(S_{\tau_{\Lambda^c}})],$$

and covariance matrix $G_\Lambda = (G_\Lambda(x,y))_{x,y\in\Lambda}$.

Proof. The proof follows the steps outlined above. The normalisation (partition function) of the finite-volume distribution is the Gaussian integral

$$Z_\Lambda(\eta)=\int \mathrm{e}^{-\frac{1}{2}\langle\varphi-u,(-\Delta_\Lambda)(\varphi-u)\rangle}\prod_{x\in\Lambda}\mathrm{d}\varphi(x)\prod_{x\in\Lambda^{\mathrm{c}}}\delta_{\eta(x)}(\mathrm{d}\varphi(x))=(2\pi)^{|\Lambda|/2}\big(\det(-\Delta_\Lambda)\big)^{-1/2}\,.$$

□

In the following, consider sequence of centred boxes $\Lambda_N:=[-N,N]^d\cap\mathbb{Z}^d$. Note that the finite-volume distribution $\gamma^\eta_{\Lambda_N}$ depends on the boundary condition $\eta\in\Omega$ only through its mean. The covariance is only sensitive to the choice of the box Λ_N,

$$G_{\Lambda_N}(x,y)=\mathbb{E}_x\Big[\sum_{k=0}^{\tau_{\Lambda_N^{\mathrm{c}}}-1}1\!\!1\{S_k=y\}\Big]\,.$$

Note that this is the Green of the SRW with killing upon leaving the box Λ_N. Indeed, this is similar to the Green function of the SRW in $\mathbb{Z}^d$, see (3.16),(3.17),(3.18),(3.19). If we increase the box Λ_N, i.e., consider $N\to\infty$, then we shall obtain the Green function of the SRW in $\mathbb{Z}^d$, see (3.18) and (3.19). For the centred boxes Λ_N we may ask about the fluctuations of the field variable φ_0, that is, what is the limit $N\to\infty$ of

$$\mathrm{Var}_{\gamma^\eta_{\Lambda_N}}(\varphi_0)=G_{\Lambda_N}(0,0)\,.$$

By monotone convergence,

$$\lim_{N\to\infty}G_{\Lambda_N}(0,0)=\mathbb{E}_0\Big[\sum_{k=0}^{\infty}1\!\!1\{S_k=0\}\Big]$$

is just the expected number if visits of the SRW at the origin. We have learnt earlier that the variance (number of visits) diverges if the SRW is recurrent ($d=1,2$). Asymptotically, as $N\to\infty$,

$$G_{\Lambda_N}(0,0)\sim\begin{cases}N & \text{if } d=1\,,\\ \log N & \text{if } d=2\,.\end{cases}$$

We therefore call the (Gaussian) random field *delocalised* if the variance of φ_0 grows unboundedly with the volume, i.e., in dimension $d=1,2$. For dimensions $d\geq 3$, the variance remains bounded, and we therefore expect the filed to remain *localised* close to its mean value in the limit $N\to\infty$, i.e.,

$$G(x,y)=\lim_{N\to\infty}G_{\Lambda_N}(x,y)=\mathbb{E}_x\Big[\sum_{k=0}^{\infty}1\!\!1\{S_k=y\}\Big]$$

is finite. Thus, for $d\geq 3$ and $m=0$, given any harmonic function η on $\mathbb{Z}^d$, there exists a Gaussian measure $\mu^\eta\in\mathcal{M}_1(\Omega)$ with mean η and covariance matrix G. The asymptotic behaviour of the Green functions is derive din (3.19) via the LCLT.

Proposition 5.17 *$d \geq 3, m = 0$, and $\eta \in \Omega$ harmonic. Then, for $x, y \in \mathbb{Z}^d$, in the limit $|x-y| \to \infty$,*

$$\mathrm{Cov}_{\mu^\eta}(\varphi_x, \varphi_y) = \frac{a_d}{|x-y|^{d-2}}(1 + o(1)) ,$$

where a_d is the dimension dependent constant from (3.19).

We highlight the connection between SRW and Gaussian random field again by revisiting Section 3.4. Recall the transition probability $P(x_1, t; x_0, t_0)$ of the SRW as the probability that the walker is at x_1 at time t when he was at x_0 at time t_0. Without loss of generality let $t_0 = 0, x_0 = 0$ and $x_1 = x$ and write

$$P(x, t; 0, 0) = \frac{1}{(2\pi)^d} \int_{\mathsf{BZ}} \widehat{P}(k, t)\, \mathrm{e}^{\mathrm{i}\langle k, x\rangle}\, \mathrm{d}k = \frac{1}{(2\pi)^d} \int_{\mathsf{BZ}} \Big(\frac{1}{d} \sum_{j=1}^d \cos(k_j)\Big)^t \mathrm{e}^{\mathrm{i}\langle k, x\rangle}\, \mathrm{d}k\, . \tag{5.13}$$

Thus the Green function is

$$\begin{aligned} G(x) = \sum_{t=0}^\infty P(x, t; 0, 0) &= \frac{1}{(2\pi)^d} \int_{\mathsf{BZ}} \Big(\sum_{t=0}^\infty \Big(\frac{1}{d} \sum_{j=1}^d \cos(k_j)\Big)^t\Big) \mathrm{e}^{\mathrm{i}\langle k, x\rangle}\, \mathrm{d}k \\ &= \frac{1}{(2\pi)^d} \int_{\mathsf{BZ}} \Big(\frac{1}{1 - \widehat{p}(k)}\Big)\, \mathrm{e}^{\mathrm{i}\langle k, x\rangle}\, \mathrm{d}k\, , \end{aligned} \tag{5.14}$$

where

$$\widehat{p}(k) = \sum_{x \in \mathbb{Z}^d} p(0, x)\, \mathrm{e}^{\mathrm{i}\langle k, x\rangle} = \frac{1}{d} \sum_{j=1}^d \cos(k_j)$$

with $P = (p(x, y))_{x,y \in \mathbb{Z}^d}$ being transition matrix of the SRW. For this one needs to justify interchanging summation with integration and the use of the geometric series. We are leave this technical details for the reader, alternatively, they can be found in [Spi01] or [Law96, LL10].

5.3 Scaling limits

Appendices

A 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 $X_n\colon \Omega \to E$ 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-σ-field (algebra) $\mathcal{B}(E)$. Let X be a random variable taking values in (E, d).

Definition A.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

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = \mathbb{P}(\{\omega \in \Omega\colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

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 A.2 (Convergence in probability) The sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X if

$$\lim_{n\to\infty} \mathbb{P}(d(X_n, X) > \varepsilon) = 0, \qquad \text{for all } \varepsilon > 0.$$

We write $X_n \xrightarrow{\text{P}} X$ for convergence in probability.

Proposition A.3 (Markov's inequality) *Let Y be a real-valued random variable and $f\colon [0,\infty) \to [0,\infty)$ 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 A.4 (Chebyshev's inequality, 1867) *For all $Y \in \mathcal{L}^2$ 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 $\mathbb{E}[d(X_n, X) \wedge 1] \to 0$ as $n \to \infty$. This is related to the almost sure convergence as follows.

Lemma A.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 $N'' \subset \mathbb{N}'$ such that $X_n \to X$ 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 A.6 (Convergence in distribution) We say that X_n converges in distribution to X, if, for every bounded continuous function $f\colon E \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f].$$

We write $X_n \xrightarrow{\text{d}} X$ for convergence in distribution.

Remark A.7 (a) $X_n \xrightarrow{\text{d}} X$ is equivalent to weak convergence of the distributions.

(b) if $X_n \xrightarrow{\text{d}} X$ and $g\colon E \to \mathbb{R}$ continuous, then $g(X_n) \xrightarrow{\text{d}} g(X)$. But note that, if $E = \mathbb{R}$ and $X_n \xrightarrow{\text{d}} X$, 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 $E = \{1, \ldots, m\}$ is finite and $d(x, y) = 1 - \mathbb{1}_{x=y}$. Then $X_n \xrightarrow{\text{d}} X$ if and only if $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$ for all $k \in E$.

(d) Let $E = [0, 1]$ and $X_n = 1/n$ almost surely. Then $X_n \xrightarrow{\text{d}} X$, where $X = 0$ almost surely. However, note that $\lim_{n\to\infty} \mathbb{P}(X_n = 0) = 0 \neq \mathbb{P}(X = 0)$.

◇

B Law of large numbers and the central limit theorem

Definition B.1 (Variance and covariance) Let $X, Y \in \mathcal{L}^2$ 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 $|XY| \le X^2 + Y^2$.

(c) If $\text{Cov}(X, Y) = 0$, then X and Y are called **uncorrelated**.

Theorem B.2 (Weak law of large numbers, $\mathcal{L}^2$-version) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of uncorrelated (e.g. independent) real-valued random variables in $\mathcal{L}^2$ with bounded variance, in that $v := \sup_{n\in\mathbb{N}} \text{Var}(X_n) < \infty$. Then for all $\varepsilon > 0$*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\Big| \ge \varepsilon\Big) \le \frac{v}{n\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

and thus $1/n\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow{\text{P}} 0$. In particular, if $\mathbb{E}[X_i] = \mathbb{E}[X_1]$ for all $i \in \mathbb{N}$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} \mathbb{E}[X_1].$$

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 B.3 (Weak law of large numbers, $\mathcal{L}^1$-version) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of pairwise independent, identically distributed real-valued random variables in $\mathcal{L}^1$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} \mathbb{E}[X_1].$$

Theorem B.4 (Strong law of large numbers) *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of pairwise uncorrelated real-valued random variables in $\mathcal{L}^2$ with $v := \sup_{n\in\mathbb{N}} \text{Var}(X_n) < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \to 0 \textit{ almost surely as } n \to \infty.$$

Theorem B.5 (Central limit theorem; A.M. Lyapunov 1901, J.W. Lindeberg 1922,P. Leévy 1922) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed real-valued random variables in $\mathcal{L}^2$ with $\mathbb{E}[X_i] = m$ and $\text{Var}(X_i) = v > 0$. Then,*

$$S_n^* := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

The normal distribution is defined in the following section.

C Normal distribution

A real-valued random variable X is **normally** distributed with mean μ and variance $\sigma^2 > 0$ if

$$\mathbb{P}(X > x) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_x^{\infty} \mathrm{e}^{-\frac{(u-\mu)^2}{2\sigma^2}}\,\mathrm{d}u, \qquad \text{for all } x \in \mathbb{R}.$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$. We say that X is standard normal distributed if $X \sim \mathsf{N}(0,1)$.

A random vector $X = (X_1, \ldots, X_n)$ is called a **Gaussian random vector** if there exits 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. $Y_i \sim \mathsf{N}(0,1)$ for $i = 1, \ldots, m$. Likewise, a random variable $Y = (Y_1, \ldots, Y_m)$ 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 C.1 *If A is an orthogonal $n \times n$ matrix, i.e. $AA^T = \mathbb{1}$, and X is a n-dimensional standard Gaussian vector, then AX is also a n-dimensional standard Gaussian vector.*

Lemma C.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 C.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 C.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 C.5 (Inequalities) *Let $X \sim \mathsf{N}(0,1)$. Then for all $x > 0$,*

$$\frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2} \le \mathbb{P}(X > x) \le \frac{1}{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,.$$

D Gaussian integration formulae

For any $a > 0$,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2}\,\mathrm{d}x = \sqrt{\pi/a}.$$

For $b \in \mathbb{C}$ and $a > 0$,

$$I(b) = \int_{-\infty}^{\infty} \mathrm{e}^{-a/2x^2+bx}\,\mathrm{d}x = \mathrm{e}^{b^2/2a}\sqrt{2\pi/a}\,.$$

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 $\varphi,\psi \in \mathbb{R}^n$.

$$\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{1}{2}\langle\varphi,A\varphi\rangle}\prod_{i=1}^n \mathrm{d}\varphi_i = (2\pi)^{n/2}\det(A^{-\frac{1}{2}}) = \det(2\pi C)^{\frac{1}{2}}.$$

For any $J \in \mathbb{C}^n$ we obtain

$$\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{1}{2}\langle\varphi,A\varphi\rangle+\langle J,\varphi\rangle}\prod_{i=1}^n \mathrm{d}\varphi_i = \det(2\pi C)^{\frac{1}{2}}\mathrm{e}^{\frac{1}{2}\langle J,CJ\rangle}\,.$$

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

$$\mu_C(\mathrm{d}\varphi) = \frac{1}{\sqrt{\det(2\pi C)}}\mathrm{e}^{-1/2\langle\varphi,C^{-1}\varphi\rangle}\prod_{i=1}^n \mathrm{d}\varphi_i,$$

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 $F \in \mathcal{L}(\mu_C)$,

$$\begin{aligned}\int_{\mathbb{R}^n} F(\varphi)\mu_C(\mathrm{d}\varphi) &= \int_{\mathbb{R}^n} \mu_{C_1}(\mathrm{d}\varphi_1)\int_{\mathbb{R}^n} \mu_{C_2}(\mathrm{d}\varphi_2)F(\varphi_1+\varphi_2)\\ &= \int_{\mathbb{R}^n} \mu_{C_1}(\mathrm{d}\varphi)\int_{\mathbb{R}^n} \mu_{C_2}(\mathrm{d}(\varphi-\varphi_1))F(\varphi).\end{aligned}$$

In other words, if $C = C_1 + C_2$, the Gaussian random variable φ 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 $X = (X_1, \ldots, X_n)$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix C reads as

$$\varphi_X(t) = \mathbb{E}\Big[\mathrm{e}^{i\langle t,\mu\rangle - \frac{1}{2}\langle t, Ct\rangle}\Big], \quad t \in \mathbb{R}^n.$$

An $\mathbb{R}^n$-valued stochastic process $X = \{X_t \colon t \geq 0\}$ 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 $(X_{t_1}, \ldots, X_{t_k})$ has a joint normal distribution. If the distribution of $(X_{t+t_1}, \ldots, X_{t+t_n})$ 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

$$\varrho(s,t) := \mathbb{E}[(X_s - m(s))(X_t - m(t))^T], \qquad s,t \geq 0.$$

If $m(t) = 0$ for all $t \geq 0$, we say that X is a zero-mean Gaussian process.

Corollary D.1 *One-dimensional BM is a zero-mean Gaussian process with covariance formula*

$$\varrho(s,t) = s \wedge t, \qquad s,t \geq 0.$$

E Some useful properties of the weak topology of probability measures

A probability measure $\mu \in \mathcal{M}_1(E)$ on a metric space (E,d) is *tight* if for each $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset E$ such that $\mu(K_\varepsilon^{\mathrm{c}}) < \varepsilon$. A family $(\mu_n)_{n\in I}$ of probability measures on the metric space (E,d) is called a tight family if the set K_ε may be chosen independently of $n \in I$, that is, for all $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset E$ and $n_0 \in I$ such that $\mu_n(K_\varepsilon^{\mathrm{c}}) < \varepsilon$ for all $n \geq n_0$.

Definition E.1 (Weak convergence of probability measures) A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on a metric space (E,d) converges weakly to $\mu \in \mathcal{M}_1(E)$ as $n \to \infty$ if

$$\int_E f(x)\,\mu_n(\mathrm{d}x) \to \int_E f(x)\,\mu(\mathrm{d}x) \quad \text{for all } f \in \mathcal{C}_{\mathrm{b}}(E) \text{ as } n \to \infty.$$

Lemma E.2 *A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on a metric space (E,d) converges weakly to $\mu \in \mathcal{M}_1(E)$ as $n \to \infty$ if*

$$\begin{aligned} \limsup_{n\to\infty} \mu_n(C) &\leq \mu(C) \quad \textit{for all closed } C \subset E, \\ \liminf_{n\to\infty} \mu_n(O) &\geq \mu(O) \quad \textit{for all open } O \subset E. \end{aligned} \tag{E.1}$$

The set of probability measures $\mathcal{M}_1(E)$ on a Polish space (E, d) is itself a Polish space. Note that $\mathcal{M}_1(E) \subset \mathcal{M}(E)$ is a closed convex subset of the (vector) - space of all finite signed measures on E. We equip $\mathcal{M}(E)$ with the topology generated by sets

$$\Big\{\beta \in \mathcal{M}(E) \colon |\int_E f(x)\,\mathrm{d}(\beta(x) - \alpha(x))| < r\Big\},$$

where $\alpha \in \mathcal{M}(E)$, $f \in \mathcal{C}_\mathrm{b}(E)$, and $r > 0$. The norm on $\mathcal{M}(E)$ is the total variation norm

$$\|\alpha\|_{\mathrm{var}} := \sup\{\int_E f(x)\alpha(\mathrm{d}x) \colon f \in \mathcal{C}_\mathrm{b}(E) \text{ with } \|f\|_\infty \le 1\}, \quad \alpha \in \mathcal{M}(E).$$

The norm $\|\cdot\|_{\mathrm{var}}$ is lower semi-continuous on $\mathcal{M}(E)$ and therefore certainly measurable on $\mathcal{M}(E)$; and clearly, $\|\cdot\|_{\mathrm{var}}$ is bounded on $\mathcal{M}_1(E)$. The Lévy metric on $\mathcal{M}_1(E)$ is a complte separable metric, which is consistent (inherited from) with the restriction of the topology on $\mathcal{M}(E)$ to the closed and convex subset $\mathcal{M}_1(E)$. Following Lévy and Prohorov, define the Lévy metric as

$$\mathrm{d}(\alpha, \nu) := \inf\Big\{\delta > 0 \colon \alpha(F) \le \nu(F^{(\delta)}) + \delta \text{ and } \nu(F) \le \alpha(F^{(\delta)}) + \delta \text{ for all closed } F \subset E\Big\},$$

$\alpha, \nu \in \mathcal{M}_1(E)$, where $F^{(\delta)}$ is defined relative to a complete metric on E, that is, $F^{(\delta)}$ is the open δ-hull of F. Since it is clear that $\mathrm{d}(\alpha, \nu) \le \|\alpha - \nu\|_{\mathrm{var}}$, all that remains to show is that the Lévy metric d is compatible with the weak topology in Definition E.1 and Lemma E.2 and that $(\mathcal{M}_1(E), \mathrm{d})$ is a Polish space. To show this one uses the tightness criterion, Lemma E.2 (the upper bound), and the following: Suppose that $\mathcal{F} \subset \mathcal{C}_\mathrm{b}(E)$ is a set of uniformly bounded test functions which is equicontinuous on every compact subset of E. Then the weak convergence $\alpha_n \Rightarrow \nu$ implies that

$$\sup\{|\int f(x)\alpha_n(\mathrm{d}x) - \int f(x)\,\nu(\mathrm{d}x)| \colon f \in \mathcal{F}\} \to 0 \text{ as } n \to \infty.$$

This is the content of the following lemma which is proved in the book by Billingsley on Convergence of probability measures.

Lemma E.3 (Lévy & Prohorov) *The Lévy metric* d *(defined above) is compatible with the weak topology on $\mathcal{M}_1(E)$, and* $(\mathcal{M}_1(E), \mathrm{d})$ *is a Polish space.*

We will frequently use the following dual space for $\mathcal{M}(E)$ (note that $\mathcal{M}_1(E)$ is not a vector space).

Lemma E.4 *The duality relation*

$$(f, \nu) \in \mathcal{C}_\mathrm{b}(E) \times \mathcal{M}(E) \mapsto \int_E f(x)\,\nu(\mathrm{d}x)$$

determines a representation of $\mathcal{M}^*$ as $\mathcal{C}_\mathrm{b}(E)$.

Theorem E.5 (Prohorov) *Let (E, d) be a Polish space, and let $\Gamma \subset \mathcal{M}_1(E)$. The $\overline{\Gamma}$ is compact iff Γ is tight.*

We shall need some version for the path space $E := \mathcal{C}([0,1];\mathbb{R})$.

Proposition E.6 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on $\mathcal{C}([0,1];\mathbb{R})$ which converges weakly to μ. Let A be a Borel set in $\mathcal{C}([0,1];\mathbb{R})$ with $\mu(\partial A) = 0$. Then $\mu_n(A) \to \mu(A)$ as $\nu \to \infty$.*

We need an adaptation of Prohorov's theorem suited to the path space $\mathcal{C}([0,1];\mathbb{R})$,

Theorem E.7 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of Borel probability measures on $\mathcal{C}([0,1];\mathbb{R})$ with the following two properties:*

(a) The finite dimensional distributions converge. That is, for any $0 \le t_1 < t_2 < \cdots < t_m \le 1$, $m \in \mathbb{N}$, there is a measure $\mu_{t^{(m)}} \in \mathcal{M}_1(\mathbb{R}^m)$ so that, as $n \to \infty$,

$$\int f(\omega(t_1),\dots,\omega(t_m))\,\mu_n(\mathrm{d}\omega) \to \int f(x_1,\dots,x_m)\,\mu_{t^{(m)}}(\mathrm{d}x) \text{ for all } f \in \mathcal{C}_\mathrm{b}(\mathbb{R}^m).$$

(b) $(\mu_n)_{n\in\mathbb{N}}$ is tight.

Then, there is a probability measure μ on $\mathcal{C}([0,1];\mathbb{R})$ so that $\mu_n \to \mu$ weakly as $n \to \infty$, and the finite dimensional distributions of μ are the $\mu_{t^{(m)}}$.

References

[BB01] H. BAUER and R. B. BURCKEL. *Measure and Integration Theory*. De Gruyter, 2001.

[Bil99] P. BILLINGSLEY. *Convergence of Probability Measures*, vol. 316 of *Wiley Series in Probability and Statistics*. Wiley, 1999.

[Bil12] P. BILLINGSLEY. *Probability and Measure:*, vol. 894 of *Wiley Series in Probability and Statistics*. Wiley, 2012.

[CK81] I. CSISZÁR and J. KÖRNER. *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadaémiai Kiadó, Budapest, 1981.

[Coh13] D. L. COHN. *Measure Theory*. Birkhäuser, 2nd rev. and ext. ed., 2013.

[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.

[Geo12] H.-O. GEORGII. *Stochastics*. De Gruyter Textbook. De Gruyter, 2nd rev. and ext. ed., 2012.

[Kal02] O. KALLENBERG. *Foundations of Modern Probability*. Springer, 2002.

[KS98] I. KARATZAS and S. E. SHREVE. *Brownian Motion and Stochastic Calculus*. Springer, 2nd ed., 1998.

[Law96] G. F. LAWLER. *Intersections of Random Walks*. Birkhauser, 1996.

[LL10] G. F. LAWLER and V. LIMIC. *Random Walks: A Modern Introduction*. Cambridge University Press, 2010.

[MP10] P. MÖRTERS and Y. PERES. *Brownian Motion*. Cambridge University Press, 2010.

[Spi01] F. SPITZER. *Principles of Randon Walk*. Springer, 2nd ed., 2001.

[Str93] D. W. STROOCK. *Probability Theory - An analytic view*. Cambridge University Press, 1993.

Index

μ-continuity set, 28
τ-topology, 27
n-step transition probbaility, 33
rth moment, 8

asymptotic event, 15

Bayes' formula, 9
Brownian motion, 50

characteristic function, 22
Chebyshev's inequality, 13
compatible, 85, 86
Conditional probability, 9
convergence in distribution, 18
convolution, 21
Covariance, 8
cummulative distribution function, 20
cumulant generating function, 61
cumulative distribution function (CDF), 5
cylinder set, 49

delocalised, 93
diffusion equation, 46
duality, 26

empirical measure, 58
entropy, 58
exponential tightness, 71
exposed point, 81
exposing hyperplane, 81

Fenchel-Legendre transform, 61
finite-dimensional probabilities, 48
Fourier transform of a probability measure, 25

Gâteaux differentiable, 81
Gaussian, 85
Gaussian field, 86
Gaussian measure (infinite volume), 86

Hamiltonian, 87, 88
heat equation, 46

Independence, 10
initial distribution, 34

Jensen's inequality, 12

Lévy's metric, 31
large deviation principle (LDP), 55
level set, 54
Linear Markov's inequality, 12
localised, 93
lower semicontinuous, 73

macro state, 57
Markov chain, 34
Markov process, 34
Markov property, 33
Markov's inequality, 13
martingale, 39
micro state, 57
Minkowski's inequality, 12
modulus of continuity, 52
moment generation function (MGF), 5

normal distribution, 20

rate function, 54
recurrent, 38
regular, 73
regular topological space, 73

simple random walk (SRW), 33
standard deviation, 8
stochastic matrix, 34
strong topology, 27

tail event, 15
tight, 27
transient, 38

upper semicontinuous, 73

Variance, 8

weak convergence, 27
weak large deviation principle (LDP), 56
Wiener measure, 53