SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Large Deviations and Weak Convergence of Measures, with applications to Monte Carlo Estimators

av

Johan Ericsson

2024 - No M3

Large Deviations and Weak Convergence of Measures, with applications to Monte Carlo Estimators

Johan Ericsson

Självständigt arbete i matematik 30 högskolepoäng, avancerad nivå

Handledare: Yishao Zhou

2024

Abstract

In this thesis we apply the large deviations principle to study the performance of Monte Carlo estimators for rare events. We introduce weak convergence of measures and study the topological structure of the collection of finite signed measures in the weak topology and in the τ-topology. We prove the law of large numbers for the empirical distributions of the importance sampling estimator and Sanov's Theorem in the τ-topology for random variables taking values in a Polish space and then more generally in a measurable space. We also introduce a version of Sanov's Theorem for the empirical distributions of importance sampling estimators. This is used to study the performance of importance sampling and crude Monte Carlo estimators.

Keywords: large deviation principle, importance sampling, rare events, weighted empirical measures, weak convergence, τ-topology.

Sammanfattning

I detta examensarbete använder vi teorin om stora avvikelser för att studera konvergensegenskaperna hos Monte Carlo estimatorer för sällsynta händelser. Vi introducerar svag konvergens av mått och studerar den topologiska strukturen hos mängden av ändliga mått bestyckade med topologin associerad med svag konvergens av mått och bestyckade med den så kallade τ-topologin. Vi bevisar de stora talens lag för de empiriska distributionerna till den så kallade *importance sampling* estimatorn och Sanov's Sats i τ-topologin för stokastiska variabler när värdemängden är ett polskt rum, och mer generellt ett mätbart rum. Vi introducerar också en version av Sanov's Sats som håller för de empiriska distributionerna till *importance sampling estimatorn*. Vi tillämpar sedan dessa metoder för att studera konvergensegenskaperna hos olika Monte Carlo estimatorer.

Nyckelord: Stora avvikelser, importance sampling, sällsynta händelser, viktade empiriska mått, svag konvergens, τ-topologin.

Acknowledgements

I would like to thank my advisor Yishao Zhou for her support and for taking on a student with a project idea of his own. Thank you for your help and feedback during the writing process. Your comments and attention to detail when reading the drafts have been invaluable.

Contents

1 Introduction

Stochastic models, which include random variables, play an important role in modern society with applications in a diverse range of areas including weather forecasting, networking systems, and physics, among others. The object of interest in these models is often the probability of some event or more generally the expected value of an unknown random variable. The techniques used to algorithmically solve these types of problems are generally referred to as *stochastic simulation* techniques, of which the most widely used class is *Monte Carlo (MC)* simulation. The heuristic idea behind Monte Carlo simulation is the following: if we are able to simulate samples from a distribution then the mean of the samples should be a good approximation of the mean of the distribution.

Consider a real valued random variable X with distribution μ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and suppose that we are interested in computing the expected value of X given by

$$\theta = \mathbb{E}[X] = \int_{\Omega} X \, \mathrm{d}\mathbb{P} \, . \tag{1}$$

The *crude Monte Carlo (CMC)* method to approximate θ consists of simulating a sequence of independent identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$ with the same distribution, μ, as X. Then the CMC estimator of θ is given by

$$\theta_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) \approx \mathbb{E}[X]. \tag{2}$$

The strong law of large numbers implies that the CMC estimator converges almost surely to the expected value $\mathbb{E}[X]$ as n goes to infinity. However, when $\mathbb{E}[X]$ is small, then the convergence rate of CMC is too slow for many applications; making it unpractical. There are two main factors that determine the convergence rate of Monte Carlo estimators: the variance of the estimator, and the sample size n. A standard technique to speed up Monte Carlo simulation is to replace the Monte Carlo estimator with another estimator that has lower variance. One such method which is widely used is *importance sampling (IS)*.

The general idea in importance sampling is to make a change of measure and sample from another distribution which better represents the region of interest for the simulation. Let μ be the distribution of X and assume that we can simulate samples of some real valued random variables Y_i with distribution π. If μ is absolutely continuous with respect to π and ρ denotes the Radon-Nikodym derivative of μ with respect to π, then the corresponding importance sampling estimator of θ is given by

$$I_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\omega) \rho(Y_i(\omega)) \approx \mathbb{E}[X]. \tag{3}$$

The choice of proposal distribution π is crucial to the performance of importance sampling, and a good choice of π can result in a much lower variance of the importance sampling estimator compared to the CMC estimator in equation (2).

Many events of interest to practitioners have very low probabilities of occurring, and importance sampling is one of the main techniques to efficiently simulate the probabilites of such events. In the stochastic simulation literature the term *rare event* is used for events with little to no probability of occurring. As an example, consider an insurance company: the business model consists of correctly pricing the insurance premiums such that all the

incoming insurance claims can be paid. If the collective value of all insurance claims is too high, then the insurance company will run the risk of insolvency. The probability of insolvency is commonly called ruin probability and the Swedish mathematician and actuary Harald Cramér (1893-1985) was one of the pioneers of ruin theory in insurance mathematics. In 1938 Cramér published a paper [22] in which he proved bounds for the probabilities that sums of independent identically distributed random variables deviate from their expected value. The results of Cramér are considered the first in the mathematical theory of large deviations.

The large deviations principle is used to estimate probabilites at an exponential scale and it is therefore especially suitable for estimating rare event probabilites. It would take almost 30 years from when Cramér published his paper until the seminal paper [52] by S.R. Srinivasa Varadhan[1] was published, which laid much of the foundation of the modern theory of large deviations. Consider a sequence of independent identically distributed random variables (X_n) taking values in a Hausdorff topological space $\mathcal{X}$ with distributions μ_n on the Borel σ-algebra $\mathcal{B}_\mathcal{X}$ of $\mathcal{X}$. The theory of large deviations is concerned with finding a lower semicontinuous function $I : \mathcal{X} \to [0, \infty]$, called a rate function, such that

$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \lim_{n\to\infty} \mu_n(A) \approx \lim_{n\to\infty} e^{-n \inf_{x\in A} I(x)}.$$

We will define and study the large deviation principle in chapter 4. More formally, a sequence of probability measures μ_n on a topological space $\mathcal{X}$ equipped with the Borel σ-algebra $\mathcal{B}_\mathcal{X}$ is said to satisfy *the large deviation principle with rate* I if

$$-\inf_{x\in A^\circ} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log[\mu_n(A)] \leq \limsup_{n\to\infty} \frac{1}{n} \log[\mu_n(A)] \leq -\inf_{x\in\overline{A}} I(x),$$

holds for every $A \in \mathcal{B}_\mathcal{X}$. The existence of the large deviation principle for the distributions of the CMC estimators, θ_n, is given by Cramér's Theorem (see e.g [25, Theorem 1.2.6]). If the logarithmic moment generating function of X is finite for every $s \in \mathbb{R}$, then the distributions of θ_n satisfy the large deviation principle with rate function I given by the Legendre-Fenchel transform of the logarithmic moment generating function of X. This can be applied to estimate the sample size required to achieve the desired precision in Monte Carlo simulations. If we want the CMC estimator, θ_n, to have relative precision ε with probability $1 - \alpha$, for some $\alpha > 0$, then this is equivalent to

$$\mathbb{P}\Big(|\theta_n - \theta| \geq \varepsilon|\theta| \Big) \leq \alpha.$$

Hence, for large enough n, the required sample size is approximately

$$n \approx \frac{\log(\alpha)}{\inf\big\{I(x) : |x - \theta| \geq \varepsilon|\theta|\big\}}.$$

Another interpretation of Monte Carlo estimators can be done through their empirical distributions, which are random variables defined on Ω and taking values in $\mathbf{M}_1(\mathcal{X})$, the space of probability measures on $\mathcal{X}$. Given a random variable $X : \Omega \to \mathcal{X}$, we can define the map $\delta_X : \Omega \to \mathbf{M}_1(\mathcal{X})$, given by $\omega \mapsto \delta_{X(\omega)}$, where

$$\delta_{X(\omega)}(A) = \begin{cases} 1, & X(\omega) \in A, \\ 0, & X(\omega) \notin A, \end{cases}$$

[1] S.R. Srinivasa Varadhan was awarded the Abel prize in 2007, partly for his contributions to the theory of large deviations.

is the Dirac measure at $X(\omega)$. When $\mathbf{M}_1(\mathcal{X})$ is equipped with a suitable σ-algebra the map δ_X is measurable and hence a random variable which takes values in the space of probability measures on $\mathcal{X}$. The empirical distribution of the CMC estimator is defined as

$$\mathbf{L}_n(\omega) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}.$$

By integrating over $\mathcal{X}$ with respect to the measure $\mathbf{L}_n(\omega)$ one gets the Monte Carlo estimator $\theta_n(\omega)$. From a theoretical perspective it is interesting to study the convergence of the empirical distributions $\mathbf{L}_n$ in the space $\mathbf{M}_1(\mathcal{X})$. Convergence is a topological concept, and there are many different topologies that $\mathbf{M}_1(\mathcal{X})$ can be equipped with. An especially important topology on $\mathbf{M}_1(\mathcal{X})$ in probability theory is the *topology of weak convergence*. This topology is the weak topology on $\mathbf{M}_1(\mathcal{X})$ that is generated by the class of linear functionals on the form

$$\langle f, \mu \rangle := \int_{\mathcal{X}} f \, \mathrm{d}\mu \,, \quad f \in C_b(\mathcal{X}).$$

We will study the topology of weak convergence of measures in detail in Chapter 3 where we also introduce another weak topology on $\mathbf{M}_1(\mathcal{X})$ known as the τ-topology. The τ-topology is stronger than the topology of weak convergence and many large deviations results hold with respect to this topology.

It was shown by Varadarajan in [50] that the empirical distributions $\mathbf{L}_n$ converge weakly to the distribution of X. Since the empirical distributions are random variables with domain Ω and taking values in the space of probability measures on $\mathcal{X}$, one may ask whether the distributions of $\mathbf{L}_n$ satisfy a large deviation principle. Sanov's Theorem (see e.g. [25, Theorem 3.2.17]) states that the distributions of $\mathbf{L}_n$, which are probability measures in $\mathbf{M}_1(\mathbf{M}_1(\mathcal{X}))$, satisfy the large deviation principle with rate function $I(\nu) = \mathbf{R}(\nu|\mu)$, where $\mathbf{R}(\cdot|\cdot)$ denotes the relative entropy[2].

In this thesis we study the space of finite signed measures and the space of probability measures equipped with the topology of weak convergence and the τ-topology. We study the large deviations of empirical distributions in these topologies and show how large deviations results can be applied to analyze the performance of IS estimators for random variables taking values in a Polish space. This leads to a recent result by Hult and Nyquist [33, Theorem 3.1] which proves that the empirical distributions of the importance sampling estimators satisfy a Laplace principle in a subspace of $\mathbf{M}_1(\mathcal{X})$ in the τ-topology. In [33] Hult & Nyquist use the weak convergence approach (see e.g. [30]) to prove their results. We take another approach to derive Sanov's Theorem based on projective systems and discuss how the full large deviation principle for the IS estimators can be derived using this framework.

1.1 Contributions of this Thesis

The main interest of this thesis is the study of large deviation principles and its applications to empirical distributions of Monte Carlo estimators. Let X be a random variable, with distribution μ, taking values in a Polish space $\mathcal{X}$, and let (Y_i) be a sequence of i.i.d. random variables taking values in $\mathcal{X}$, with distribution π. We assume that μ is absolutely continuous with respect to π, and denote this by $\mu \ll \pi$. The empirical distributions of the importance

[2] Also known as Kullback-Leibler divergence and we introduce it in Chapter 3.

sampling estimator is given by

$$\mathbf{I}_n(\omega) := \sum_{i=1}^{n} \rho\big(Y_i(\omega)\big)\delta_{Y_i(\omega)}.$$

In contrast to the empirical distributions of the CMC estimator the map $\mathbf{I}_n$ takes values in the space of nonnegative finite measures $\mathbf{M}_+(\mathcal{X})$. There are two main contributions of this thesis:

1. In Chapter 3 we study the topological and linear structure of the space of finite signed measures equipped with different topologies, concentrating on the topology of weak convergence and the τ-topology. We collect several important results related to measurability and continuity in the τ-topology and the weak convergence topology previously scattered across the research literature. We also give a proof that $\mathbf{I}_n$ converges to μ weakly whenever $\mathcal{X}$ is a separable metrizable space. The following theorem is proved:

   Theorem 3.17. *Let $\mathcal{X}$ be a separable metrizable space and (Y_i) a sequence of i.i.d. random variables taking values in $\mathcal{X}$ with distribution π. If X is another random variable on $\mathcal{X}$ with distribution $\mu \ll \pi$, then the empirical distributions of the IS estimator, $\mathbf{I}_n$, converge weakly to μ almost surely, i.e.*

   $$\mathbb{P}(\{\omega \in \Omega \ : \ \mathbf{I}_n(\omega) \implies \mu\}) = 1.$$

   Here we use $\implies$ to denote weak convergence of measure. This is a known extension of the results by Varadarajan in [50], but a proof is hard to come by.

2. In Chapter 4 we introduce the theory of large deviations necessary to understand and analyze the required sample sizes of the CMC and IS estimator and show how the projective systems approach of de Acosta [3] can be used to prove Sanov's Theorem in the τ-topology. This theory is applied to Monte Carlo estimators in Chapter 4.6.

1.2 Outline

We assume that the reader has background knowledge in measure theory, functional analysis, and point set topology comparable to introductory courses at advanced level. For the sake of completeness we have included some well known results from probability theory, measure theory and functional analysis in the appendix.

The outline of this thesis is as follows:

Chapter 2 review the main definition and results of measure theoretic probability and introduce importance sampling. A reader with a solid background in probability and Monte Carlo methods can skip the first two sections, however we put a large emphasis on absolute continuity of measures and the Radon-Nikodym Theorem which is not standard in the Monte Carlo textbooks. The same goes for our presentation of importance sampling where we work with measures rather than probability density functions. This may seem like an unnecessary abstraction but is the correct framework for proving weak convergence and large deviations results in later chapters.

Chapter 3 is devoted to the study of topologies on the space, $\mathbf{M}(\mathcal{X})$, of positive finite measures and the space of probability measures $\mathbf{M}_1(\mathcal{X})$. The concept of weak

convergence of measures is introduced as convergence in the weak topology induced by integration with respect to $C_b(\mathcal{X})$, and the main results from the theory of weak convergence of measures on separable metric spaces are proved. Two more topologies on are introduced: the τ-topology, and the topology corresponding to the total variation norm. Relative entropy (also known as Kullback-Leibler divergence) is also introduced and the chapter is finished with a section discussing continuity and measurability with respect to the different topologies introduced.

Chapter 4 presents the theory of large deviations. We state and prove two classical results in the theory: Cramér's, and Sanov's Theorems. We then proceed by introducing a recent result of Hult and Nyquist [33] which extend Sanov's Theorem to empirical measures for importance sampling estimators. We end this chapter with a section that goes more into depth of the applications of large deviations to Monte Carlo estimators.

2 Background

In this chapter, we present some background material from probability theory and Monte Carlo methods necessary to follow the developments in this thesis. In section 1, we review some measure theoretic probability and the Radon-Nikodym Theorem. In section 2, we give an introduction to the crude Monte Carlo estimator and its convergence rate, and in section 3, we introduce importance sampling and show how it can be used to reduce the variance of the Monte Carlo estimator and hence speed up Monte Carlo simulations.

In Appendix A.2 we have included some more results from measure theoretic probability which will be used in later chapters. There are many textbooks on probability theory which cover the material presented in the next section, among them [10], [28], and [34]. The Radon-Nikodym Theorem is also covered in most Real Analysis textbooks, see for instance [43], [31], and [32].

2.1 Probability and Notation

The formal setting of our study in this text is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables (i.e. measurable transformations) with domain Ω taking values in a Polish space $\mathcal{X}$ equipped with the Borel σ-algebra $\mathcal{B}_{\mathcal{X}}$. A Polish space is a separable topological space which is metrizable with a complete metric. Polish spaces have enough topological structure to make them useful when working with certain limits of measures. We will discuss this more in section 3.3. The collection of all probability measures on $\mathcal{B}_{\mathcal{X}}$ will be denoted by $\mathbf{M}_1(\mathcal{X})$. We use the capital letters X, Y to denote $\mathcal{X}$ valued random variables with domain Ω. The *distribution* of a random variable $X : \Omega \to \mathcal{X}$ is a probability measure on $\mathcal{B}_{\mathcal{X}}$ defined by

$$\mu(E) := \mathbb{P} \circ X^{-1}(E), \quad \text{for every } E \in \mathcal{B}_{\mathcal{X}}.$$

For a real valued random variable, X, with domain Ω, we use $\mathbb{E}$ to denote the *expected value* of X, which is defined as the integral

$$\mathbb{E}[X] := \int_{\Omega} X \, \mathrm{d}\mathbb{P} \, .$$

The variance of X is defined as

$$\mathbb{V}[X] := \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^2\big] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

The Radon-Nikodym Theorem

Even if $\mathcal{X}$ is different from $\mathbb{R}$ we can construct an integrable real valued random variable by taking a Borel measurable and integrable function $g : \mathcal{X} \to \mathbb{R}$ and composing it with X. The function $g(X)$ is a real valued random variable with domain $(\Omega, \mathcal{F})$, and a standard measure theoretic argument using the monotone convergence theorem and simple approximation shows that

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X) \, \mathrm{d}\mathbb{P} = \int_{\mathcal{X}} g \, \mathrm{d}\mu \, . \tag{4}$$

The right hand side of equation (4) can also be interpreted as the expected value of the random variable g on the probability space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu)$. Whenever we take expected values with respect to some measure other than $\mathbb{P}$, we will clarify this by sub-scripting the expected

value operator $\mathbb{E}$. This notation is standard in the probabilistic community and using this the right hand side of equation (4) can be expressed as

$$\int_{\mathcal{X}} g \, \mathrm{d}\mu = \mathbb{E}_{\mu}[g].$$

These types of change of measure will play an important role in the forthcoming developments. We will now review absolute continuity of measures and the Radon-Nikodym Theorem. This result and its implications are central to understanding both importance sampling and relative entropy which play an important role in the theory of Large Deviations.

Definition 2.1. A measure μ is said to be *absolutely continuous* with respect to a measure π on a σ-algebra $\mathcal{B}$ if $\mu(A) = 0$ whenever $A \in \mathcal{B}$ and $\pi(A) = 0$.

We write $\mu \ll \pi$ to denote that μ is absolutely continuous with respect to π.

Theorem 2.1 (Radon-Nikodym Theorem)**.** *Let μ and π be σ-finite measures on a measurable space $(\mathcal{X}, \mathcal{B})$, and $\mu \ll \pi$. Then there exists a μ-integrable function ρ, such that*

$$\mu(A) = \int_A \rho \, \mathrm{d}\pi \,, \quad \text{for every } A \in \mathcal{B}.$$

The function ρ is unique μ-a.e. and is called the *Radon-Nikodym derivative* of μ with respect to π, this is written as $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}$. The usefulness of the Radon-Nikodym Theorem becomes more apparent in the following two corollaries.

Corollary 2.1 (Radon-Nikodym change of measure)**.** *Let μ and π be σ-finite measures on a measurable space $(\mathcal{X}, \mathcal{B})$, and $\mu \ll \pi$. If $g \in L^1(\mathcal{X}, \mu)$, then*

$$\int_{\mathcal{X}} g \, \mathrm{d}\mu = \int_{\mathcal{X}} g \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \, \mathrm{d}\pi \,.$$

Using the Radon-Nikodym Theorem we define the *probability density function*, f_X of a random variable X with distribution μ with respect to measure $\mathrm{d}x$ as the Radon-Nikodym derivative

$$f = \frac{\mathrm{d}\mu}{\mathrm{d}x}.$$

The continuous distributions known from introductory probability theory courses correspond to real valued random variables with distributions which are absolutely continuous with respect to Lebesgue measure. Similarly the discrete distributions correspond to random variables which takes values in $\mathbb{N}$ and with distributions absolutely continuous with respect to counting measure. Using Corollary 2.1 we get the classical formula for the expected value of a continuous random variable

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, \mathrm{d}\mu = \int_{\mathbb{R}} x f(x) \, \mathrm{d}x \,.$$

The next corollary states two very useful properties of the Radon-Nikodym derivatives.

Corollary 2.2 (Radon-Nikodym derivative properties)**.** *Let ν, μ, and π, be σ-finite measures on a measurable space $(\mathcal{X}, \mathcal{B})$.*

If $\nu \ll \mu \ll \pi$, then the following chain rule holds for the Radon-Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\pi} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \frac{\mathrm{d}\mu}{\mathrm{d}\pi}.$$

Furthermore, if it also holds that $\pi \ll \mu$, then

$$\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\mu} = 1 \text{ a.e.}$$

By rearranging the second part of Corollary 2.2 we get the relation between the two different Radon-Nikodym derivatives

$$\frac{\mathrm{d}\pi}{\mathrm{d}\mu} = \frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} \text{ a.e.} \tag{5}$$

We will revisit this identity when discussing optimal choices of measures in importance sampling.

2.2 Monte Carlo

In this section we give an in introduction to Monte Carlo simulation. The material is well known and there are several classical reference- and textbooks which treat the same material (see e.g. [39], [6], [42], and [44]). Our exposition is inspired by and follows that of Asmussen and Glynn in [6] the most closely. However, we differ from most of the texts mentioned above in our strong focus on working with probability measures instead of probability distribution functions. This abstract approach to the subject will be necessary when working with large deviations theory.

We assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that X is random variable taking values in the measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$. If $g : \mathcal{X} \to \mathbb{R}$ is a Borel measurable and integrable function, then $g(X)$ is a real valued random variable with domain $(\Omega, \mathcal{F})$. Given such a function we are interested in computing expected values of the type

$$\theta(g) := \mathbb{E}[g(X)]. \tag{6}$$

One of the most simple estimators to form for θ is given by

$$\theta_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(X_i). \tag{7}$$

The method to approximate θ by θ_n is often referred to as *vanilla* Monte Carlo or *crude* Monte Carlo (CMC) and we use the two terms interchangeably. We shall refer θ_n as the CMC estimator of $g(X)$. The key motivation for the Monte Carlo estimate is the strong law of large numbers, which asserts that $\theta_n \to \theta$ almost surely as $n \to \infty$. However the strong law of large numbers does not provide us with any insight about the rate of convergence of the CMC estimator θ_n. The Monte Carlo convergence rate can be explored by means of the central limit theorem and confidence intervals. If the random variables $g(X_i)$ are square integrable with finite variance $\sigma^2 = \mathbb{V}[g(X)]$ then it follows from the central limit theorem that

$$\sqrt{n}\big(\theta_n - \theta\big) \implies \mathcal{N}\big(0, \sigma^2\big), \quad \text{as } n \to \infty. \tag{8}$$

Here we use $\implies$ to denote convergence in distribution. Thus it follows from (8) that the estimation error $\theta_n - \theta$ converges to a normal distribution with variance σ^2/n. Hence, for large n the estimator θ_n is approximately $\mathcal{N}(\theta, \sigma^2/n)$-distributed, and consequently we can create a $1-\alpha$ two-sided confidence interval for the CMC estimator by

$$I_\alpha(\theta_n) \approx \left(\theta_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\,,\ \theta_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right), \tag{9}$$

This means that for large enough n

$$\mathbb{P}\left(\theta \in I_\alpha(\theta_n)\right) \gtrapprox 1-\alpha.$$

We use the symbol $\gtrapprox$ to mean that the inequality holds approximately for large n (where n depends on the context). From equation (9) we can deduce that given a fixed confidence level α the absolute error of the Monte Carlo estimate is proportional to to the half width of the confidence interval,

$$HW_\alpha := \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}. \tag{10}$$

Equation (10) captures an important fact about the Monte Carlo method: the error convergence rate is $O(1/\sqrt{n})$. This is often called the *canonical Monte Carlo convergence rate*. Rigorous analysis of the convergence of the CMC estimator can be done using asymptotic confidence intervals. In practice the variance σ^2 is generally unknown and must also be estimated, for more on these topics a good starting point is [6, §III.2].

Performance Of MC Estimators

Generally speaking we want the confidence interval in equation (9) to be as narrow as possible, however a *good* value for the half width HW_α will likely depend on the magnitude of θ. Two of the most simple measures for the error of the Monte Carlo estimators are the

1. *absolute precision* $\varepsilon_a = |\theta_n - \theta|$
2. *relative precision* $\varepsilon_r = \frac{|\theta_n - \theta|}{|\theta|}$

In the statistical literature the terms precision and accuracy are often used for the above entities whilst other areas of applied mathematics usually refer to them as absolute and relative approximation errors. To get the absolute precision less than ε with confidence at least $1-\alpha$ is in mathematical terms equivalent to

$$\mathbb{P}\big(|\theta_n - \theta| < \varepsilon\big) \geq 1-\alpha.$$

The necessary sample size n to achieve this precision can be derived from the confidence interval for the CMC estimator given in equation (9), which yields the formula

$$n \gtrapprox \frac{z_{1-\alpha/2}^2 \sigma^2}{\varepsilon_a^2}. \tag{11}$$

We can use the above formula together with the fact that $\varepsilon_a = |\theta|\varepsilon_r$ to get the expression for the required sample size

$$n \gtrapprox \frac{z_{1-\alpha/2}^2 \sigma^2}{\varepsilon_r^2 \theta^2}. \tag{12}$$

This shows that in order to achieve the same relative precision as absolute precision given by n requires a sample size which is scaled by a factor of $|\theta|^{-2}$. It is clear that this number can grow very large for small θ. Furthermore, the relative precision is the preferred method of the two for evaluating the effectiveness of an estimator when $\theta \ll 1$, which is the case when working with rare events. In practice there is one complication with using the formulas in equation (11) and (12); the variance σ^2 is generally unknown. Furthermore the formula in equation (12) involves the expected value θ, which we are trying to approximate. There are workarounds to this problem and one common solution is to use the sample mean and

variance in combination with sequential algorithms that update the samples every iteration (see e.g. [42]).

Another common measure for the performance of an estimator which is commonly used in statistics is the *Mean Square Error* (MSE). The MSE of the estimator $\hat{\theta}$ of $g(X)$ is defined as

$$MSE(\hat{\theta}) := \mathbb{E}\big[\big(g(X) - \hat{\theta}\big)^2\big].$$

Whenever the estimator $\hat{\theta}$ is unbiased, i.e.

$$\mathbb{E}\big[\hat{\theta}\big] = \mathbb{E}\big[g(X)\big],$$

then the expression for the mean square error simply reduces to the variance of the estimator $\hat{\theta}$. This is the case for the CMC estimator θ_n and will also be true for the importance sampling estimator introduced in the next section. Hence the terms variance and MSE will be used interchangeably for this entity.

Rare Events

The term *rare event* is used in the stochastic simulation literature denote events $A \in \mathcal{F}$ that satisfy $\mathbb{P}(A) \ll 1$. This is usually used to mean probabilities below an order of 10^{-3} or 10^{-4} in magnitude (see e.g. [6] or [37]). Oftentimes, the probability p, of a rare event, is unknown and what the practitioner is seeking to estimate. In that case the CMC estimate for the random variable $\mathbb{1}_{\mathbb{A}}$ can be used to find the probability

$$p := \mathbb{P}(A) = \int_A \mathrm{d}\mathbb{P} = \int_\Omega \mathbb{1}_A \,\mathrm{d}\mathbb{P}\,.$$

The CMC estimator is in this context be given by

$$\theta_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_i \sim \mathrm{Be}(p),$$

where $\mathrm{Be}(p)$ denotes the Bernoulli distribution with parameter p. The variance is given by the expression

$$\mathbb{V}[\mathbb{1}_A] = p(1-p).$$

For rare events the variance and expectation are almost the same, this is clear if we consider the quotient

$$\frac{\mathbb{V}[\mathbb{1}_A]}{\mathbb{E}[\mathbb{1}_A]} = 1 - p.$$

If we consider the factor which depends on p required for a given relative precision given in equation (12), we see that it can be approximated by

$$\frac{\sigma^2}{\theta^2} = \frac{\mathbb{V}[\mathbb{1}_A]}{p^2} = \frac{1-p}{p} \approx \frac{1}{p}, \quad p \ll 1.$$

Thus, for rare events the number of simulations required to get the the relative precision $\varepsilon_r < 1$ with CMC is greater than p^{-1}. When small probabilities are of interest this can lead to huge simulation costs. In [6] they mention that probabilities in telecommunications can be about 10^{-9} in magnitude. If one would like to simulate a probability of that size with a relative precision of 0.1 and confidence level 0.05 it would require more than 10^{11} samples. This goes to show that the CMC method is very inefficient for rare events and in many cases it may not even be possible to get an answer from the CMC method in a reasonable amount of time. We present an example of this below by computing the probability of a tail event for a normal distributed random variable.

Example 1 (Probability of tail event for normal distributed r.v.)**.** Consider a standard normal random variable $Z \sim N(0,1)$ and the event $A = \{Z \geq a\}$. For large $a > 0$ the event A can be considered a rare event and the probability is given by

$$p_a := \mathbb{P}(A) = \mathbb{E}[\mathbb{1}_{\{Z \geq a\}}] = \int_a^\infty f(x)\,\mathrm{d}x\,, \tag{13}$$

where $f(x)$ is the p.d.f. for a standard normally distributed random variable given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

The integral in equation (13) cannot be solved analytically and it is usually expressed as

$$p_a = 1 - \Phi(a),$$

where Φ denotes the cumulative distribution function of a $N(0,1)$ random variable, which can be approximated numerically. Similarly, we get the expression for the variance of Z, which is given by

$$\mathbb{V}[Z] = p_a(1 - p_a) = \big(1 - \Phi(a)\big)\Phi(a).$$

If we let $a = 4$, then the probability $p_a \approx 3.17 \times 10^{-5}$. A natural question to ask is how many samples we need to get a relative precision ε_r with a certain confidence level α. If $\alpha = 0.01$ then $z_{1-\alpha/2} = 2.58$ and if we want a relative precision $\varepsilon_r = 0.1$ equation (12) leads to a sample size

$$n \geq 2.1 \times 10^7.$$

Figure 1 below show the values of HW_α for the CMC estimator of the rare event A for varying sample sizes, and Figure 2 shows the relevant parts where HW_α is close to $\varepsilon_r p_a$.

It is clear from the example above that the simulation costs can become very large, even when simulating the probabilities of rare events under very simple distributions. However, increasing the sample size is not the only option to increase the precision of the simulation. The half width HW_α can also be reduced by reducing the variance σ^2. This is the idea behind importance sampling, which is introduced in the next section. For more on rare event simulation two good starting points are Chapter VI of [6] and Chapter 10 of [37], which provide additional references.

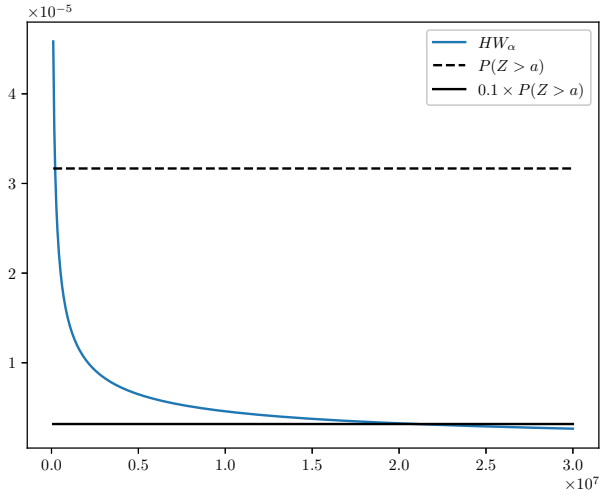

Figure 1: Half width HW_α for the CMC estimator of the rare event A for varying sample sizes.

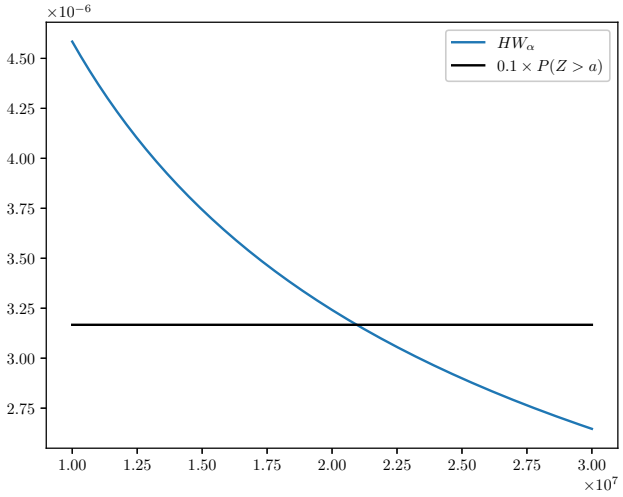

Figure 2: Half width HW_α for the CMC estimator displayed for the rare event A for sample sizes where HW_α is close to $p \times 10^{-1}$.

2.3 Importance Sampling

The convergence rate of vanilla Monte Carlo is $O(1/\sqrt{n})$ which in many cases in unfeasible from a computational stand point, especially when it comes to rare event simulations. There are two main strategies to increase the precision and reduce the MSE of Monte Carlo estimates. Firstly, it is always possible to increase the sample size n, secondly we may use another estimator with lower variance. The general technique of replacing a Monte Carlo

estimator with another estimator that has lower variance is well established, and these methods are commonly referred to as *variance reduction techniques*. Importance sampling solves the convergence rate problem by the latter approach, thus it belongs to the class of variance reduction techniques. The idea is to make a change of measure and sample from another distribution which results in a lower variance of the estimator. Let $\mu, \pi \in \mathbf{M}_1(\mathcal{X})$ satisfy $\mu \ll \pi$, and ρ denote the Radon-Nikodym derivative of μ with respect to π, i.e.

$$\rho = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}.$$

Suppose we are interested in computing θ as in equation (6) and have a random variable Y taking values in $\mathcal{X}$ with domain $(\Omega, \mathcal{F}, \mathbb{P})$, and the distribution, μ, of X, is absolutely continuous with respect to to the distribution, π, of Y. Then it follows from the Radon-Nikodym Theorem (Corollary 2.1) that

$$\theta = \mathbb{E}[g(X)] = \int_{\mathcal{X}} g \, \mathrm{d}\mu = \int_{\mathcal{X}} g\rho \, \mathrm{d}\pi = \mathbb{E}[g(Y)\rho(Y)]. \tag{14}$$

This identity is called the the *importance sampling fundamental identity* [42] and we refer to the measure μ as the *target distribution* and π as the *proposal distribution*[3]. Even though equation (14) shows an equality in expectation between the right and left hand side it is still possible that the variance of $g(Y)\rho(Y)$ is lower than the variance of $g(X)$. That is clear by the following lemma.

Lemma 2.1. *Assume that $g(X)$ has finite variance, then the variance of $g(Y)\rho(Y)$ is given by*

$$\mathbb{V}(g(Y)\rho(Y)) = \int_{\mathcal{X}} g^2\rho \, \mathrm{d}\mu - \theta^2.$$

Proof. The proof is a straight forward application of the Radon-Nikodym Theorem. Let $g(X)$ have finite variance, then a change of measure yields

$$\mathbb{V}[g(Y)\rho(Y)] = \mathbb{E}_\pi[g^2\rho^2] - \mathbb{E}_\pi[g\rho]^2 = \int_{\mathcal{X}} g^2\rho \, \mathrm{d}\mu - \theta^2.$$

■

The implications of Lemma 2.1 are very important. It follows that the variance of $g(Y)\rho(Y)$ may be different from the variance of $g(X)$, which is given by

$$\mathbb{V}[g(X)] = \mathbb{E}_\mu\big[g^2\big] - \mathbb{E}_\mu[g]^2 = \int_{\mathcal{X}} g^2 \, \mathrm{d}\mu - \theta^2,$$

We express this in the corollary below.

Corollary 2.3. *The difference in variance between $g(X)$ and $g(Y)\rho(Y)$ is given by*

$$\mathbb{V}[g(X)] - \mathbb{V}[g(Y)\rho(Y)] = \int_{\mathcal{X}} g^2 \, (1-\rho) \, \mathrm{d}\mu \, .$$

Hence, by using a change of measure, it is possible to preserve the expectation of interest whilst reducing the variance. By Corollary 2.3, the variance is reduced whenever the inequality

$$\int_{\mathcal{X}} g^2(1-\rho) \, \mathrm{d}\mu > 0 \tag{15}$$

[3]The proposal distribution π goes under many names in the literature and it is also commonly called the *sampling distribution* or *importance distribution*.

is satisfied. Furthermore, equation (15) gives some insight into what properties a good sampling distribution should have in terms of the Radon-Nikodym derivative $\rho = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$. Preferably ρ should be small around the areas where g^2 is the largest. In terms of measures this means that π should be concentrated on the areas where g^2 is large. It is not very hard to show which distribution π that is optimal when it comes to reducing the variance, and it is shown in the following theorem.

Theorem 2.2. *Let $Z : \Omega \to \mathcal{X}$ be a random variable with distribution $\sigma \in \mathbf{M}_1(\mathcal{X})$ that satisfies $\mu \ll \sigma$. If*

$$\frac{\mathrm{d}\mu}{\mathrm{d}\sigma} = \frac{\mathbb{E}_\mu(|g|)}{|g|}, \tag{16}$$

then σ is the optimal measure in the sense that it yields the lowest variance of all IS-measures. I.e. if $Y : \Omega \to \mathcal{X}$ is a random variable with distribution $\pi \in \mathbf{M}_1(\mathcal{X})$ and $\mu \ll \pi$, then

$$\mathbb{V}\left[g(Z)\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}(Z)\right] \leq \mathbb{V}\left[g(Y)\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(Y)\right].$$

Proof. Let μ, π and σ satisfy the assumptions of Theorem 2.2. By Lemma 2.1 it follows that

$$\mathbb{V}\left[g(Z)\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}(Z)\right] - \mathbb{V}\left[g(Y)\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(Y)\right] = \int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\right)^2 \mathrm{d}\sigma - \int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi$$

Thus, we need to show that

$$\int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\right)^2 \mathrm{d}\sigma \leq \int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi\,.$$

Using the expression for the Radon-Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}$ we get that

$$\int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\right)^2 \mathrm{d}\sigma = \int_{\mathcal{X}} g^2 \left(\frac{\mathbb{E}_\mu(|g|)}{|g|}\right)^2 \mathrm{d}\sigma = \mathbb{E}_\mu(|g|)^2 \tag{17}$$

By a change of measure and then applying the Cauchy-Schwarz inequality we get that

$$\mathbb{E}_\mu(|g|)^2 = \left(\int_{\mathcal{X}} |g| \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi\right)^2 \leq \int_{\mathcal{X}} g^2 \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi \tag{18}$$

Thus, by combining equation (17) and (18) the inequality follows.

■

This *change of measure technique* is the main idea behind the importance sampling estimate of θ. Instead of using the target distribution μ importance sampling works by sampling from the proposal distribution π. The only difference between the crude Monte Carlo estimator and the importance sampling estimator is a change of measure which makes it possible to simulate random variables with law π instead of μ.

Definition 2.2. Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables with domain $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\mathcal{B}_\mathcal{X}$. Assume that the law π, of Y_i, is absolutely continuous w.r.t. the law of X, μ. Then the *importance sampling (IS) estimator* of θ is defined as

$$I_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(Y_i)\rho(Y_i), \quad \rho = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}. \tag{19}$$

Lemma 2.2. *The variance of the IS estimator I_n is given by*

$$\mathbb{V}(I_n) = \frac{1}{n}\left(\int_{\mathcal{X}} g^2\rho \,\mathrm{d}\mu - \theta^2\right) = \frac{1}{n}\left(\int_{\mathcal{X}} g^2\rho^2 \,\mathrm{d}\pi - \theta^2\right).$$

Proof. The random variables $g(Y_i)\rho(Y_i)$ are i.i.d. real valued random variables. It follows from independence that

$$\mathbb{V}[I_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[g(Y_i)\rho(Y_i)] = \frac{1}{n}\mathbb{V}[g(Y_i)\rho(Y_i)].$$

Thus, Lemma 2.1 implies that

$$\mathbb{V}[I_n] = \frac{1}{n}\left(\int_{\mathcal{X}} g^2\rho^2 \,\mathrm{d}\pi - \theta^2\right) = \frac{1}{n}\left(\int_{\mathcal{X}} g^2\rho \,\mathrm{d}\mu - \theta^2\right).$$

■

The expression for the variance of the IS estimator given by Lemma 2.2 is simply the expression of the variance for $g(Y)\rho(Y)$ scaled by $1/n$. Hence, the optimal proposal distribution is given by the measure satisfying equation (16) from Theorem 2.2. However, in practice equation (16) is not very helpful for choosing the proposal distribution π since it involves $\theta = \mathbb{E}_\mu[g]$ which is the unknown value we are trying to compute.

2.4 Autonormalised Importance Sampling

It is possible to extend the IS-algorithm for cases when the Radon-Nikodym derivative, ρ, is known only up to a multiplicative constant. We say that a function $w : \mathcal{X} \to \mathbb{R}$ is an *un-normalised* Radon-Nikodym derivative of μ with respect to π if it satisfies

$$w(x) = \alpha\rho(x) \tag{20}$$

for some positive constant α. Note, since μ is a probability measure we get that

$$\mu(\mathcal{X}) = \int_{\mathcal{X}} \rho \,\mathrm{d}\pi = \frac{1}{\alpha}\int_{\mathcal{X}} w \,\mathrm{d}\pi = 1.$$

The normalising constant α can be found by integration, which yields $\alpha = \int_{\mathcal{X}} w\mathrm{d}\pi$. Hence, we can express θ as

$$\int_{\mathcal{X}} g \,\mathrm{d}\mu = \frac{\int_{\mathcal{X}} g w \,\mathrm{d}\pi}{\int_{\mathcal{X}} w \,\mathrm{d}\pi}.$$

This is the fundamental identity of the *autonormalised IS estimator* J_n which is defined as

$$J_n(g) := \frac{\sum_{i=1}^{n} g(Y_i)w(Y_i)}{\sum_{i=1}^{n} w(Y_i)} \tag{21}$$

It is important to note that the autonormalised IS estimator is biased. This, may not be apparent from equation (21) at first glance, however, the autonormalised IS estimator is a quotient of random variables and hence it is not possible to simply use linearity of expectation to guarantee that the expected value is preserved for the estimator. This can be compared with the regular IS estimator which is unbiased. The autonormalised IS estimator plays an important role in many advanced Monte Carlo schemes due to the fact that it is sufficient to know the distributions only up to a normalising constant. These type of situations naturally occur in Bayesian filtering problems which are commonly solved with Sequential Monte Carlo methods, also known as particle filters. For more on these topics the reader is referred to the books [20], [27], and [47].

2.5 Importance Functions

When using importance sampling, the assumption that $\mu \ll \pi$ on all of $\mathcal{X}$ may seem unnecessarily restrictive. If we are interested in computing $\mathbb{E}[g(X)]$, then it should suffice that $\mu \ll \pi$ on the region $\{g \neq 0\}$, since

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g \, \mathrm{d}\mu = \int_{g \neq 0} g \, \mathrm{d}\mu \, .$$

We can extend this idea even further by defining an *importance function* $f : \mathcal{X} \to [0, \infty)$ which is specifically designed to capture the importance of different regions of $\mathcal{X}$. The idea is that we want the measure π to be a good approximation on μ in the regions of greatest interest. When computing $\mathbb{E}[g(X)]$ a suitable importance function could be the indicator function $\mathbb{1}_{\{g \neq 0\}}$.

Using the idea of an importance function we can restrict the criteria that $\mu \ll \pi$ to a subset of $\mathcal{X}$ rather than on all of $\mathcal{X}$. If $F \subset \mathcal{X}$, then the restriction of the σ-algebra $\mathcal{B}$ to F is given by

$$\mathcal{B}_F = \{A \cap F : A \in \mathcal{B}\},$$

and is a $\sigma-$algebra of subsets of F. Let μ_F and π_F denote the restrictions of μ and π to $\mathcal{B}_F$. If it holds that $\mu_F \ll \pi_F$ on $\mathcal{B}_F$, then it follows from the Radon-Nikodym Theorem that the Radon-Nikodym derivative

$$\rho_F = \frac{\mathrm{d}\mu_F}{\mathrm{d}\pi_F}$$

exists, and that

$$\int_F h \, \mathrm{d}\mu_F = \int_F h \rho_F \, \mathrm{d}\pi_F$$

whenever h is a μ_F-integrable function. Hence, since $\mathbb{1}_F$ is integrable, it follows that if $\mu \ll \pi$ on all of $\mathcal{X}$ then ρ_F is simply the restriction of the Radon-Nikodym derivative ρ to F. Given an importance function, f, we define $\rho_f : \mathcal{X} \to [0, \infty)$ by

$$\rho_f := \begin{cases} \frac{\mathrm{d}\mu_{\{f \neq 0\}}}{\mathrm{d}\pi_{\{f \neq 0\}}}, & f \neq 0 \\ 0, & f = 0, \end{cases}$$

and let μ_f denote the measure on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ given by

$$\mu_f(A) := \int_A f \, \mathrm{d}\mu.$$

Then it follows that

$$\int_{\mathcal{X}} g f \, \mathrm{d}\mu = \int_{\mathcal{X}} g \, \mathrm{d}\mu_f = \int_{\mathcal{X}} g \rho_f \, \mathrm{d}\pi \, , \tag{22}$$

for every integrable g that satisfies $\mathrm{supp}(g) \subset \mathrm{supp}(f)$. If $f = \mathbb{1}_{\{g \neq 0\}}$, then equation (22) reduces to the importance case

$$\int_{\mathcal{X}} g \, \mathrm{d}\mu_f = \int_{\mathcal{X}} g \rho_g \, \mathrm{d}\pi \, .$$

The main idea behind the importance function is that we do not need the proposal distribution to satisfy $\mu \ll \pi$ on all of $\mathcal{X}$, but only the important regions. Hence, we may consider a larger class of admissible proposal distributions. We will not take this idea further, but it is good to have this in mind when designing an IS estimator.

3 Topology of Measures and Relative Entropy

In this chapter we study topologies on the space $\mathbf{M}(\mathcal{X})$ of finite signed measures on $\left(\mathcal{X}, \mathcal{B}_{\mathcal{X}}\right)$ and some of its subspaces. The two subspaces of most interest to us are the space nonnegative measures, $\mathbf{M}_+(\mathcal{X})$, and the space of probability measures $\mathbf{M}_1(\mathcal{X})$. We introduce three topologies: the total variation norm, the $\tau-$topology, and the topology of weak convergence. The term *weak convergence* of measures is very natural from a functional analytic perspective, as we will see the corresponding topology is the weak topology on $\mathbf{M}(\mathcal{X})$ generated by the space of bounded continuous functions on $\mathcal{X}$. The general idea will be that the space of bounded measurable functions and $\mathbf{M}(\mathcal{X})$ can be paired by the dual relation

$$\langle f, \mu \rangle = \int_{\mathcal{X}} f \, \mathrm{d}\mu \,,$$

and $\mathbf{M}(\mathcal{X})$ is a subset of the dual to the space of bounded measurable functions. When the functions f are taken to be bounded measurable functions, then the strong topology induced by this dual pairing is the total variation norm and the weak* topology is the τ-topology. In the final two sections of this chapter we study the measurability and continuity properties with respect to the τ-topology and topology of weak convergence on $\mathbf{M}(\mathcal{X})$. We also introduce the relative entropy which can be thought of as a distance between to probability measures and play an important role as a rate function in the large deviations theory presented in chapter 4.

3.1 Spaces of measures

In this section we study the topological and linear structure of different spaces of measures. We start with the basic definitions and by introducing some notation. Let $\mathcal{X}$ be a set, then we use $\mathcal{B}$ to denote a σ-algebra of subsets of $\mathcal{X}$ and $\mathcal{A}$ to denote an algebra of subsets of $\mathcal{X}$. If $\mathcal{X}$ is a topological space, then we will use $\mathcal{B}_{\mathcal{X}}$ to denote the Borel σ-algebra of subsets of $\mathcal{X}$ and similarly $\mathcal{A}_{\mathcal{X}}$ to denote the Borel algebra of subsets of $\mathcal{X}$, i.e. the smallest $\sigma-$algebra and algebra of subsets of $\mathcal{X}$ containing all open sets.

Definition 3.1. Let $\mathcal{A}$ be an algebra of subsets of $\mathcal{X}$, then a function $\mu : \mathcal{A} \to [-\infty, \infty]$ is said to be a *finitely additive signed measure*[4] if

1. $\mu(\emptyset) = 0$.
2. μ is finitely additive
3. μ takes at most one of the values $-\infty/\infty$.

Finite additivity means that whenever $\{A_i\}$ is a finite disjoint collection of elements of $\mathcal{A}$, then

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i).$$

It is clear that countable additivity implies finite additivity, hence every signed measure is a finitely additive signed measure. A nonnegative finitely additive signed measure will simply

[4] Some authors refer to finitely additive measures as charges, see e.g. Aliprantis and Border [5].

be called a finitely additive measure. The *total variation* of a finitely additive signed measure is the finitely additive measure $|\mu|$ defined by

$$|\mu|(A) := \sup\left\{\sum |\mu(A_i)| \ : \{A_i\} \text{ is a finite disjoint partition of } A \text{ of elements of } \mathcal{A}\right\}.$$

If $|\mu(X)| < \infty$ we say that μ is of *bounded variation* or simply *finite*. The Hahn-Jordan Decomposition Theorem (see e.g. [29, §III.1.8]) states that any finitely additive finite signed measure μ can be decomposed into a positive and negative part μ^+ and μ^- that satisfy

$$\mu(A) = \mu^+(A) - \mu^-(A), \quad |\mu|(A) = \mu^+(A) + \mu^-(A),$$

for every $A \in \mathcal{A}$. Explicitly, the expression for the positive and negative part of the Jordan decomposition is given by

$$\mu^+(A) := \sup_{U \subset A} \mu(U), \quad \mu^-(A) := \inf_{A \subset F} \mu(F),$$

where the supremum and infimum are taken over sets $U, F \in \mathcal{A}$. Given a measurable space $(\mathcal{X}, \mathcal{B})$ we use the notation $\mathbf{M}(\mathcal{X})$ for the collection of finite signed measures on $\mathcal{X}$. Similarly, if $\mathcal{A}$ is an algebra of subsets of $\mathcal{X}$ we write $\mathbf{M}^{ba}(\mathcal{X})$ for the collection of all finite finitely additive signed measures on $(\mathcal{X}, \mathcal{A})$. Since every measure is finitely additive, it follows that if $\mathcal{A} = \mathcal{B}$, then we get the following inclusion

$$\mathbf{M}_1(\mathcal{X}) \subset \mathbf{M}(\mathcal{X}) \subset \mathbf{M}^{ba}(\mathcal{X}).$$

Regular Measures

An important class of measures on topological space are the regular[5] measures.

Definition 3.2. A finitely additive measure, μ, defined on a topological space is said to be *regular* if for every $A \in \mathcal{A}$, μ satisfy

$$\begin{aligned}\mu(A) &= \sup\{\mu(C) \ : \ C \subset A, \ C \text{ closed}\} \\ &= \inf\{\mu(U) \ : \ A \subset U, \ U \text{ open}\}.\end{aligned}$$

We write $\mathbf{M}^r(\mathcal{X})$ to denote the collection of Borel regular measures on $\left(\mathcal{X}, \mathcal{B}_X\right)$ and $\mathbf{M}^{rba}(\mathcal{X})$ for the collection of finitely additive signed regular measures on $\left(\mathcal{X}, \mathcal{A}_\mathcal{X}\right)$. Note that $\mathbf{M}^{rba}(\mathcal{X})$ *is defined on the Borel algebra and not the Borel σ-algebra*, thus $\mathbf{M}^{rba}(\mathcal{X}) \not\subset \mathbf{M}(\mathcal{X})$. The relations between these spaces is non-trivial, however in some cases every element of $\mathbf{M}^{rba}(\mathcal{X})$ can be uniquely extended to an element of $\mathbf{M}^r(\mathcal{X})$ (see e.g. [5, section 14.4] note that their notation is different from ours).

One of the main properties of regular measures, which follows directly from the definition, is that they are completely determined by their values on open (respectively closed) sets. and this is the distinguishing feature which make them very useful when working with convergence. Another equivalent characterization of regular measures which also follows directly from the definition is given below.

[5] There are several definitions of a regular measure existing in the measure theoretic literature and we define a Borel regular measure the same way as Partasarathy [41], Bogachvev [13], and Dunford & Schwartz [29]. Another definition, sometimes found in real analysis and geometric measure theory texts, defines a regular measure to be what we call a regular and tight measure. See for instance Royden [43] and Mattila [40].

Lemma 3.1. *Let $\mathcal{X}$ be a topological space $\mathcal{X}$ and μ be a Borel measure on $\mathcal{X}$. Then μ is regular if and only if for every $A \in \mathcal{B}_{\mathcal{X}}$ and $\varepsilon > 0$ there exists an open set U_ε and a closed set C_ε such that $C_\varepsilon \subset A \subset U_\varepsilon$ and*

$$\mu\big(U_\varepsilon \setminus C_\varepsilon\big) < \varepsilon.$$

A useful consequence Lemma 3.1 is that whenever a measure is regular on a topological space and $A \in \mathcal{A}$ there exists an increasing sequence of closed subsets (C_n) of A such that

$$\lim_{n\to\infty} \mu(C_n) = \mu(A),$$

and a decreasing sequence of open sets U_n all containing A such that

$$\lim_{n\to\infty} \mu(U_n) = \mu(A).$$

It is useful to know that in general topological spaces a Borel measure may not be regular but on metric spaces every Borel measure is regular and we will study measures on metric spaces more in detail towards the end of next section.

Representation Theorems

The collection of signed measures has a natural linear structure. The addition of two measures μ, ν, and scalar multiplication with $\alpha \in \mathbb{R}$ are given by

$$(\mu + \nu)(A) := \mu(A) + \nu(A), \quad (\alpha\mu)(A) := \alpha\mu(A),$$

It is clear that $\mathbf{M}^{ba}(\mathcal{X})$ and $\mathbf{M}^{rba}(\mathcal{X})$ are linear spaces under these operations and that $\mathbf{M}(\mathcal{X})$ is a linear subspace of $\mathbf{M}^{ba}(\mathcal{X})$. The collection $\mathbf{M}_1(\mathcal{X})$ on the other hand is not a linear subspace since the sum of two probability measure is not a probability measure. The total variation induces a norm on these linear spaces called the *total variation norm*, given by

$$\|\mu\|_{TV} := |\mu|(\mathcal{X}) = \mu^+(\mathcal{X}) + \mu^-(\mathcal{X}).$$

Under this norm the spaces $\mathbf{M}^{ba}(\mathcal{X})$, $\mathbf{M}^{rba}(\mathcal{X})$, and $\mathbf{M}(\mathcal{X})$ are Banach spaces[6] (see e.g [29, pp. 240, IV.2.15–16]). We are now going to state some well known representation theorems for $\mathbf{M}^{ba}(\mathcal{X})$ and $\mathbf{M}^{rba}(\mathcal{X})$. The advantage of having these representations is that we are able to use functional analytic methods to place much weaker (weak*) topologies on these spaces.

Let $B(\mathcal{X})$ denote the space of real valued functions with domain $\mathcal{X}$ which are bounded and Borel measurable. The space $B(\mathcal{X})$ equipped with the uniform norm

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|,$$

is a Banach space and the strong dual space $B^*(\mathcal{X})$ can be represented by $\mathbf{M}^{ba}(\mathcal{X})$ equipped with the total variation norm.

Theorem 3.1 (see e.g. [29, §IV.5.1])**.** *Let $\mathcal{A}$ be an algebra of subsets of $\mathcal{X}$, then the space $\mathbf{M}^{ba}(\mathcal{X}, \mathcal{A})$ with the total variation norm is linearly isometric to $B^*(\mathcal{X})$ in the strong dual topology, by the map*

$$\mu \mapsto \int_{\mathcal{X}} f \, \mathrm{d}\mu\,, \quad \text{for every } f \in B(\mathcal{X}).$$

[6] In fact they are all Banach lattices, specifically AL-spaces (see e.g. [5, pp. 10.10–11 & 12.2]).

Remark. This holds for any algebra of subsets of $\mathcal{X}$, thus it also holds for any σ-algebra of subsets of $\mathcal{X}$. An important case is when $\mathcal{B}$ is a σ-algebra and $\mathbf{M}(\mathcal{X}, \mathcal{B}) \subset \mathbf{M}^{ba}(\mathcal{X}, \mathcal{B})$.

Using the fact that $\mathbf{M}^{ba}(\mathcal{X})$ is isomorphic to the strong dual of $B(\mathcal{X})$, we get an equivalent norm to the total variation norm induced by the strong operator topology, given by

$$\|\mu\| = \sup_{\substack{f \in B(\mathcal{X}) \\ \|f\|_\infty \leq 1}} |\langle f, \mu \rangle|$$

Definition 3.3. Let $\mathcal{X}$ be a topological space, then we define $C_b(\mathcal{X})$ to be the space of bounded continuous real valued function on $\mathcal{X}$, equipped with the uniform norm

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

The space $C_b(\mathcal{X})$ is a Banach Space (see e.g. [29, §IV.6]) , and when $\mathcal{X}$ is normal the dual space of $C_b(\mathcal{X})$ is isomorphic to $\mathbf{M}^{rba}(\mathcal{X}, \mathcal{A}_\mathcal{X})$. But, before we state the exact result we give the following useful result which we will use later.

Theorem 3.2 (see e.g. [5, Thm 14.8])**.** *Let $\mathcal{X}$ be a normal topological space and Λ a positive linear functional on $C_b(\mathcal{X})$, then there exists a unique element $\mu \in \mathbf{M}^{rba}(\mathcal{X}, \mathcal{A}_\mathcal{X})$ that satisfy $\mu(\mathcal{X}) = \Lambda(1)$ and*

$$\int_\mathcal{X} f \, \mathrm{d}\mu = \Lambda(\mu), \quad \text{for every } f \in C_b(\mathcal{X}).$$

Theorem 3.3 (see e.g. [29, §IV.6.2])**.** *Let $\mathcal{X}$ be a normal topological space, then the space $\mathbf{M}^{rba}(\mathcal{X}, \mathcal{A}_\mathcal{X})$ with the total variation norm is linearly isometric to $C_b(\mathcal{X})$ in the strong dual topology, by the map*

$$\mu \mapsto \int_\mathcal{X} f \, \mathrm{d}\mu \,, \quad \text{for every } f \in C_b(\mathcal{X}).$$

In the special case when $\mathcal{X}$ is a compact Hausdorff space the dual is isomorphic to the regular Borel measures defined on the Borel σ-algebra. The following result is generally known as Riesz Representation Theorem (see e.g [29, §IV.6.3]).

Theorem 3.4 (Riesz Representation Theorem)**.** *Let $\mathcal{X}$ be a compact Hausdorff space, then the space of all regular signed Borel measures on $\mathcal{X}$ equipped with the total variation norm is linearly isometric to $C_b(\mathcal{X})$ in the strong dual topology, by the map*

$$\mu \mapsto \int_\mathcal{X} f \, \mathrm{d}\mu \,, \quad \text{for every } f \in C_b(\mathcal{X}).$$

An important example of this is when $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is a compact metric space, then all Borel measures are regular and $\mathbf{M}(\mathcal{X})$ is linearly isometric to the dual space of $C_b(\mathcal{X})$.

The total variation norm

We have already seen that that the total variation norm makes $\mathbf{M}^{ba}(\mathcal{X}, \mathcal{B}_\mathcal{X})$ into a Banach space and it is linearly isometric to the strong dual of $B(\mathcal{X})$ by Theorem 3.1. The collection of finite signed measures, $\mathbf{M}(\mathcal{X})$, is a closed linear subspace and hence also a Banach space in this topology (see e.g. [13, Theorem 4.6.1]).

However, the topology generated by the total variation norm on $\mathbf{M}(\mathcal{X})$ is too strong for most probabilistic applications. An example, which highlights this fact is the following. Consider the Dirac measures δ_x defined by

$$\delta_x(A) := \begin{cases} 1, x \in A, \\ 0, x \notin A. \end{cases}$$

These are elements of $\mathbf{M}(\mathcal{X})$, however, if (x_α) is a sequence of distinct elements in $\mathcal{X}$, then the sequence of Dirac measures (δ_{x_α}) never converges to δ_x, even if $\mathcal{X}$ is a metric space and (x_α) converges to x. Thus the map $x \mapsto \delta_x$ is not continuous from $\mathcal{X}$ to $\mathbf{M}(\mathcal{X})$ equipped with the the total variation norm. Especially, the empirical distributions which we are interested in are not continuous in the total variation norm. Generally, the total variation norm is a too strong topology on $\mathbf{M}(\mathcal{X})$ and its subspaces for probabilistic purposes, which motivates the τ-topology that is a much weaker topology.

The τ-topology

In this section we assume that $\mathcal{X}$ is a measurable space with $\sigma-$algebra $\mathcal{B}$ and that all spaces of measures are taken over $(\mathcal{X}, \mathcal{B})$. Using the dual correspondence between $B(\mathcal{X})$ and $\mathbf{M}^{ba}(\mathcal{X})$ given by Theorem 3.1 it is possible to equip $\mathbf{M}^{ba}(\mathcal{X})$ with the weak* topology induced by $B(\mathcal{X})$.

Definition 3.4. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, then the topology on $\mathbf{M}(\mathcal{X})$ inherited from $\mathbf{M}^{ba}(\mathcal{X})$ equipped with the weak* topology is called the *τ-topology* [7].

Thus, the τ-topology is the weakest topology for which the evaluation maps

$$\mu \longmapsto \int_{\mathcal{X}} f \, \mathrm{d}\mu$$

are continuous for every $f \in B(\mathcal{X})$. The definition and basic properties of weak topologies are given in Appendix A.4. It follows from the definition of the weak* topology that a neighborhood basis of $\mu \in \mathbf{M}(\mathcal{X})$ in the $\tau-$topology is given by

$$\left\{ \nu \in \mathbf{M}(\mathcal{X}) \ : \ \left| \int_{\mathcal{X}} f_i \, \mathrm{d}\mu - \int_{\mathcal{X}} f_i \, \mathrm{d}\nu \right| < \varepsilon, \quad f_1, \ldots, f_n \in B(\mathcal{X}), \quad \varepsilon > 0 \right\}.$$

The τ-topology is much weaker than the total variation norm; however, in general the maps $x \mapsto \delta_x$ are not continuous with respect to the τ-topology. This can be seen by noting that if a net $x_\alpha \to x$ in $\mathcal{X}$ then $\delta_{x_\alpha} \to \delta_x$ in the τ-topology if and only if $f(x_\alpha) \to f(x)$ for every $f \in B(\mathcal{X})$. However, every function satisfying this convergence criteria is continuous. Thus the map $x \mapsto \delta_x$ is τ-continuous if and only if $B(\mathcal{X}) \subset C_b(\mathcal{X})$. Furthermore, the τ-topology is in general not metrizable and nor are the subspaces of greatest interest to us in this topology: $\mathbf{M}(\mathcal{X}), \mathbf{M}_+(\mathcal{X}), \mathbf{M}_1(\mathcal{X})$. In fact $\mathbf{M}_1(\mathcal{X})$ is not metrizable in the τ-topology if there exists an element $\mu \in \mathbf{M}_1(\mathcal{X})$ that satisfies $\mu(x) = 0$ for every $x \in \mathcal{X}$. A proof of this fact is outlined in exercise 9.1.15 of [48].

An interesting observation is that the only topological information about $\mathcal{X}$ which is transferred to $\mathbf{M}_+(\mathcal{X})$ in the τ-topology is the one contained in the σ-algebra $\mathcal{B}_{\mathcal{X}}$. However,

[7] Some authors refer to this topology on $\mathbf{M}(\mathcal{X})$ as the strong topology which can be misleading as it is not the topology inherited from the strong dual of $B^*(\mathcal{X})$, i.e. $\mathbf{M}^{ba}(\mathcal{X}, \mathcal{B})$, with the total variation norm.

the spaces we study will in general be metrizable, hence contain a lot of topological structure. Continuous functions provide a lot more topological information than the bounded measurable functions. Furthermore if $x_\alpha \to x$ then

$$\int_{\mathcal{X}} f \, \mathrm{d}\delta_{x_\alpha} = f(x_\alpha) \to f(x) = \int_{\mathcal{X}} f \, \mathrm{d}\delta_x \,, \quad \text{for every } f \in C_b(\mathcal{X}),$$

which motivates using the maps $\langle f, \cdot \rangle$, where $f \in C_b(\mathcal{X})$, to induce a weak topology on $\mathbf{M}(\mathcal{X})$ instead.

The topology of weak convergence

In this section we place a topology on $\mathbf{M}(\mathcal{X}, \mathcal{B}_\mathcal{X})$ by noting that for every $\mu \in \mathbf{M}(\mathcal{X})$ and $C_b^*(\mathcal{X})$ the map

$$\langle f, \mu \rangle := \int_{\mathcal{X}} f \, \mathrm{d}\mu \,, \quad \text{for every } f \in C_b(\mathcal{X}) \tag{23}$$

defines a dual pairing. However, the representation of $C_b^*(\mathcal{X})$ given in Theorem 3.3 involves finitely additive measures defined on the Borel algebra of $\mathcal{X}$. Therefore, we cannot simply identify $\mathbf{M}(\mathcal{X})$ with a subspace of $\mathbf{M}^{rba}(\mathcal{X})$. [8] We will assume that $\mathcal{X}$ is a metric space, then all Borel measures are regular. Regular measures are uniquely determined by their values on closed sets and on normal spaces Urysohn's Lemma can be applied to show that the maps

$$\mu \mapsto \int_{\mathcal{X}} f \, \mathrm{d}\mu \,.$$

are separating on $\mathbf{M}^r(\mathcal{X})$. Hence, the dual pairing above induces a weak topology on $\mathbf{M}(\mathcal{X})$.

Definition 3.5. Let $\mathcal{X}$ be a metric space, then the weak topology generated by the dual pairing (23) is called the *topology of weak convergence* or simply the *weak topology* on $\mathbf{M}(\mathcal{X})$.

A neighbourhood basis for $\mu \in \mathbf{M}(\mathcal{X})$ in this topology is given by

$$\left\{ \nu \in \mathbf{M}(\mathcal{X}) \;:\; \left| \int_{\mathcal{X}} f_i \, \mathrm{d}\mu - \int_{\mathcal{X}} f_i \, \mathrm{d}\nu \right| < \varepsilon, \quad f_1, \ldots, f_n \in C_b(\mathcal{X}), \quad \varepsilon > 0 \right\}$$

It follows from the properties of weak topologies (see Theorem A.7) that a net (μ_α) converges to μ in this topology iff $\int f \, \mathrm{d}\mu_\alpha \to \int_{\mathcal{X}} f \, \mathrm{d}\mu$ for every $f \in C_b(\mathcal{X})$. This type of convergence is called *weak convergence of measure*. Weak convergence in metric spaces are studied more in detail in section 3.3. We shall make some quick remarks regarding closed subspaces of $\mathbf{M}(\mathcal{X})$. The subspaces of greatest interest to us are $\mathbf{M}_1(\mathcal{X})$, $\mathbf{M}_{\leq 1}(\mathcal{X})$, and $\mathbf{M}_+(\mathcal{X})$, which are closed subspaces of $\mathbf{M}(\mathcal{X})$ in the topology of weak convergence. This is easily seen since the constant function $1 \in C_b(\mathcal{X})$, and therefore if a net (μ_α) converges to μ

$$\mu_\alpha(\mathcal{X}) \to \mu(\mathcal{X}),$$

which shows that $\mathbf{M}_1(\mathcal{X})$ is closed under limits. Borel regularity of the measures ensure that the limit of nonnegative measures are nonnegative and the subspaces $\mathbf{M}_{\leq 1}(\mathcal{X})$ and $\mathbf{M}_+(\mathcal{X})$ are also closed. Many important topological properties such as separability and completeness are inherited for $\mathbf{M}_+(\mathcal{X})$ in the the topology of weak convergence. In Section 3.3, we prove that $\mathbf{M}_+(\mathcal{X})$ is a Polish space if and only if $\mathcal{X}$ is a Polish space. The next section

[8] In the special case when $\mathcal{X}$ is a compact metric space, then Riesz Representation Theorem states that $\mathbf{M}(\mathcal{X})$ is linearly isomorphic to the dual of $C_b(\mathcal{X})$.

covers the theory of Borel measures and continuous functions on metrizable spaces. The theory presented will be essential to our study of weak convergence on metrizable spaces in Section 3.3.

3.2 Metrizable Spaces

This section reviews some of the topological properties of of metrizable spaces. The aim is to introduce the topological results necessary for studying weak convergence of measures on metrizable spaces. A topological space $(\mathcal{X}, \mathcal{T})$ is said to be *metrizable* if there exists a metric d such that the metric space $(\mathcal{X}, d)$ is homeomorphic to $(\mathcal{X}, \mathcal{T})$. The notion of equivalent metrics has different different definitions in the literature and in order to avoid confusion we say that two metrics are *topologically equivalent* if they generate the same topology. There are many topologically equivalent metrics on any metrizable space and given a metric d on $\mathcal{X}$ it is always possible to define an equivalent bounded metric given by

$$d_b = \frac{d}{1+d}.$$

A special class of metrizable spaces which are of great interest in probability theory are the Polish spaces.

Definition 3.6. A topological space $(\mathcal{X}, \mathcal{T})$ is said to be a *Polish space* if it is separable and metrizable with a complete metric.

Remark. A metric that is topologically equivalent to a complete metric may not be complete, an example of this is given below.

Example 2 (Equivalent complete and non-complete metrics)**.** Consider the set $(-1, 1)$ in the subspace topology inherited from $\mathbb{R}$ with the standard euclidean topology. Then, the euclidean distance induces a metric on $(-1, 1)$ which is not complete, since it is possible to construct Cauchy sequences converging to the limit points -1 and 1. On the other hand, a topologically equivalent complete metric on $(-1, 1)$ is given by

$$d(x, y) = \frac{|x - y|}{1 - |x - y|^2}.$$

Whilst completeness depends on the choice of metric, separability is a topological property which is preserved by homeomorphisms, and therefore not affected by the choice of metric. It is well known that a metric space is compact if and only if it is complete and totally bounded (see e.g. [5, Theorem 3.28]). Furthermore, as a consequence of Urysohn's metrization theorem, on metrizable spaces separability is equivalent to the existence of a totally bounded metric.

Theorem 3.5. *A metrizable space, $\mathcal{X}$, is separable if and only if $\mathcal{X}$ admits a totally bounded metric.*

Proof. Let $\mathcal{X}$ be a separable metrizable space. Then by Urysohn's metrization theorem (see e.g. [5, Theorem 3.40]) there exists an isometric embedding φ of $\mathcal{X}$ into the Hilbert Cube $\mathbf{H} = [0, 1]^{\mathbb{N}}$. The unit interval $[0, 1]$ is a compact metric space with the usual euclidean distance metric and Tychonoff's Theorem implies that the Hilbert cube is compact. Furthermore, since $\mathbf{H}$ is a countable product of metric spaces it is metrizable (see e.g. [5, Thm

3.36]) and a metric on $\mathbf{H}$ is given by

$$d_{\mathbf{H}}(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{|x_n - y_n|}{1 + |x_n - y_n|}, \quad \mathbf{x}, \mathbf{y} \in \mathbf{H}.$$

The space $\mathbf{H}$ is compact, thus $d_{\mathbf{H}}$ is a totally bounded metric on $\mathbf{H}$, which induces a a totally bounded metric on $\mathcal{X}$ given by

$$d(x, y) := d_{\mathbf{H}}\big(\varphi(x), \varphi(y)\big).$$

Next, we show the reverse implication. Assume that $\mathcal{X}$ is a metrizable space and that d is a totally bounded metric on $\mathcal{X}$, i.e. for every $\varepsilon > 0$ there exists a finite collection of points $(x_i)_{i=1}^k$ in $\mathcal{X}$ satisfying

$$\mathcal{X} = \bigcup_{i=1}^{k} B(x_i, \varepsilon).$$

Thus, for each $n \in \mathbb{N}$ there exists a finite collection of points D_n satisfying

$$\mathcal{X} = \bigcup_{x_i \in D_n} B(x_i, 1/n).$$

Define $D = \cup_{n=1}^{\infty} D_n$, then D is a countable subset of $\mathcal{X}$ and we claim that $\overline{D} = \mathcal{X}$. Let $x \in \mathcal{X}$, then it follows from the construction of the sets D_n that we can create a sequence of points (x_n) such that $x_n \in D_n$ and

$$x \in B(x_n, 1/n), \quad \text{for every } n \in \mathbb{N}.$$

Thus $x_n \to x$, which proves that x is a limit point of D. It follows that D is dense in $\mathcal{X}$, and that $\mathcal{X}$ is separable.

■

The previous theorem shows that Polish spaces share some properties with compact metric spaces; they admit a complete metric and a topologically equivalent totally bounded metric. However, a as example 2 shows, a Polish space is in general not compact, the complete metric may not be the totally bounded metric. In fact, since compactness is a topological property it follows that every topologically consistent metric on a compact metrizable space is totally bounded and complete.

The Distance Function

Let $\mathcal{X}$ be a metric space with metric d and $A \subset \mathcal{X}$, then we define the *distance function* $d(x, A) : \mathcal{X} \to [0, \infty]$ by

$$d(x, A) := \inf_{y \in A} d(x, y).$$

For fixed $x \in \mathcal{X}$ we call $d(x, A)$ the distance between x and A. Similarly, we define the distance between two sets as

$$d(A, B) := \inf_{x \in A, y \in F} d(x, y)$$

A very useful property of the distance function is that it is uniformly continuous on $\mathcal{X}$ and this holds for every metric d and every $A \subset \mathcal{X}$.

Lemma 3.2. *Let $(\mathcal{X}, d)$ be a metric space and $A \subset \mathcal{X}$, then the distance function $d(\cdot, A)$ is uniformly continuous on $\mathcal{X}$.*

Proof. Let $A \subset \mathcal{X}$ and $x, y \in A$, then the triangle inequality implies that

$$\begin{aligned} d(x, A) &\leq d(x, y) + d(y, A) \\ d(y, A) &\leq d(x, y) + d(x, A) \end{aligned} \tag{24}$$

Rearranging, we get that

$$-d(x, y) \leq d(x, A) - d(y, A) \leq d(x, y). \tag{25}$$

Let $\varepsilon > 0$ and choose $\delta = \varepsilon$, then by equation (25) it follows that if $d(x, y) < \delta$

$$|d(x, A) - d(y, A)| < \varepsilon,$$

which shows that $d(\cdot, A)$ is uniformly continuous.

■

It follows directly from the proof that $d(\cdot, A)$ is not only uniformly continuous but also Lipschitz continuous with Lipschitz constant 1. Next we present a strengthening of Urysohn's Lemma for closed sets in metric spaces with positive distance between them.

Lemma 3.3 (Urysohn's Lemma in Metric Spaces)**.** *Let $\mathcal{X}$ be a metric space and $A, B \subset \mathcal{X}$ be two disjoint nonempty closed subsets of $\mathcal{X}$. If there exists a $\delta > 0$ such that $d(A, B) \geq \delta$, then there exists a bounded uniformly continuous function $f : \mathcal{X} \to [0, 1]$ which satisfy*

$$f(x) = \begin{cases} 1 & x \in A, \\ 0 & x \in B. \end{cases}$$

A direct proof is presented below. Another proof can be found in [11, Lemma 2.1] that relies on special properties of uniformly continuous functions.

Proof. Let A, B be disjoint nonempty closed subsets of $\mathcal{X}$, and define the function

$$f(x) := \frac{d(x, B)}{d(x, A) + d(x, B)}.$$

Then it is clear from this definition that $f = 0$ on B, and $f = 1$ on A. We are going to show that f is uniformly continuous. Let $\delta > 0$ and $d(A, B) \geq \delta$ and $x, y \in \mathcal{X}$, then

$$\begin{aligned} |f(x) - f(y)| &= \left| \frac{d(x, B)}{d(x, A) + d(x, B)} - \frac{d(y, B)}{d(y, A) + d(y, B)} \right| \\ &= \left| \frac{d(x, B)[d(y, A) + d(y, B)] - d(y, B)[d(x, A) + d(x, B)]}{(d(x, A) + d(x, B))(d(y, A) + d(y, B))} \right| \\ &\leq \frac{|d(x, B)[d(y, A) + d(y, B)] - d(y, B)[d(x, A) + d(x, B)]|}{\delta^2} \\ &= \frac{|d(x, B)d(y, A) - d(y, B)d(x, A)|}{\delta^2}. \end{aligned}$$

By subtracting and adding the term $d(y,A)d(y,B)$ and an application of the triangle inequality, we get that

$$\begin{aligned}|f(x)-f(y)| &\leq \frac{|d(x,B)d(y,A)-d(y,A)d(y,B)+d(y,A)d(y,B)-d(y,B)d(x,A)|}{\delta^2}\\ &= \frac{|(d(x,B)-d(y,B))d(y,A)+(d(y,A)-d(x,A))d(y,B)|}{\delta^2}\\ &\leq \frac{1}{\delta^2}|d(x,A)-d(y,A)|d(y,B)+\frac{1}{\delta^2}|d(x,B)-d(y,B)|d(y,A)|.\end{aligned} \tag{26}$$

By Lemma 3.2 it follows that the functions $d(\cdot,A), d(\cdot,B)$ are bounded and uniformly continuous. Let

$$M = \max\left(\sup_{x\in\mathcal{X}} d(x,A), \sup_{x\in\mathcal{X}} d(x,B)\right) < \infty,$$

and $\varepsilon > 0$. Then, by the uniform continuity there exists a $\eta > 0$ such that $|x-y| < \eta$ implies that

$$|d(x,A)-d(y,A)| \leq \varepsilon\frac{\delta^2}{2M}, \quad |d(x,B)-d(y,B)| \leq \varepsilon\frac{\delta^2}{2M}.$$

Combined with equation (26) this shows that that

$$|x-y| < \eta \implies |d(x,A)-d(y,A)| < \varepsilon,$$

hence, f is uniformly continuous.

■

Spaces of Bounded Continuous Functions

Definition 3.7. Let $\mathcal{X}$ be metrizable space, then we define $C_b(\mathcal{X})$ to be the space of bounded continuous real valued function on $\mathcal{X}$, equipped with the uniform norm

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|.$$

The space $C_b(\mathcal{X})$ is a Banach Space and if $\mathcal{X}$ is a metrizable space, then $C_b(\mathcal{X})$ is compact if and only if $\mathcal{X}$ is compact.

Theorem 3.6 ([21, Theorem 6.6].)**.** *Let $\mathcal{X}$ be a metrizable space, then $C_b(\mathcal{X})$ is separable if and only if $\mathcal{X}$ is compact.*

A class of very useful subspaces of $C_b(\mathcal{X})$ are those consisting of uniformly bounded continuous functions.

Definition 3.8. Let $\mathcal{X}$ be a metrizable space and d a metric on $\mathcal{X}$, then we define $U_b(\mathcal{X},d)$ to be *the space of bounded uniformly continuous functions* on $(\mathcal{X},d)$.

When the metric d is clear from context we drop it from the notation and simply write $U_b(\mathcal{X})$. Since uniform continuity is dependent on the metric d, there is not a unique subspace of uniformly continuous functions, but every admissible metric d on $\mathcal{X}$ has its corresponding subspace $U_b(\mathcal{X},d) \subset C_b(\mathcal{X})$. The next lemma shows that for every metric d on a metrizable space the collection of uniformly bounded continuous functions $U_b(\mathcal{X})$ forms a closed subspace of $C_b(\mathcal{X})$.

Theorem 3.7. *Let $\mathcal{X}$ be a metric space, then $U_b(\mathcal{X})$ is a closed subspace of $C_b(\mathcal{X})$ under the uniform norm.*

Proof. We will show that $U_b(\mathcal{X})$ is closed under a limits, hence a closed subset of $C_b(\mathcal{X})$. Let f be a limit point of $U_b(\mathcal{X})$ and (f_n) a sequence in $U_b(\mathcal{X})$ converging to f. It is clear that f must be bounded since it is an element of the space $C_b(\mathcal{X})$. It remains to show that the limit is uniformly continuous. This is showed by an application of the triangle inequality and a classical $\varepsilon/3$ argument. By the triangle inequality

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(y) - f(y)| + |f_n(x) - f_n(y)|,$$

The convergence of (f_n) to f in the uniform norm implies that for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies \|f - f_n\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - f_n(x)| < \frac{\varepsilon}{3}.$$

Fix any $n \geq N$, then it follows that

$$|f(x) - f(y)| \leq |f_n(x) - f_n(y)| + \frac{2\varepsilon}{3}.$$

By the uniform continuity of f_n it follows that there exists $\delta > 0$ such that $|x - y| < \delta$ implies

$$|f_n(x) - f_n(y)| < \frac{\varepsilon}{3}.$$

Thus, for $|x - y| < \delta$ we get

$$|f(x) - f(y)| < \varepsilon,$$

which shows that the limit f is uniformly continuous.

■

In contrast to the space $C_b(\mathcal{X})$, compactness is not a necessary condition for $U_b(\mathcal{X}, d)$ to be separable and it suffices that the space $(\mathcal{X}, d)$ is totally bounded for $U_b(\mathcal{X}, d)$ to be separable.

Theorem 3.8. *Let $(\mathcal{X}, d)$ be a totally bounded metric space, then $U_b(\mathcal{X}, d)$ is separable.*

The proof below follows the proof given in Parthasarathy [41, Lemma 6.3]. The main idea is to show that $U_b(\mathcal{X}, d)$ is isometric to a subset of the separable metric space $C(\hat{\mathcal{X}})$ and therefore it is separable.

Proof. Let $(\mathcal{X}, d)$ be totally bounded and $\hat{\mathcal{X}}$ denote its completion, then $\hat{\mathcal{X}}$ is totally bounded and complete and therefore compact. Any function $f \in U_b(\mathcal{X}, d)$ can be uniquely extended to a function $\hat{f} \in U_b(\hat{\mathcal{X}}, d)$ such that

$$\left\| \hat{f} \right\|_\infty = \|f\|_\infty,$$

see e.g. [5, Lemma 3.11]. The map which takes f to $\hat{f}$ is an isometric embedding of $U_b(\mathcal{X}, d)$ into $U_b(\hat{\mathcal{X}}, d)$ and $U_b(\mathcal{X}, d)$ is homeomorphic to a subset of $U_b(\hat{X}, d)$. Furthermore, as the space $\hat{\mathcal{X}}$ is compact, it follows that $U_b(\hat{X}, d) = C_b(\hat{\mathcal{X}})$, which is a separable metric space by Theorem 3.6. Thus $U_b(\mathcal{X}, d)$ is isometric to a subset of a separable metric which implies that $U_b(\mathcal{X}, d)$ is separable.

■

By Theorem 3.5 a metrizable space is separable if and only it admits a totally bounded metric, hence Theorem 3.8 implies the following.

Corollary 3.1. *Let $\mathcal{X}$ be a separable metrizable space, then $\mathcal{X}$ admits a metric d which makes $U_b(\mathcal{X}, d)$ separable.*

Measures on Metrizable Spaces

In this section we present some important results about Borel measures defined on metrizable spaces. A very useful property of Borel measures on metrizable spaces is that they are regular, i.e. the measure of a Borel set can be approximated from within by closed sets and approximated from without by open sets. We start with the definitions of regular and tight measures.

Definition 3.9. A Borel measure μ defined on a topological space is to be *tight* if for every $A \in \mathcal{B}_{\mathcal{X}}$, and every $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset A$ such that

$$\mu\big(A \setminus K_\varepsilon\big) < \varepsilon.$$

Similarly to regular measure a tight measure is completely determined by its value on compact sets, and if μ is tight and $A \in \mathcal{B}_{\mathcal{X}}$, then there exists a sequence of compact sets $K_n \subset A$ such that

$$\lim_{n\to\infty} \mu(K_n) = \mu(A).$$

Tightness is a stronger criterion than regularity of measures on metrizable spaces and the concept is closely related to weak convergence of probability measures on Polish spaces. In fact all Borel measures on metrizable spaces are regular and in Polish spaces every Borel measure is tight. This result will be used throughout this chapter and a proof of this can be found in Bogachev [13, Thm 7.1.7].

Theorem 3.9. *Let $\mathcal{X}$ be a metrizable space and μ a Borel measure on $\mathcal{X}$, then μ is regular, furthermore if $\mathcal{X}$ is Polish, then μ is tight.*

Whenever the space $\mathcal{X}$ is normal it is possible to achieve an integral variant of the criterion for equivalence of regular Borel measures. Two regular Borel measures are equivalent if their integrals over the space $\mathcal{X}$ agree for every bounded continuous function on $\mathcal{X}$. In metrizable spaces an even smaller class of functions suffices to determine whether two measures agree.

Theorem 3.10. *Let $(\mathcal{X}, d)$ be a metric space and μ, ν be two Borel measures on $\mathcal{X}$. Then $\mu = \nu$ if*

$$\int_{\mathcal{X}} f \,\mathrm{d}\mu = \int_{\mathcal{X}} f \,\mathrm{d}\nu\,, \quad \text{for every } f \in U_b(\mathcal{X}, d).$$

The general idea of the proof is to show that μ and ν agree on closed sets, and since Borel measures on metric spaces are regular this implies that they agree on all Borel measurable sets.

Proof. Let $C \subset \mathcal{X}$ be closed and μ, ν be two regular Borel measures satisfying the statements of the Theorem. Define the sets $C_n = \{x \in \mathcal{X} : d(x, C) < \frac{1}{n}\}$ to be the open neighbourhoods of C consisting of points which lie within distance $1/n$ from C. The sets C_n are open since they can be expressed as a union of open sets:

$$C_n = \bigcup_{x\in C} B(x, 1/n).$$

The complements C_n^c are disjoint with C and satisfy $d(C, C_n^c) \geq \frac{1}{n}$ thus we may apply Urysohn's Lemma for metric spaces (Lemma 3.3). Let f_n be a sequence of bounded uniformly continuous function taking values in $[0, 1]$ and satisfying

$$f_n(x) := \begin{cases} 1, x \in C, \\ 0, x \in C_n^c. \end{cases}$$

The sets $\{C_n\}$ form a decreasing sequence of measurable sets with $\cap C_n = C$, hence, by continuity of measure

$$\mu(C) \leq \lim_{n\to\infty} \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \leq \lim_{n\to\infty} \int_{C_n} \mathrm{d}\mu = \mu(C). \tag{27}$$

The same holds for the measure ν,

$$\nu(C) \leq \lim_{n\to\infty} \int_{\mathcal{X}} f_n \, \mathrm{d}\nu \leq \lim_{n\to\infty} \int_{C_n} \mathrm{d}\nu = \nu(C). \tag{28}$$

Since, the integral of of μ and ν agrees for every uniformly bounded continuous function it follows from equations (27) and (28) that $\mu(C) = \nu(C)$. This holds for every closed set, thus it follows from regularity of Borel measures on metric spaces that $\mu = \nu$.

■

3.3 Weak Convergence of Measures on Metric Spaces

This section studies the properties of $\mathbf{M}(\mathcal{X})$ and its subspaces $\mathbf{M}_+(\mathcal{X})$ and $\mathbf{M}_1(\mathcal{X})$ in the topology of weak convergence when $\mathcal{X}$ is a metrizable space. There are many great references that treat weak convergence of probability measures: Partasarathy [41], Bertsekas & Shreve [9], Billingsley [12], Stroock [48], and Aliprantis & Border [5] are just a few. However, in some situations the measures of interest are not guaranteed to be probability measures. An example of this is the empirical distribution of the IS estimator, which we study more closely in section 3.4. Many of the results from the theory of weak convergence of probability measure holds in the space $\mathbf{M}_+(\mathcal{X})$ of nonnegative finite measures. In fact, most of the results in Varadarajan's groundbreaking paper [51] on weak convergence in separable metric spaces are proven in the latter more general setting.

In this section we prove that many known weak convergence results, such as the Portmanteau Theorem hold in $\mathbf{M}_+(\mathcal{X})$. We show that much of the structure of $\mathcal{X}$ carries over to $\mathbf{M}_+(\mathcal{X})$ and that $\mathcal{X}$ is a separable metrizable space if and only if $\mathbf{M}_+(\mathcal{X})$ is separable metrizable space. Most proofs are based on the proofs given in the references above, but modified to hold in our more general setting. However, the main ideas underlying most of this chapter can be traced back to [51] and also [4] and [11] for the Portmanteau Theorem.

Since we have not shown that the weak topology on $\mathbf{M}(\mathcal{X})$ is first countable we cannot use sequences to characterize the convergence in this topology and we will instead work with nets for most of this section. In appendix A.3 the definition of nets can be found together with some facts about convergence in topological spaces. Let us restate the definition of weak convergence.

Definition 3.10. Let $\mathcal{X}$ be a metrizable space, then a net (μ_α) in $\mathbf{M}(\mathcal{X})$ *converges weakly* to $\mu \in \mathbf{M}(\mathcal{X})$ if

$$\lim_\alpha \int_{\mathcal{X}} f \, \mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f \, \mathrm{d}\mu \,, \quad \text{for every } f \in C_b(\mathcal{X}).$$

We write $\mu_\alpha \implies \mu$ to denote that μ_α converges to μ weakly.

Remark. In the definition of weak convergence we have used nets to characterize convergence and not sequences. The reason behind this is that we have not proved that the topology of weak convergence on $\mathbf{M}_+(\mathcal{X})$ is first countable, and thus sequences cannot be used to characterize important topological concepts such as continuity or limit points. However, we will see later that $\mathcal{X}$ is a separable metrizable space if and only if the space $\mathbf{M}_+(\mathcal{X})$ is metrizable and separable in which case working with sequences is perfectly fine.

The following theorem, known as the *Portmanteau Theorem*, gives several useful equivalent criteria for weak convergence of measures. Our proof is adapted from the one given by Partasarathy in [41, Theorem 6.1] and Billingsley in [11], which only prove the result for probability measures. The main ideas behind the Portmanteau Theorem go back to the work by Alexandroff [4].

Theorem 3.11 (Portmanteau Theorem)**.** *Let $\mathcal{X}$ be a metrizable space and (μ_α) be a net in $\mathbf{M}_+(\mathcal{X})$ and $\mu \in \mathbf{M}_+(\mathcal{X})$, then the statements below are equivalent.*

1. $\mu_\alpha \implies \mu$.
2. $\lim_\alpha \int_{\mathcal{X}} f \, \mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f \, \mathrm{d}\mu$ *for every* $f \in U_b(\mathcal{X})$.

Furthermore, if $\lim_\alpha \mu_\alpha(\mathcal{X}) = \mu(\mathcal{X})$*, then each of the following is equivalent to weak convergence.*

3. $\limsup_\alpha \mu_\alpha(C) \leq \mu(C)$ *for every closed* $C \subset \mathcal{X}$.
4. $\liminf_\alpha \mu_\alpha(U) \geq \mu(U)$ *for every open* $U \subset \mathcal{X}$.
5. $\lim_\alpha(A) = \mu(A)$ *for every* $A \in \mathcal{B}_{\mathcal{X}}$ *satisfying* $\mu(\partial A) = 0$.

Note that if all measures involved are probability measures, then $\mu_\alpha(\mathcal{X}) = \mu(\mathcal{X}) = 1$ for all α and all of statements above are equivalent. Normally the Portmanteau Theorem is stated only for the subspace $\mathbf{M}_1(\mathcal{X})$, but this generalized version will be useful to us later. The extra criterion for the equivalence of points 3-5 above to hold is not as limiting as it may seem, since it only concerns the limit over the entire space $\mathcal{X}$.

Proof. The implication 1 $\implies$ 2 follows directly from the definition of weak convergence and the implication 2 $\implies$ 1 follows from Theorem 3.10.

Now, assume that $\lim_\alpha \mu_\alpha(\mathcal{X}) = \mu(\mathcal{X})$. To prove that 2 $\implies$ 3 we use Urysohn's Lemma in metric spaces, Lemma 3.3. Let $C_n = \{x \in \mathcal{X} : d(x, C) < \frac{1}{n}\}$ be the open sets of points which lie within distance $1/n$ of C. Then, the complements C_n^{c} are disjoint from C and satisfy $d(C, C_n^{\mathrm{c}}) \geq \frac{1}{n}$. Thus, we can create a sequence of bounded uniformly continuous functions (f_n) taking values in $[0, 1]$ and

$$f_n := \begin{cases} 1, x \in C, \\ 0, x \in C_n^{\mathrm{c}}. \end{cases}$$

The sets $\{C_n\}$ form a decreasing sequence of measurable sets with $\cap C_n = C$, hence, by continuity of measure,

$$\lim_{n\to\infty} \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \leq \lim_{n\to\infty} \int_{C_n} \mathrm{d}\mu = \mu(C). \tag{29}$$

The functions $f_n \in U_b(\mathcal{X})$, thus by property 2 we get that

$$\limsup_{\alpha} \mu_\alpha(C) = \limsup_{\alpha} \int_C \mathrm{d}\mu_\alpha \leq \limsup_{\alpha} \int_{\mathcal{X}} f_n \, \mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \, .$$

Hence, by applying the limit in equation (29) the third statement follows. Next, we show that 3 $\implies$ 4. Let U be an open set, then the complement is closed and using 3 combined with the fact that $\mu_\alpha(\mathcal{X}) = \mu(\mathcal{X})$, for every α, leads to the inequality

$$\begin{aligned}\liminf_{\alpha} \mu_\alpha(U) &= \liminf_{\alpha} [\mu_\alpha(\mathcal{X}) - \mu_\alpha(U^{\mathrm{c}})] \geq \liminf_{\alpha} \mu_\alpha(\mathcal{X}) - \liminf_{\alpha} \mu_\alpha(U^{\mathrm{c}}) \\ &\geq \mu(\mathcal{X}) - \limsup_{\alpha} \mu_\alpha(U^{\mathrm{c}}) \geq \mu(\mathcal{X}) - \mu(U^{\mathrm{c}}) = \mu(U).\end{aligned}$$

From the above, it is apparent that 3 and 4 are equivalent, and combined they imply 5. To see this, let $A \in \mathcal{B}_{\mathcal{X}}$ satisfy $\mu(\partial A) = 0$, then

$$\mu(A^{\mathrm{o}}) = \mu(\overline{A}) = \mu(A).$$

3 and 4 gives the inequalities

$$\begin{aligned}\limsup_{\alpha} \mu_\alpha(A) \leq \limsup_{\alpha} \mu_\alpha(\overline{A}) \leq \mu(\overline{A}) = \mu(A). \\ \liminf_{\alpha} \mu_\alpha(A) \geq \liminf_{\alpha} \mu_\alpha(A^{\mathrm{o}}) \geq \mu(A^{\mathrm{o}}) = \mu(A).\end{aligned}$$

combining, these prove that

$$\liminf_{\alpha} \mu_\alpha(A) = \limsup_{\alpha} \mu_\alpha(A) = \mu(A).$$

Finally, we show that 5 $\implies$ 1. Let $f \in C_b(\mathcal{X})$, and (a, b) be an open interval that contains the image of f. The distribution of f, given by $\mu_f = \mu \circ f^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, hence finite, which implies that the collection of points such that $\mu(\{f = t\}) > 0$ is at most countable. For any $\varepsilon > 0$ we can make a partition $\Pi(\varepsilon)$ of (a, b) such that

(a) $a = t_0 < t_1 < t_2 < \cdots < t_n = b$

(b) $\mu(\{f = t_i\}) = 0$ for every $t_i \in \Pi(\varepsilon)$

(c) $t_i - t_{i-1} < \varepsilon$ for every $i = 1, \ldots, n$.

Then we can approximate f by the simple function

$$\varphi = \sum_{i=1}^{n} t_{i-1} \mathbb{1}_{F_i}, \quad F_i = \{x \in \mathcal{X} \, : \, t_{i-1} \leq f(x) < t_i\}.$$

By the construction of these sets it follows from (b) that

$$\mu(\partial F_i) = \mu(\{f = t_{i-1}\}) + \mu(\{f = t_i\}) = 0,$$

and furthermore (c) implies that $|f-\varphi|<\varepsilon$ on all of $\mathcal{X}$. Consequently, since the measures μ_α and μ are all finite measures, we get that

$$\int_{\mathcal{X}} |f-\varphi|\,\mathrm{d}\mu_\alpha < \varepsilon\,\mu_\alpha(\mathcal{X}), \quad \int_{\mathcal{X}} |\varphi-f|\,\mathrm{d}\mu < \varepsilon\mu(\mathcal{X}).$$

Thus, by the triangle inequality we obtain the bound

$$\begin{aligned}\left|\int_{\mathcal{X}} f\,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f\,\mathrm{d}\mu\right| &\le \left|\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} \varphi \mathrm{d}\mu\right| + \varepsilon\mu_\alpha(\mathcal{X}) + \varepsilon\mu(\mathcal{X}) \\ &\le \sum_{i=1}^{n} t_{i-1}|\mu_\alpha(F_i)-\mu(F_i)| + \varepsilon\mu_\alpha(\mathcal{X}) + \varepsilon\mu(\mathcal{X}).\end{aligned}$$

Since the sets F_i all have boundaries with measure zero property 5 implies that

$$\lim_\alpha \sum_{i=1}^{n} t_{i-1}|\mu_\alpha(F_i)-\mu(F_i)| = 0,$$

and

$$\lim_\alpha \mu_\alpha(\mathcal{X}) = \mu(\mathcal{X}).$$

Hence,

$$\lim_\alpha \left|\int_{\mathcal{X}} f\,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f\,\mathrm{d}\mu\right| \le 2\varepsilon\mu(\mathcal{X}),$$

and this holds for arbitrary $\varepsilon>0$, which implies that $\mu_\alpha \implies \mu$.

■

Following up on the previous remark, the Portmanteau Theorem is commonly stated for sequences of measures only and not for nets. Then it is possible to show that 5 implies either 3 or 1 by using Lebesgue's Bounded Convergence Theorem or Fatou's Lemma for some sequence of functions created from μ_n. However, that is not possible when working with nets as those results are only valid for sequences. There are several versions of the Portmanteau Theorem and some of the equivalences hold under weaker criteria in subsets of the space of finite measures and even when $\mathcal{X}$ is not a metrizable space. Results of this nature in more general topological spaces can be found in the work by Topsøe [49, see especially Thm 8.1 page 40] and in Chapter 8 of Bogachev [13].

We will now move on with our study of nonnegative finite measures in the weak topology by using the second part of the Portmanteau Theorem to show that if $U_b(\mathcal{X},d)$ is separable, then in order to prove weak convergence it is sufficient to show that the integral convergence criteria of weak convergence holds for a countable dense subset of $U_b(\mathcal{X},d)$.

Theorem 3.12. *Let $(\mathcal{X},d)$ be a metric space. If $U_b(\mathcal{X},d)$ is separable, then there exists a countable collection $\{f_n\}\subset U_b(\mathcal{X},d)$ such that any net $\mu_\alpha \implies \mu$ in $\mathbf{M}_+(\mathcal{X})$ if*

$$\lim_\alpha \int_{\mathcal{X}} f_n\,\mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f_n\,\mathrm{d}\mu\,, \quad \textit{for every } n\in\mathbb{N}.$$

Proof. Let $U_b(\mathcal{X},d)$ be separable and $\{f_n\}\subset U_b(\mathcal{X},d)$ a countable dense subset. Let (μ_α) be a net in $\mathbf{M}_+(\mathcal{X})$, $\mu\in\mathbf{M}_+(\mathcal{X})$, and assume that

$$\lim_\alpha \int_{\mathcal{X}} f_n\,\mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f_n\,\mathrm{d}\mu\,, \quad \text{for every } n\in\mathbb{N}. \tag{30}$$

Let $f \in U_b(\mathcal{X}, d)$, then

$$\begin{aligned}\left|\int_{\mathcal{X}} f \,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f \,\mathrm{d}\mu\right| \leq \left|\int_{\mathcal{X}} f_n \,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f_n \,\mathrm{d}\mu\right| \\ + \int_{\mathcal{X}} |f - f_n| \,\mathrm{d}\mu_\alpha + \int_{\mathcal{X}} |f_n - f| \,\mathrm{d}\mu\end{aligned} \tag{31}$$

For every ε, there exists $n \in \mathbb{N}$ such that

$$\|f - f_n\|_\infty < \varepsilon.$$

Thus, it follows from equation (31) that

$$\left|\int_{\mathcal{X}} f \,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f \,\mathrm{d}\mu\right| < \left|\int_{\mathcal{X}} f_n \,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f_n \,\mathrm{d}\mu\right| + \varepsilon\mu_\alpha(\mathcal{X}) + \varepsilon\mu(\mathcal{X}),$$

for every $\varepsilon > 0$. The separability of $\{f_n\} \subset \mathcal{X}$ combined with equation (30) implies that

$$\lim_\alpha \mu_\alpha(\mathcal{X}) = \mu(\mathcal{X}).$$

Hence, by taking limits with respect to α it follows that

$$\lim_\alpha \left|\int_{\mathcal{X}} f \,\mathrm{d}\mu_\alpha - \int_{\mathcal{X}} f \,\mathrm{d}\mu\right| < 2\varepsilon\mu(\mathcal{X}),$$

for every $\varepsilon > 0$. Thus

$$\lim_\alpha \int_{\mathcal{X}} f \,\mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f \,\mathrm{d}\mu\,, \quad \text{for every } f \in U_b(\mathcal{X}, d),$$

and the Portmanteau Theorem (Theorem 3.11) therefore implies that $\mu_\alpha \implies \mu$.

■

A useful property of the topology of weak convergence is that the map $x \longmapsto \delta_x$ which maps points of $\mathcal{X}$ to their Dirac measures in $\mathbf{M}(\mathcal{X})$ is not only continuous, but also an embedding. Thus $\mathcal{X}$ is homeomorphic to $\{\delta_x : x \in \mathcal{X}\} \subset \mathbf{M}(\mathcal{X})$.

Lemma 3.4. *Let $\mathcal{X}$ be a metric space, then $\mathcal{X}$ can be embedded into $\mathbf{M}(\mathcal{X})$ by the map $x \mapsto \delta_x$.*

Proof. Let $\varphi : \mathcal{X} \to \mathbf{M}(\mathcal{X})$ be the map defined by $x \mapsto \delta_x$. We want to show that φ is a homeomorphism of $\mathcal{X}$ onto $\varphi(\mathcal{X}) = D$. We start by showing that φ is injective. Let $x, y \in \mathcal{X}$ and $x \neq y$, then it is clear that $\delta_x \neq \delta_y$ since

$$1 = \delta_x(\{x\}) \neq \delta_y(\{x\}) = 0.$$

In order to prove that the map φ is bicontinuous is suffices to show that $x_\alpha \to x$ in $\mathcal{X}$ if and only if $\delta_{x_\alpha} \to \delta_x$ in $\mathbf{M}(\mathcal{X})$ for every net in (x_α) in $\mathcal{X}$ (see Theorem A.6).

Let $x_\alpha \to x$, and $f \in C_b(\mathcal{X})$ then $f(x_\alpha) \to f(x)$ and therefore

$$\int_{\mathcal{X}} f \,\mathrm{d}\delta_{x_\alpha} \to \int_{\mathcal{X}} f \,\mathrm{d}\delta_x\,,$$

which shows that φ is continuous. We prove the converse by proving the contrapositive. Let $x_\alpha \not\to x$, and $f \in C_b(\mathcal{X})$, then $f(x_\alpha) \not\to f(x)$, which implies that

$$\int_{\mathcal{X}} f \,\mathrm{d}\delta_{x_\alpha} \not\to \int_{\mathcal{X}} f \,\mathrm{d}\delta_x\,.$$

Hence, $\delta_{x_\alpha} \not\Longrightarrow \delta_x$ proving that φ^{-1} is continuous. It follows that φ is a homeomorphism onto its image.

■

Metrizability of $\mathbf{M}_+(\mathcal{X})$

We are now ready to prove one of the main theorems regarding the topological properties of the space $\mathbf{M}_+(\mathcal{X})$, which states $\mathbf{M}_+(\mathcal{X})$ is separable and metrizable if and only if $\mathcal{X}$ is so as well. This result is originally by Varadarajan [51] and we present his proof below.

Theorem 3.13. *The space $\mathbf{M}_+(\mathcal{X})$ is separable and metrizable if and only if $\mathcal{X}$ is separable and metrizable.*

Proof. The idea is to prove that $\mathbf{M}_+(\mathcal{X})$ can be embedded into a separable metric space and hence is separable and metrizable. Let $\mathcal{X}$ be a separable metrizable space, then there exists a totally bounded metric d on $\mathcal{X}$ which makes the space $U_b(\mathcal{X}, d)$ separable (see Theorems 3.5 and 3.8). By Theorem 3.12 there exists a countable collection $\{f_n\} \subset U_b(\mathcal{X}, d)$ such that a net $(\mu_\alpha) \in \mathbf{M}_+(\mathcal{X})$ converges to $\mu \in \mathbf{M}_+(\mathcal{X})$ if and only if

$$\lim_\alpha \int_{\mathcal{X}} f_n \, \mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \,, \quad \text{for every } n \in \mathbb{N}. \tag{32}$$

Let $T : \mathbf{M}_+(\mathcal{X}) \to \mathbb{R}^{\mathbb{N}}$ be the map defined by

$$T(\mu) = \left(\textstyle\int_{\mathcal{X}} f_1 \, \mathrm{d}\mu \,, \int_{\mathcal{X}} f_2 \, \mathrm{d}\mu \,, \ldots\right)$$

We are going to show that T is an embedding. We start by showing that T is injective. Let $\mu \neq \nu$, then by Theorem 3.10 there exists a function $f \in U_b(\mathcal{X}, d)$ such that

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu \neq \int_{\mathcal{X}} f \, \mathrm{d}\nu \,.$$

Since the collection $\{f_n\}$ is dense in $U_b(\mathcal{X}, d)$ it follows that there exists $f_k \in \{f_n\}$ such that

$$\int_{\mathcal{X}} f_k \, \mathrm{d}\mu \neq \int_{\mathcal{X}} f_k \, \mathrm{d}\nu \,,$$

which implies that $T(\mu) \neq T(\nu)$. Next, we show that T is continuous. Let $\mu_\alpha \to \mu$ in $\mathbf{M}_+(\mathcal{X})$, then it follows from equation (32) that $T(\mu_\alpha) \to T(\mu)$ coordinate-wise, which implies that that $T(\mu_\alpha) \to T(\mu)$ in the product topology on $\mathbb{R}^{\mathbb{N}}$.

It remains to show that T^{-1} is continuous. Let (μ_α) be a net in $\mathcal{X}$ and $T(\mu_\alpha) \to T(\mu)$. Then by Theorem 3.12 $\mu_\alpha \implies \mu$, which shows that T^{-1} is continuous. Thus T is an embedding and it follows that $\mathbf{M}_+(\mathcal{X})$ is homeomorphic to a subset of a separable metrizable space, hence $\mathbf{M}_+(\mathcal{X})$ is separable and metrizable.

To prove the converse we assume that $\mathbf{M}_+(\mathcal{X})$ is separable and metrizable. By Lemma 3.4 $\mathcal{X}$ is homeomorphic to the subset of Dirac measures in $\mathbf{M}_+(\mathcal{X})$. Hence, $\mathcal{X}$ is separable and metrizable.

■

The embedding, T, of $\mathbf{M}_+(\mathcal{X})$ into the separable metric space $\mathbb{R}^{\mathbb{N}}$ which is used in the proof of Theorem 3.13 is very useful and we will use it several times in this section. It follows from Theorem 3.13 that $\mathbf{M}_+(\mathcal{X})$ is metrizable, and hence second countable, whenever $\mathcal{X}$ is separable and metrizable. With this in mind we will work with sequences of measures instead of nets, whenever we know that $\mathbf{M}_+(\mathcal{X})$ is metrizable.

Compactness in $\mathbf{M}(\mathcal{X})$

In this subsection we study compact subsets of $\mathbf{M}(\mathcal{X})$ and prove that if $\mathcal{X}$ is a metric space, then $\mathbf{M}_1(\mathcal{X})$ is compact if and only if $\mathcal{X}$ is compact. The general idea to show this relies on Riesz Representation Theorem, Theorem 3.4, which states that when $\mathcal{X}$ is compact, then the dual of $C_b^*(\mathcal{X})$ is linearly isometric to $\mathbf{M}(\mathcal{X})$. Thus, for compact metric space the topology of weak convergence is the weak* topology induced by $C_b(\mathcal{X})$ on its dual $\mathbf{M}(\mathcal{X})$, and we can use the Banach-Alaoglu Theorem to gain some insight into the compact subsets of $\mathbf{M}(\mathcal{X})$.

Lemma 3.5. *Let $\mathcal{X}$ be a compact metrizable space and a, b be real numbers that satisfy $0 \leq a \leq b$, then the set $\{\mu \in \mathbf{M}(\mathcal{X}) \ : \ a \leq \mu(\mathcal{X}) \leq b\}$ is a compact subset of $\mathbf{M}(\mathcal{X})$ in the topology of weak convergence.*

Proof. Let $0 \leq a \leq b$ and let $K = \{\mu \in \mathbf{M}_+(\mathcal{X}) \ : \ a \leq \mu(\mathcal{X}) \leq b\}$. It follows from the Banach-Alaoglu Theorem (see e.g. [29, p. V.4.3]) $K \subset \mathbf{M}_+(\mathcal{X})$ is compact in the weak* topology if K is bounded in the total variation norm and closed in the weak* topology. It is clear that the sets K are bounded in the total variation norm, since

$$\|\mu - \nu\|_{TV} \leq |\mu|(\mathcal{X}) + |\nu|(\mathcal{X}) \leq 2b, \quad \text{for every } : \ \mu, \nu \in K.$$

Furthermore it is clear from the definition of weak convergence that if $\mu_\alpha \implies \mu$, then by using the constant function 1 as a test function, $\mu_\alpha(\mathcal{X}) \to \mu(\mathcal{X})$, which implies that K is closed under limits and thus closed. Hence K is compact by the Banach-Alaoglu Theorem.

■

Especially, it follows that the spaces $\mathbf{M}_1(\mathcal{X})$ and $\mathbf{M}_{\leq 1}(\mathcal{X})$ are compact in the topology of weak convergence. Combining the above with the embedding of $\mathcal{X}$ into $\mathbf{M}(\mathcal{X})$ given in Lemma 3.4 we get the following useful result.

Lemma 3.6. *The space $\mathbf{M}_1(\mathcal{X})$ is compact and metrizable if and only if $\mathcal{X}$ is compact and metrizable.*

Proof. Let $\mathcal{X}$ be compact and metrizable, then it follows from the previous Lemma that $\mathbf{M}_1(\mathcal{X})$ is compact. Conversely, assume that $\mathbf{M}_1(\mathcal{X})$ is compact. By Lemma 3.4 the space $\mathcal{X}$ is homeomomorphic to the set of Dirac measures in $\mathcal{X}$, which is a subset of the compact and metrizable space $\mathbf{M}_1(\mathcal{X})$. Hence, $\mathcal{X}$ is compact and metrizable.

■

Separability and Completeness

In this section we study the separability and completeness properties of $\mathbf{M}_+(\mathcal{X})$ in greater detail. The main result of this section , due to Varadarajan [51], is given in Theorem 3.14 below which states that $\mathcal{X}$ is Polish if and only if $\mathbf{M}_+(\mathcal{X})$ is Polish.

Lemma 3.7. *Let $\mathcal{X}$ be a compact metrizable space, then $\mathbf{M}_+(\mathcal{X})$ is Polish.*

Proof. Let $\mathcal{X}$ be compact and metrizable with d any admissible metric for the topology on $\mathcal{X}$. Then $C_b(\mathcal{X}) = U_b(\mathcal{X}, d)$ is separable by Theorem 3.6 and therefore, by Theorem 3.12 there exists a countable dense subset $\{f_n\} \subset U_b(\mathcal{X}, d)$ such that a net (μ_α) converges to $\mu \in \mathbf{M}_+(\mathcal{X})$ if and only if

$$\lim_\alpha \int_{\mathcal{X}} f_n \, \mathrm{d}\mu_\alpha = \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \,, \quad \text{for every } n \in \mathbb{N}. \tag{33}$$

Thus, the map $T : \mathbf{M}_+(\mathcal{X}) \to \mathbb{R}^{\mathbb{N}}$ defined by

$$T(\mu) = \left(\int_{\mathcal{X}} f_1 \,\mathrm{d}\mu \,, \int_{\mathcal{X}} f_2 \,\mathrm{d}\mu \,, \ldots\right)$$

is a continuous embedding of $\mathbf{M}_+(\mathcal{X})$ into $\mathbb{R}^{\mathbb{N}}$, which is a Polish space. A closed subset of a Polish space is Polish, thus it suffices to show that the image $T(\mathbf{M}_+(\mathcal{X}))$ is a closed subset of $\mathbb{R}^{\mathbb{N}}$. Let $(\mathbf{x}_k)$ be a sequence in $T(\mathbf{M}_+(\mathcal{X}))$ converging to some limit $\mathbf{x} \in \mathbb{R}^{\mathbb{N}}$, then the sequence of measures $\mu_k = T^{-1}(\mathbf{x}_k) \in \mathbf{M}_+(\mathcal{X})$ satisfies

$$\lim_{k\to\infty} \int_{\mathcal{X}} f_n \,\mathrm{d}\mu_k = x_n \in \mathbb{R}, \quad \text{for every } n \in \mathbb{N}.$$

We need to show that there exists a measure $\mu \in \mathbf{M}_+(\mathcal{X})$ such that $\mu = T^{-1}(\mathbf{x})$. By the exact same reasoning as in the proof of Theorem 3.12 the limits

$$\Lambda(f) := \lim_{k\to\infty} \int_{\mathcal{X}} f \,\mathrm{d}\mu_k \in \mathbb{R}$$

exist for every $f \in C_b(\mathcal{X})$. It is clear that Λ is positive functional on $C_b(\mathcal{X})$, furthermore by linearity of integration it follows that Λ is linear. Hence, the Riesz Representation Theorem implies that there exists a measure $\mu \in \mathbf{M}_+(\mathcal{X})$ such that

$$\Lambda(f) = \int_{\mathcal{X}} f \,\mathrm{d}\mu \,, \quad \text{for every } f \in C_b(\mathcal{X}).$$

By the definition of Λ we see that $\mu_k \implies \mu$, and it follows that the image of $\mathbf{M}_+(\mathcal{X})$ under T is closed and therefore Polish.

■

Before we prove the main result of this section we introduce the following useful Lemma. It is a generalization of Theorem 15.14 from [5] extended from probability measures to non-negative finite measures.

Lemma 3.8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $\varphi : \mathcal{X} \to \mathcal{Y}$ be an embedding. Then the map $\varphi_* : \mathbf{M}_+(\mathcal{X}) \to \mathbf{M}_+(\mathcal{Y})$ defined by $\mu \mapsto \mu \circ \varphi^{-1}$ is an embedding.*

Proof. Let φ be an embedding of $\mathcal{X} \hookrightarrow \mathcal{Y}$. Then φ is a homeomorphism onto its image, and it follows that $A \in \mathcal{B}_{\mathcal{X}}$ if and only if $\varphi(A) \in \mathcal{B}_{\mathcal{Y}}$. Thus φ_* and ${\varphi_*}^{-1}$ are well defined. Next, we show that φ_* is continuous. If $f \in C_b(\mathcal{Y})$, then it follows from the continuity of φ that $f \circ \varphi$ is a bounded continuous function on $\mathcal{X}$ and that

$$\int_{\mathcal{X}} f \circ \varphi \,\mathrm{d}\mu = \int_{\mathcal{Y}} f \,\mathrm{d}\varphi_*(\mu) \,,$$

for every $\mu \in \mathbf{M}(\mathcal{X})$. Thus, if (μ_α) is a net converging to μ in $\mathbf{M}_+(\mathcal{X})$ and $f \in C_b(\mathcal{Y})$, then

$$\lim_{n\to\infty} \int_{\mathcal{X}} f \,\mathrm{d}\varphi_*(\mu_n) = \lim_{n\to\infty} \int_{\mathcal{X}} f \circ \varphi \,\mathrm{d}\mu_n = \int_{\mathcal{X}} f \circ \varphi \,\mathrm{d}\mu = \int_{\mathcal{Y}} f \,\mathrm{d}\varphi_*(\mu) \,.$$

Hence, φ_* is continuous.

Next, we show that φ_* is injective. Let $\mu \neq \nu$ be elements of $\mathbf{M}_+(\mathcal{X})$. Since $\mathcal{X}$ is Polish it follows from Theorem 3.9 that μ and ν are tight. Thus, there exists a compact set $K \subset \mathcal{X}$ such that $\mu(K) \neq \nu(K)$. The compactness of K ensures that the image $\varphi(K)$ is compact in $\mathcal{Y}$ and therefore closed and Borel measurable. Hence,

$$\varphi_*(\varphi(K)) = \mu(K) \neq \nu(K) = \varphi_*(\varphi(K)),$$

which shows that φ_* is injective.

It remains to show that ${\varphi_*}^{-1}$ is a continuous map from the image of $\mathbf{M}_+(\mathcal{X})$ under φ_*. Let $\nu = \mu \circ \varphi^{-1} \in \varphi_*\big(\mathbf{M}_+(\mathcal{X})\big)$, and $g \in C_b(\mathcal{X})$, then $g\varphi^{-1} \in C_b(\mathcal{Y})$ and

$$\int_{\mathcal{X}} g \, \mathrm{d}\mu = \int_{\mathcal{Y}} g \circ \varphi^{-1} \, \mathrm{d}\nu$$

It follows from the above equality that if $\nu_n \implies \nu$ in $\varphi_*\big(\mathbf{M}_+(\mathcal{X})\big)$, then $\mu_n \implies \mu$ in $\mathbf{M}(\mathcal{X})$ and ${\varphi_*}^{-1}$ is continuous. Which proves that φ_* is homeomorphism onto its image, i.e. an embedding.

■

We are now ready to prove the main result of this section.

Theorem 3.14. *$\mathbf{M}_+(\mathcal{X})$ is a Polish space if and only if $\mathcal{X}$ is a Polish space.*

Proof. Let $\mathcal{X}$ be a Polish space, then it follows from the separability of $\mathcal{X}$ that $\mathcal{X}$ admits a totally bounded metric d. Let $\hat{\mathcal{X}}$ denote the completion of $(\mathcal{X}, d)$, which is also totally bounded, hence compact. Thus, Lemma 3.7 implies that $\mathbf{M}_+(\hat{\mathcal{X}})$ is Polish. Now, let φ be the embedding of $\mathcal{X}$ into $\hat{\mathcal{X}}$ induced by the completion. Since $\mathcal{X}$ is Polish it follows that $\hat{\mathcal{X}}$ also is Polish and therefore Lemma 3.8 implies that the map φ_* defined by $\mu \mapsto \mu \circ \varphi^{-1}$ is an embedding of $\mathbf{M}_+(\mathcal{X})$ into $\mathbf{M}_+(\hat{\mathcal{X}})$. It follows from Alexandroff's Lemma (see e.g. [53, Theorem 24.12]) that every G_δ subset of $\mathbf{M}_+(\hat{\mathcal{X}})$ is Polish.

Now, we show that $\varphi_*(\mathbf{M}_+(\mathcal{X}))$ is G_δ in $\mathbf{M}_+\big(\hat{\mathcal{X}}\big)$. Since $\mathcal{X}$ is Polish it is G_δ in $\hat{\mathcal{X}}$ (see e.g. [53, Theorem 24.13]), i.e. there exists a sequence G_n of open sets such that

$$\varphi(\mathcal{X}) = \bigcap_{n=1}^{\infty} G_n \tag{34}$$

Note that $\varphi^{-1}(\hat{\mathcal{X}}) = \varphi^{-1}(\mathcal{X}) = \mathcal{X}$, thus

$$\varphi^{-1}(\hat{\mathcal{X}} \smallsetminus \mathcal{X}) = \varphi^{-1}(\hat{\mathcal{X}}) \smallsetminus \varphi^{-1}(\mathcal{X}) = \emptyset,$$

and

$$\varphi_*(\mathbf{M}_+(\mathcal{X})) = \{\mu \in \mathbf{M}_+(\hat{\mathcal{X}}) : \mu(\hat{\mathcal{X}} \smallsetminus \mathcal{X}) = 0\}.$$

Combining this with equation (34), we get that

$$\begin{aligned}
\varphi_*(\mathbf{M}_+(\mathcal{X})) &= \{\mu \in \mathbf{M}_+(\hat{\mathcal{X}}) : \mu(\hat{\mathcal{X}} \smallsetminus \mathcal{X}) = 0)\} \\
&= \bigcap_{n=1}^{\infty} \left\{\mu \in \mathbf{M}_+(\hat{\mathcal{X}}) : \mu\left(\hat{\mathcal{X}} \smallsetminus G_n\right) = 0\right\} \\
&= \bigcap_{n=1}^{\infty} \bigcap_{k=1}^{\infty} \left\{\mu \in \mathbf{M}_+(\hat{\mathcal{X}}) : \mu\left(\hat{\mathcal{X}} \smallsetminus G_n\right) < \frac{1}{k}\right\}.
\end{aligned}$$

We want to show that the sets

$$U_{k,n} = \bigcap_{n=1}^{\infty} \bigcap_{k=1}^{\infty} \left\{\mu \in \mathbf{M}_+(\hat{\mathcal{X}}) : \mu\left(\hat{\mathcal{X}} \smallsetminus G_n\right) < \frac{1}{k}\right\}$$

are open for every $n, k \in \mathbb{N}$ or, equivalently, that the sets $U_{k,n}^{\mathrm{c}}$ are closed. Let (μ_α) be a net in $\mathbf{M}_+(\hat{\mathcal{X}})$ satisfying $\mu\left(\hat{\mathcal{X}} \smallsetminus G_n\right) \geq 1/k$ that converges weakly to some $\mu \in \mathbf{M}_+(\hat{\mathcal{X}})$.

Then, since $\hat{\mathcal{X}} \setminus G_n$ is closed, it follows from the Portmanteau Theorem that

$$\mu\left(\hat{\mathcal{X}} \setminus G_n\right) \geq \limsup_{\alpha} \mu\left(\hat{\mathcal{X}} \setminus G_n\right) \geq \frac{1}{k},$$

which shows that $\mu \in U_{k,n}^{c}$. It follows that the sets $U_{k,n}$ are open and that $\varphi_*\left(\mathbf{M}_+(\mathcal{X})\right)$ is G_δ in $\mathbf{M}_+\left(\hat{\mathcal{X}}\right)$, hence Polish. Thus, $\mathbf{M}_+(\mathcal{X})$ is Polish, since φ_* is homeomorphic onto its image.

Conversely, assume that $\mathbf{M}_+(\mathcal{X})$ is Polish. Then, since $\mathcal{X}$ is homeomorphic to the collection of Dirac measures in $\mathbf{M}_+(\mathcal{X})$ by Lemma 3.4, it suffices to show that $\{\delta_x : x \in \mathcal{X}\}$ is closed with respect to weak convergence. Let (δ_{x_n}) be a sequence in $\mathbf{M}_+(\mathcal{X})$ that converges weakly to some $\mu \in \mathbf{M}_+(\mathcal{X})$, and assume for contradiction that $\mu \neq \delta_x$ for every $x \in \mathcal{X}$. It follows from Lemma 3.4 that a sequence $x_n \to x$ in $\mathcal{X}$ if and only if $\delta_{x_n} \implies \delta_x$ in $\mathbf{M}_+(\mathcal{X})$. Any convergent subsequence of (δ_{x_n}) converges to μ, thus (x_n) cag have any convergent subsequence (x_{n_k}). This implies for every $x \neq x_n$ there exists $\varepsilon_x > 0$ such that $B(x, \varepsilon_x) \cap x_n = \emptyset$ for every $n \in \mathbb{N}$. The union of these balls is open, hence the complement, which is $\{x_n : n \in \mathbb{N}\}$, is closed. By assumption $\delta_{x_n} \implies x$, hence the Portmanteau Theorem implies that

$$\mu\big(\{x_m : m \in \mathbb{N}\}\big) \geq \limsup_{n\to\infty} \delta_{x_n}\left(\{x_m : m \in \mathbb{N}\}\right) = 1. \tag{35}$$

However, the same is true for every infinite subset of $\{x_m : m \in \mathbb{N}\}$. Especially, by considering the disjoint subsets $\{x_{2m} : m \in \mathbb{N}\}$ and $\{x_{2m+1} : m \in \mathbb{N}\}$ we get that $\mu(\{x_m : m \in \mathbb{N}\}) \geq 2$. But then equation (35) cannot hold, which is a contradiction.

Thus, there exists a subsequence x_{n_k} that converges to x, which implies that $\delta_{x_{n_k}} \implies \delta_x$, and that $\{\delta_x : x \in \mathcal{X}\}$ is closed.

■

Tightness and Compact Subsets of Measures

In Polish spaces the relatively compact subsets of $\mathbf{M}(\mathcal{X})$ can be characterized by as exactly those which are uniformly tight.

Definition 3.11. A collection $\mathcal{F} \subset \mathbf{M}(\mathcal{X})$ is *(uniformly) tight* if for every $\varepsilon > 0$ there exists a compact set K such that

$$|\mu|(\mathcal{X} \setminus K) < \varepsilon, \quad \text{for every } \mu \in \mathbf{F}.$$

The following theorem due to Prokhorov give a characterization of the relatively compact subsets of $\mathbf{M}(\mathcal{X})$ (see e.g. [13, Theorem 8.6.2]).

Theorem 3.15 (Prokhorov's Theorem)**.** *Let $\mathcal{X}$ be a Polish space and $\mathcal{K} \subset \mathbf{M}(\mathcal{X})$. then $\mathcal{K}$ is tight if and only if $\mathcal{K}$ is relatively compact.*

Remark. Recall $K \subset \mathcal{X}$ is relatively compact if $\overline{K} = \mathcal{X}$ and this is true if and only if every sequence in K contains a weakly convergent subsequence.

For nonnegative measures we also have the following useful result (see e.g. [13, Theorem 8.3.4] for a slightly more general statement).

Theorem 3.16. *Let $\mathcal{X}$ be a Polish space and (μ_n) a sequence of measures in $\mathbf{M}_+(\mathcal{X})$ that converges weakly to $\mu \in \mathbf{M}_+(\mathcal{X})$, then the collection $\{\mu_n\}$ is uniformly tight.*

3.4 Empirical Distributions

Another interpretation of Monte Carlo estimators can be given through their empirical distributions. We define the empirical distribution of a sequence of i.i.d. random variables (X_i) taking values in $\mathcal{X}$ to be the map $\mathbf{L}_n : \Omega \to \mathbf{M}_1(\mathcal{X})$ defined by

$$\mathbf{L}_n(\omega)(E) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}(E), \quad \delta_{X_i}(E) = \mathbb{1}_E[X_i(\omega)], \quad E \in \mathcal{B}_{\mathcal{X}}, \quad \omega \in \Omega. \tag{36}$$

Note that the empirical distributions are not measures, as the name may suggest, however for any fixed $\omega \in \Omega$, the value of $\mathbf{L}_n$ is a probability measure on $\mathcal{X}$. Thus, we may think of $\mathbf{L}_n$ as a random variable taking values in $\mathbf{M}_1(\mathcal{X})$. Sanov's Theorem, which we prove in the next chapter, states that the distributions of the random variables $\mathbf{L}_n$ satisfy a large deviation principle. A theorem proved by Varadarajan in [50] establishes an important fact about the convergence of the empirical distributions: in separable metric spaces the empirical distributions converge weakly to μ almost surely. We are now going to extend this result to the empirical distributions of the IS estimators defined in 2.2. The empirical distribution of the IS estimator is given by

$$\mathbf{I}_n(\omega) := \frac{1}{n}\sum_{i=1}^{n} \rho\big(Y_i(\omega)\big)\delta_{Y_i(\omega)}.$$

Note that $\mathbf{I}_n$ is in general not a probability measure, since

$$\mathbf{I}_n(\omega)(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}\mathbf{I}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} \rho\big(Y_i(\omega)\big)\delta_{Y_i(\omega)}. \tag{37}$$

Thus, $\mathbf{I}_n(\omega)$ is a probability measure if and only if $\rho(Y_i(\omega)) = 1$ for every $i = 1, \ldots, n$, and the empirical distribution of the IS estimator is map from Ω into the space of all positive signed measures $\mathbf{M}_+(\mathcal{X})$. The result of Varadarajan given in [50] can be extended to $\mathbf{I}_n$ with the proof practically unmodified.

Theorem 3.17. *Let $\mathcal{X}$ be a separable metrizable space and (Y_i) a sequence of i.i.d. random variables taking values in $\mathcal{X}$ with distribution π. If X is another random variable on $\mathcal{X}$ with distribution $\mu \ll \pi$, then the empirical distributions of the IS estimator, $\mathbf{I}_n$, converge weakly to μ almost surely, i.e.*

$$\mathbb{P}(\{\omega \in \Omega \ : \ \mathbf{I}_n(\omega) \implies \mu\}) = 1.$$

Proof. Let $f \in B(\mathcal{X})$, then

$$\int_{\mathcal{X}} f \, \mathrm{d}\mathbf{I}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} f\big(Y_i(\omega)\big)\rho\big(Y_i(\omega)\big)\delta_{Y_i(\omega)}.$$

The strong law of large numbers (Theorem A.4) implies that

$$\lim_{n\to\infty} \int_{\mathcal{X}} f \, \mathrm{d}\mathbf{I}_n(\omega) = \int_{\mathcal{X}} f\rho \, \mathrm{d}\pi = \int_{\mathcal{X}} f \, \mathrm{d}\mu\,, \quad \text{a.s.} \tag{38}$$

Since $\mathcal{X}$ is separable we may choose a totally bounded metric d on $\mathcal{X}$, by Theorem 3.5. In this metric, Theorem 3.12 implies that there exists a countable dense subset $\{f_k\}$ of $U_b(\mathcal{X}, d)$ such that $\mathbf{I}_n(\omega) \implies \mu$ if

$$\lim_{n\to\infty} \int_{\mathcal{X}} f_k \, \mathrm{d}\mathbf{I}_n(\omega) = \int_{\mathcal{X}} f_k \, \mathrm{d}\mu\,, \quad \text{for every } k \in \mathbb{N}. \tag{39}$$

Equation (38) implies that for each $k \in \mathbb{N}$ there exists a null set $N_k \subset \Omega$ such that $\int_{\mathcal{X}} f_k \, \mathrm{d}\mathbf{I}_n(\omega) \to \int_{\mathcal{X}} f_k \, \mathrm{d}\mu$ on N_k. Define $N = \bigcup N_k$, then $\mathbb{P}(N) = 0$ and (39) holds for every $\omega \in \Omega \setminus N$, which implies that $\mathbf{I}_n \implies \mu$ almost surely on Ω.

■

3.5 Measurability and Continuity in the Weak- and τ-Topologies

In this section we go more into detail about the relationships between the τ-topology and the topology of weak convergence on $\mathbf{M}(\mathcal{X})$. We write $\mathcal{T}_\tau$ to denote the τ-topology and $\mathcal{T}_w$ to denote the topology of weak convergence. Since $C_b(\mathcal{X}) \subset B(\mathcal{X})$, the definition of a weak topology implies that $\mathcal{T}_\tau$ is a stronger topology than $\mathcal{T}_w$ on $\mathbf{M}(\mathcal{X})$, i.e.

$$\mathcal{T}_w \subset \mathcal{T}_\tau.$$

Let $(\mathcal{Y}, \mathcal{T}_\mathcal{Y})$ be a topological space with Borel $\sigma-$algebra $\mathcal{B}_\mathcal{Y}$. Some of the immediate consequences of the fact that $\mathcal{T}_\tau$ is stronger than $\mathcal{T}_w$ are:

1. If $\Phi : \mathcal{Y} \to \mathbf{M}(\mathcal{X})$ is τ-continuous, then it is weakly continuous.
2. If $\Psi : \mathbf{M}(\mathcal{X}) \to \mathcal{Y}$ is weakly continuous, then it is τ-continuous.
3. If $K \subset \mathbf{M}(\mathcal{X})$ is τ-compact, then K is compact in the weak topology.

Hence there are more continuous maps from $\mathcal{Y}$ to $\mathbf{M}(\mathcal{X})$ in the weak topology than in the τ-topology, and conversely there are fewer continuous maps from $\mathbf{M}(\mathcal{X})$ to $\mathcal{Y}$ in the weak topology than in the τ-topology. An example of this is given by the maps $x \mapsto \delta_x$, which we have seen are continuous maps when $\mathbf{M}(\mathcal{X})$ is equipped with the weak topology, but in general not continuous in the $\tau-$topology. The maps $x \mapsto \delta_x$ may not even be measurable with respect to the Borel σ-algebra generated by the τ-topology. Hence the empirical distributions are not measurable with respect to this σ-algebra. The general solution to this problem is to consider a smaller σ-algebra on $\mathbf{M}(\mathcal{X})$. Given $f \in B(\mathcal{X})$ we define $\Psi_f : \mathbf{M}(\mathcal{X}) \to \mathbb{R}$ to be the evaluation map

$$\Psi_f(\mu) := \langle f, \mu \rangle = \int_{\mathcal{X}} f \, \mathrm{d}\mu \, .$$

We will consider the σ-algebra generated by evaluation maps Ψ_f.

Definition 3.12. We define the *cylinder σ-algebra* on $\mathbf{M}(\mathcal{X})$ as

$$\mathcal{B}_\Psi := \sigma\left(\Psi_f : f \in B(\mathcal{X})\right). \tag{40}$$

Let $\mathcal{B}_w$ denote the Borel σ-algebra on $\mathbf{M}(\mathcal{X})$ equipped with the weak topology and $\mathcal{B}_\tau$ denote the Borel σ-algebra on $\mathbf{M}(\mathcal{X})$ with the τ-topology. This notation is not standard and some authors refer to $\mathcal{B}_\Psi$ as the cylinder $\sigma-$algebra which can be somewhat confusing, considering that the term cylinder $\sigma-$algebra is widely used when working with products of measure spaces. The fact that the τ-topology is stronger than the weak topology implies that

$$\mathcal{B}_w \subset \mathcal{B}_\tau.$$

The next Lemma shows that working with $\mathcal{B}_\Psi$ instead of $\mathcal{B}_\tau$ can solve the measurability issues that arise when working with $\mathcal{B}_\tau$, because the maps $x \mapsto \delta_x$ are $\mathcal{B}_\Psi$-measurable.

Lemma 3.9. *The map $x \mapsto \delta_x$ is $\mathcal{B}_\Psi$-measurable.*

Proof. Since $\mathcal{B}_\Psi$ is generated by the maps Ψ_f it follows from Lemma A.2 that φ is $\mathcal{B}_\Psi$-measurable if

$$\varphi^{-1}\left(\bigcup_{f\in B(\mathcal{X})} {\Psi_f}^{-1}(\mathcal{B}_\mathbb{R})\right) \subset \mathcal{B}_\mathcal{X}. \tag{41}$$

Let $\varphi : \mathcal{X} \to \mathbf{M}(\mathcal{X})$ be the map defined by $x \mapsto \delta_x$. For every $f \in B(\mathcal{X})$ it holds that

$$f = \Psi_f \circ \varphi.$$

By the measurability of f it follows that

$$\varphi^{-1}(M) \in \mathcal{B}_\mathcal{X}, \quad \text{for every } M \in {\Psi_f}^{-1}(\mathcal{B}_\mathbb{R}).$$

Hence, the inclusion in equation (41) holds.

■

The measurability of the map $x \mapsto \delta_x$ implies that the empirical distributions of the CMC and IS estimators are $\mathcal{B}_\Psi$-measurable. Therefore, when studying the large deviations of a sequence of measures on $\mathbf{M}_1(\mathcal{X})$ equipped with the τ-topology, the $\sigma-$algebra $\mathcal{B}_\Psi$ is often used. If the underlying space $\mathcal{X}$ has enough topological structure this is not very restricting. In fact, in separable metric spaces it actually holds that $\mathcal{B}_\Psi = \mathcal{B}_w$. The following result follows directly from [9, Proposition 7.25].

Lemma 3.10. *Let $\mathcal{X}$ be a separable metrizable space, then $\mathcal{B}_\Psi = \mathcal{B}_w$ on $\mathbf{M}(\mathcal{X})$.*

When $\mathcal{X}$ is a Polish space the above result is often attributed to [15, Lemma 2.1] in the large deviations literature. However, it follows trivially from [9, Proposition 7.25] and the first edition of [9] was published more than 10 years before [15].

3.6 Relative Entropy

In this section we introduce the *relative entropy*, also known as *Kullback-Leibler divergence (KL-divergence)*, which occurs as a rate function in the theory of large deviations. We will make use of many of the results presented here in the next chapter. The relative entropy is also widely used in information theory. We give the definition of relative entropy below.

Definition 3.13. The *relative entropy* is a mapping $\mathbf{R}(\cdot|\cdot)$ from $\mathbf{M}_1(\mathcal{X}) \times \mathbf{M}_1(\mathcal{X})$ to $[0,\infty]$ defined as

$$\mathbf{R}(\mu|\pi) := \begin{cases} \int_\mathcal{X} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \mathrm{d}\mu = \int_\mathcal{X} \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \mathrm{d}\pi\,, & \text{if } \mu \ll \pi, \\ \infty, & \text{otherwise.} \end{cases} \tag{42}$$

The relative entropy can be thought of as a measure of the similarity of two probability measures in $\mathbf{M}_1(\mathcal{X})$. However, the relative entropy is not a metric; it is trivial to see from the definition of relative entropy that it is not symmetric with respect to its arguments. The relative entropy has several very useful properties which may not be obvious from the definition. The following result gives a variational representation of the relative entropy and it is known as the *Donsker-Varadhan variational formula*. The original result by Donsker and Varadhan [26, Lemma 2.1] was proven for a smaller class of nonnegative bounded continuous functions defined on a metric space. The version of the variational formula we state here can be found in [30, Lemma 1.4.3 (a)].

Theorem 3.18 (Donsker-Varadhan variational formula)**.** *Let $\mathcal{X}$ be a Polish space. Then, for every $\mu, \pi \in \mathbf{M}_1(\mathcal{X})$, it holds that*

$$\begin{aligned}\mathbf{R}(\mu|\pi) &= \sup_{g \in C_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g \, \mathrm{d}\mu - \log \left[\int_{\mathcal{X}} e^g \, \mathrm{d}\pi \right] \right\} \\ &= \sup_{g \in B(\mathcal{X})} \left\{ \int_{\mathcal{X}} g \, \mathrm{d}\mu - \log \left[\int_{\mathcal{X}} e^g \, \mathrm{d}\pi \right] \right\}\end{aligned}$$

In the next chapter, we will see that there are several criteria that good rate functions in the theory of large deviations must satisfy. The next lemma, which was originally proven for the relative entropy in [2] and extended to hold more generally in [24], shows that a large class of functions on $\mathbf{M}(\mathcal{X})$ satisfy these properties. The term *level set* of a function will be used frequently from now on, and we will use the following definition of a level set, also known as a sublevel set.

Definition 3.14. Let $\mathcal{X}$ be a set and $\varphi : \mathcal{X} \to [0, \infty]$, then the sets

$$\{x \in \mathcal{X} : \varphi(x) \leq t\}, \quad t \in [0, \infty],$$

are said to be the *level sets* of φ.

Lemma 3.11 ([24, Lemma 6.2.16])**.** *Let $\varphi : \mathbb{R} \to [0, \infty]$ be a nonnegative convex lower semicontinuous function with compact level sets that satisfies*

$$\frac{|\varphi(x)|}{|x|} \to \infty, \quad \text{as } |x| \to \infty. \tag{43}$$

If $\pi \in \mathbf{M}_+(\mathcal{X})$, then the function $I_\varphi : \mathbf{M}(\mathcal{X}) \to \mathbb{R}$, defined by

$$I_\varphi(\mu) := \begin{cases} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \mathrm{d}\pi \,, & \text{if } \mu \ll \pi, \\ \infty, & \text{otherwise,} \end{cases}$$

is a nonnegative lower semicontinuous function with compact level sets on $\mathbf{M}(\mathcal{X})$ equipped with the τ-topology.

The condition in equation (43) may seem strange, but is there to ensure uniform integrability. The proof of Lemma 3.11 uses several deep results from functional analysis and we have chosen to not include it here. The full proof can be found in [24, page 266]. Note that if $\varphi(x) = x \log x$, then $I_\varphi(\mu) = \mathbf{R}(\mu|\pi)$, and we get the following corollary of Lemma 3.11, which we will use in the next chapter.

Corollary 3.2. *Let $\pi \in \mathbf{M}_1(\mathcal{X})$, then the function $R(\cdot|\pi)$ is a nonnegative lower semicontinuous function with compact levels sets on $\mathbf{M}(\mathcal{X})$ equipped with the τ-topology.*

4 Large Deviations Theory

This chapter is concerned with the study of the large deviation principle. The focus is on large deviations results that can be applied to analyze the performance of the empirical distributions of the CMC and IS estimator. We will introduce Cramér's & Sanov's Theorems and version of Sanov's Theorem that holds for the empirical distributions of the IS estimators due to Hult & Nyquist [33]. The methods used and the exposition of this chapter is greatly inspired by the book by Dembo & Zeitouni [24], the articles by de Acosta [2], [3], [1], and the article by Hult & Nyquist [33].

The outline of this chapter is as follows. In section 4.1 we give the basic definitions related to the theory of large deviations. This is followed by some existence and uniqueness result in section 4.2, and sub-additivity techniques developed by Ruelle [46] and Lanford [38] are introduced. In section 4.3 we show how large deviation principle can be moved between spaces by different transformations. We state the contraction principle which motivates the rate function in Sanov's Theorem for the empirical distributions of the IS estimator. In section 4.4 we prove Cramér's Theorem in Polish spaces and this is combined in section 4.5 with projective systems to prove the classical version of Sanov's Theorem in Polish spaces and the space of measures equipped with τ-topology. We end this chapter with some example applications of how the large deviations principle can be applied to analyze the CMC and the IS estimator.

4.1 Definition and Basic Properties

Definition 4.1. A function $f : \mathcal{X} \to [0, \infty]$ is said to be a *rate function* if it is lower semicontinuous. Furthermore, if I is a rate function and the level sets

$$\{x \in \mathcal{X} \ : \ f(x) \leq t\}$$

are compact for every $t \in [0, \infty]$ then I is said to be a *good rate function.*

Definition 4.2. Let $\mathcal{X}$ be a topological space, $\mathcal{B}$ be a σ-algebra on $\mathcal{X}$ containing all open sets, and I a rate function on $\mathcal{X}$. Then a sequence of probability measures (μ_n) is said to satisfy the *large deviation principle* with rate function I if

$$-\inf_{U} I \leq \liminf_{n \to 0} \frac{1}{n} \log [\mu_n(U)], \tag{44}$$

for every open $U \subset \mathcal{X}$, and

$$-\inf_{C} I \geq \limsup_{n \to 0} \frac{1}{n} \log [\mu_n(C)], \tag{45}$$

for every closed set C.

Equation (46) is referred to the *large deviation lower bound* and equation (47) is referred to as the *large deviation upper bound.* In practice the full term large deviations principle is commonly replaced by the acronym *LDP* (see e.g. [25] or [24]). Note the similarity between the the large deviations bounds and the third and fourth statement in the Portmanteau Theorem. An equivalent formulation of the LDP that follows directly from the definition is given below.

Lemma 4.1. *Let $\mathcal{X}$ be a topological space and I a rate function on $\mathcal{X}$. Then a sequence of probability measures (μ_n) satisfies the large deviation principle with rate I if and only if the following two equations hold for every $A \in \mathcal{B}_{\mathcal{X}}$,*

$$-\inf_{A^\circ} I \leq \liminf_{n\to\infty} \frac{1}{n} \log\left[\mu_n(A)\right], \tag{46}$$

$$-\inf_{\overline{A}} I \geq \limsup_{n\to\infty} \frac{1}{n} \log\left[\mu_n(A)\right]. \tag{47}$$

Especially, if $\inf_{A^\circ} I = \inf_{\overline{A}} I$, then the limit

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\mu_n(A)\right]$$

exists.

Note that we assume that the Borel $\sigma-$algebra on $\mathcal{X}$ is contained in the σ-algebra $\mathcal{B}$. It is possible to use the formulation given in Lemma 4.1 to define the large deviation principle when $\mathcal{B}_{\mathcal{X}} \not\subset \mathcal{B}$, but we will work under the assumption that $\mathcal{B}_{\mathcal{X}} \subset \mathcal{B}$ and from now on it is always assumed that this condition is met. If $\mathcal{X}$ is a regular topological space and a sequence of probability measures satisfy a large deviation principle on $\mathcal{X}$ with rate function I, then the sequence cannot satisfy the large deviation principle with another rate function (see e.g. [24, Lemma 4.1.4]). If we replace the closed sets in the upper bound with compact sets we get the definition of the weak large deviation principle.

Definition 4.3. Let $\mathcal{X}$ be a topological space and I a rate function on $\mathcal{X}$. Then a family of probability measures $\{\mu_n\}$ is said to satisfy *a weak large deviation principle* with rate function I if it satisfies the lower bound (46) and

$$-\inf_{K} I \geq \limsup_{n\to\infty} \frac{1}{n} \log\left[\mu_n(K)\right] \tag{48}$$

for every compact set $K \subset \mathcal{X}$.

It is generally much easier to prove the weak upper bound than the regular upper bound. However, when (μ_n) satisfy a weak large deviation principle then the rate function may not be unique[9]. The strengthening of a weak large deviation principle to the full large deviation principle is often achieved by showing that the sequence (μ_n) is exponentially tight.

Definition 4.4. A sequence of probability measures (μ_n) is *exponentially tight* if for every $\alpha > 0$ there exists a compact set K such that

$$\limsup_{n\to\infty} \frac{1}{n} \log\left[\mu_n(\mathcal{X} \setminus K)\right] < -\alpha$$

If the sequence satisfy a weak large deviation principle and is exponentially tight, then the following Lemma can be used to gain a full large deviation principle (see e.g. [25, Lemma 2.1.5] or [24, Lemma 1.2.18]).

Lemma 4.2. *Let (μ_n) satisfy a weak large deviation principle on $\mathcal{X}$ with rate I. If the sequence (μ_n) is exponentially tight, then I is a good rate function and (μ_n) satisfy the full large deviation principle with rate I.*

[9] This explains the usage of the definite article *the* in conjunction with the full large deviation principle and the usage of the indefinite article *a* in conjunction with the weak large deviation principle.

We summarize the technique described above below.

It is a commonly used technique in large deviations to first that prove that (μ_n) satisfy a weak large deviation principle, and then show that (μ_n) safisfy the full large deviation principle by showing that the sequence (μ_n) is exponentially tight.

Before we move on, it is worth noting that it is possible to work with arbitrary nets of measures instead of sequences. The most common alternative is to replace the sequence (μ_n) with a net of measures indexed by $\varepsilon \to 0$. Then, the upper and lower bounds become

$$\begin{aligned} -\inf_{A^o} I &\leq \liminf_{\varepsilon \to 0} \varepsilon \log\left[\mu_\varepsilon(A)\right], \\ -\inf_{\overline{A}} I &\geq \limsup_{\varepsilon \to 0} \varepsilon \log\left[\mu_\varepsilon(A)\right]. \end{aligned}$$

We will only present large deviation results for sequences of measures in this chapter, but much of the theory also holds for more arbitrary nets. Especially, apart from the subsection on sub-additivity all of the theory in section 4.2 hold for when (μ_n) is replaced by (μ_ε) (see [24]). The same is true for the results in section 4.4 on transformations of large deviation principles and projective systems.

4.2 Existence of a LDP

In this section we introduce two results which are useful for proving the existence of a large deviation principle in topological spaces which can be found in chapter 4 of [24]. The first result, Theorem 4.1, gives a criterion for the existence of a large deviation principle in topological spaces based on the existence of large deviations limits over all basis elements. The second result, Theorem 4.2, provides a criterion for the convexity of the rate function in topological linear spaces.

Definition 4.5. We define the *lower large deviation limit* of a sequence (μ_n) of probability measures as the function $\underline{L} : \mathcal{B} \to [0, \infty]$ defined by

$$\underline{L}(A) := -\liminf_{n \to \infty} \frac{1}{n} \log\left[\mu_n(A)\right],$$

and the *upper large deviation limit* as

$$\overline{L}(A) := -\limsup_{n \to \infty} \frac{1}{n} \log\left[\mu_n(A)\right],$$

If the two limits agree we say that $L = \overline{L} = \underline{L}$ is the *large deviation limit* of μ_n. The name upper and lower large deviation limit is non-standard and they are not to be confused with the upper and lower large deviation bound.

Theorem 4.1 (Topological Existence Theorem [24, Thm 4.1.11])**.** *Let $\mathcal{X}$ be a topological space and $\mathcal{U}$ be a base for the topology on $\mathcal{X}$ and assume that $\mathcal{B}_\mathcal{X} \subset \mathcal{B}$. Define the upper and lower rate functions taking values in $[0, \infty]$ by*

$$\begin{aligned} \underline{I}(x) &:= \sup\{\, \underline{L}(B) \;:\; B \subset \mathcal{U},\ x \in B\}, \\ \overline{I}(x) &:= \sup\{\, \overline{L}(B) \;:\; B \subset \mathcal{U},\ x \in B\}. \end{aligned} \tag{49}$$

If $\overline{I} = \underline{I}$ on $\mathcal{X}$, then $(\mu_n)_n$ satisfies a weak large deviation principle with rate function

$$I = \underline{I} = \overline{I}.$$

Theorem 4.1 is very useful in combination with the following Lemma that can be used to show that the rate function is convex. If one can show that the rate function is convex, then tools from convex analysis may be used to identify the rate function with the limit of the Legendre-Fenchel transform of the logarithmic moment generating function of the measures μ_n. We will give such a result later in this section.

Theorem 4.2 ([24, Lemma 4.1.21]). *Let $\mathcal{X}$ be a topological linear space and $\mathcal{U}$ be a base for the topology on $\mathcal{X}$ and assume that $\mathcal{B}_{\mathcal{X}} \subset \mathcal{B}$. If $\underline{I} = \overline{I}$ and $\overline{L}$ satisfy*

$$\overline{L}\left(\frac{B_1 + B_2}{2}\right) \leq \frac{1}{2}\big(\overline{L}(B_1) + \overline{L}(B_2)\big), \quad \text{for every } B_1, B_2 \in \mathcal{U}, \tag{50}$$

then the rate function I is convex.

In order to utilize Theorem 4.1 one must show that the upper and lower rate functions agree on the full space $\mathcal{X}$, which is the case if

$$\overline{L}(B) = \underline{L}(B) = \lim_{n\to\infty} \frac{1}{n} \log\big[\mu_n(B)\big],$$

for every $B \in \mathcal{U}(\mathcal{T})$. In the next section we show how sub-additivity can be used to prove the existence of the limit above.

Sub-Additivity

Definition 4.6. A function $f : \mathbb{N} \to [0, \infty]$ is *sub-additive* if

$$f(m + n) \leq f(m) + f(n), \quad \text{for every } m, n \in \mathbb{N},$$

and *super-additive* if

$$f(m) + f(n) \leq f(m + n), \quad \text{for every } m, n \in \mathbb{N},$$

The following property of sub-additive functions follows directly from the definition.

Lemma 4.3. *Let f be a sub-additive function and $m, n \in \mathbb{N}$, then*

$$f(mn) \leq mf(n), \text{ and } \quad f(mn) \leq nf(m).$$

The main application of sub-additivity to the theory of large deviation is to prove the existence of the large deviation limit L. The usefulness comes from the following result which can be found in [25, Lemma 3.1.3][10] and [24, Lemma 6.1.11].

Lemma 4.4. *Let $f : \mathbb{N} \to [0, \infty]$ be a sub-additive function. If there exists $N \in \mathbb{N}$ such that $f(n) < \infty$ for every $n \geq N$, then*

$$\lim_{n\to\infty} \frac{f(n)}{n} = \inf_{n\geq N} \frac{f(n)}{n} < \infty.$$

Proof. Let f be sub-additive and $n \geq m \geq N$. Then we may express n as

$$n = n - m\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right) + m\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right).$$

[10] The proof in [25] contains an error, but the general idea is correct.

Therefore, by sub-additivity and Lemma 4.3 it follows that

$$f(n) \leq f\left(n - m\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right)\right) + f(m)\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right).$$

The product of the second term satisfy the inequalities

$$\frac{1}{m} - \frac{2}{n} \leq \frac{\frac{n}{m}(m-2)}{mn} \leq \frac{\left\lfloor \frac{n}{m} \right\rfloor - 1}{n} \leq \frac{n-m}{nm} = \frac{1}{m} - \frac{1}{n}.$$

And since m is fixed it follows that

$$\lim_{n\to\infty} \frac{f(m)}{n}\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right) = \frac{f(m)}{m}.$$

We shall now give a bound for the first term. We may express n as

$$n = m\left\lfloor \frac{n}{m} \right\rfloor + r.$$

for some $0 \leq r < m$, which shows that

$$m \leq n - m\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right) \leq 2m - 1.$$

Let $M := \max\{f(k) \ : \ m \leq k \leq 2m-1\}$, then

$$f\left(n - m\left(\left\lfloor \frac{n}{m} \right\rfloor - 1\right)\right) \leq M,$$

Especially, we get that

$$\limsup_{n\to\infty} \frac{f(n)}{n} \leq \frac{f(m)}{m}, \quad \text{for every } m \geq N.$$

At the same time

$$\liminf_{n\to\infty} \frac{f(n)}{n} = \sup_{k\geq 1} \inf_{m\geq k} \frac{f(m)}{m} \geq \inf_{m\geq N} \frac{f(m)}{m}.$$

By combining the last two inequalities the Lemma follows.

■

Lemma 4.4 combined with Theorem 4.1 can be used to prove the existence of a large deviation principle for a sequence, (μ_n), of measures. If one can show that for every basis element, B, the function $f(n) := -\log[\mu_n(B)]$ is sub-additive and bounded for n greater than some $N \in \mathbb{N}$, then Lemma 4.4 implies that $L = \underline{L} = \overline{L}$. Hence, the assumptions in Theorem 4.1 are satisfied and the sequence of measure satisfy the large deviation principle. The application of sub-additivity to the theory of large deviations was largely driven by its applications to statistical mechanics and is generally attributed to Lanford [38] and Ruelle [46].

Convexity

We have introduced methods to prove the existence of a large deviation principle and Theorem 4.1 gives an expression for the rate function as a supremum of the large deviation limits. Furthermore, if $\mathcal{X}$ is a topological linear space and the conditions in Theorem 4.2 are satisfied, then this rate function is convex. Using tools from convex analysis it is many cases possible to get a better representation for the rate function, and in this subsection

we present some of the main results from this theory. This section is largely based on [24, Section 4.5], and also [25, Chapter 2]. In the language of large deviations practitioners the derivation of *nicer* expressions of I is called *identification of the rate function.* Throughout this subsection $\mathcal{X}$ is assumed to be a Hausdorff locally convex topological linear space, and we use $\mathcal{X}^*$ to denote the topological dual space of $\mathcal{X}$.

Definition 4.7. Let $\mu \in \mathbf{M}_1(\mathcal{X})$, then we define the *logarithmic moment generating function* of μ to be the function $\Lambda_\mu : \mathcal{X}^* \to (-\infty, \infty]$ given by

$$\Lambda_\mu(\lambda) := \log\left[\int_{\mathcal{X}} e^{\langle \lambda, x\rangle}\, \mathrm{d}\mu(x)\right].$$

Note that if X is an $\mathcal{X}$-valued random variable with distribution μ, then

$$\Lambda_\mu(\lambda) = \log \mathbb{E}\left[e^{\langle \lambda, X\rangle}.\right]$$

When X is a real valued random variable we usually write $\Lambda_X(s)$ in stead of $\Lambda_\mu(\lambda)$, and as the name suggests this is simply the logarithm of the moment generating function of X:

$$\Lambda_X(s) = \log M_X(s) = \log\left[\int_{\mathbb{R}} e^{sx}\, \mathrm{d}\mu(x)\right].$$

Lemma 4.5. *For any $\mu \in \mathbf{M}_1(\mathcal{X})$ the function Λ_μ is convex, and for fixed $\lambda \in \mathcal{X}^*$ the function $\Lambda_\mu(\lambda t) : \mathbb{R} \to \mathbb{R}$ is lower semicontinuous.*

Proof. Let $t \in [0,1]$ and $\lambda_1, \lambda_2 \in \mathcal{X}^*$, then

$$\begin{aligned}\Lambda_\mu(t\lambda_1 + (1-t)\lambda_2) &= \log\left[\int_{\mathcal{X}} e^{\langle t\lambda_1 + (1-t)\lambda_2, x\rangle}\, \mathrm{d}\mu(x)\right] \\ &= \log\left[\int_{\mathcal{X}} e^{t\langle \lambda_1, x\rangle} e^{(1-t)\langle \lambda_2, x\rangle}\, \mathrm{d}\mu(x)\right].\end{aligned}$$

Hence, it follows from Hölder's inequality, applied with Hölder conjugates $1/t$ and $1/(1-t)$, that

$$\begin{aligned}\Lambda_\mu(t\lambda_1 + (1-t)\lambda_2) &\le \log\left[\left(\int_{\mathcal{X}} e^{\langle \lambda_1, x\rangle}\, \mathrm{d}\mu(x)\right)^t \left(\int_{\mathcal{X}} e^{\langle \lambda_1, x\rangle}\, \mathrm{d}\mu(x)\right)^{1-t}\right] \\ &= t\Lambda_\mu(\lambda_1) + (1-t)\Lambda_\mu(\lambda_2).\end{aligned}$$

Which shows that Λ_μ is convex. Next, let $t \in \mathbb{R}$ and (t_n) be a sequence that converges to t. Then, it follows from the continuity of λ that

$$\lim_{n\to\infty} e^{t_n\langle \lambda, x\rangle} = e^{t\langle \lambda, x\rangle}.$$

Thus, Fatou's Lemma implies

$$\Lambda_\mu(\lambda t) = \int_{\mathcal{X}} e^{t\langle \lambda, x\rangle}\, \mathrm{d}\mu(x) \le \liminf_{n\to\infty} \int_{\mathcal{X}} e^{t_n\langle \lambda, x\rangle}\, \mathrm{d}\mu(x) = \liminf_{n\to\infty} \Lambda_\mu(\lambda t_n),$$

which proves that $\Lambda_\mu(\lambda t)$ is lower semicontinuous with respect to t.

■

The logarithmic moment generating play an important role in the theory of large deviations as it is closely related to the rate function I. In the case of Cramér's theorem, which we state in the next section, the rate function is given by the Legendre-Fenchel transform of the moment generating function.

Definition 4.8. Let $f : \mathcal{X} \to [-\infty, \infty]$, then the *Legendre-Fenchel transform*[11] of f is the function $f : \mathcal{X}^* \to [-\infty, \infty]$ defined by

$$\begin{aligned} f^*(\lambda) &= \sup\{\langle \lambda, x\rangle - f(x) \ : \ x \in \mathcal{X}\} \\ &= -\inf\{f(x) - \langle \lambda, x\rangle \ : \ x \in \mathcal{X}\}. \end{aligned}$$

For many common distributions the logarithmic moment generating function exists and is known. Furthermore, if X is a real valued random variable, then the Legendre-Fenchel transform can be found by solving the continuous 1-dimensional optimization problem

$$\Lambda_X^*(x) = \sup_{s\in\mathbb{R}}\{sx - \Lambda_X(s).\}$$

This is easily solved by taking the derivative the logarithmic moment generating function twice (see e.g. [7]).

Example 3. Let X be a $N(\theta, \sigma^2)$ real valued random variable, then the logarithmic moment generating function of X is given by

$$\Lambda(s) = s\theta + \frac{s^2\sigma^2}{2}.$$

The expression for the Legendre transform is given by

$$\Lambda^*(x) = \frac{(x-\theta)^2}{2\sigma^2},$$

which is the exponent appearing in the probability density function of X scaled by -1.

Example 4. Let X be a $\mathrm{Po}(\lambda)$ random variable, then the logarithmic moment generating function of X is given by

$$\Lambda(s) = \lambda(e^s - 1).$$

The expression for the Legendre transform is given by

$$\Lambda^*(x) = \lambda - x + x\log\left(\frac{x}{\lambda}\right)$$

There are two important results that play an essential role in identifying the rate function as the Legendre-Fenchel transform of the logarithmic moment generating function. The first is a well known result in the theory of large deviations from [52] which is known as *Varadhan's Lemma.* The version we present below can be found in [48, Theorem 2.1.10] and [24, Theorem 4.3.1]

Theorem 4.3 (Varadhan's Lemma)**.** *Let $\mathcal{X}$ be a regular topological space and (μ_n) satisfy the large deviation principle with good rate function I. If $f \in C_b(\mathcal{X})$ satisfies either*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log\left[\int_{f\geq M} e^{nf} \, \mathrm{d}\mu_n\right] = -\infty, \tag{51}$$

[11] Also known as the *convex conjugate* of f or simply the *Legendre transform* of f.

or

$$\limsup_{n\to\infty} \frac{1}{n} \log\left[\int_{\mathcal{X}} e^{\alpha n f} \, \mathrm{d}\mu_n\right] < \infty, \tag{52}$$

for some $\alpha > 1$*, then*

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\int_{\mathcal{X}} e^{nf} \, \mathrm{d}\mu_n\right] = \sup_{x\in\mathcal{X}}\{f(x) - I(x)\}. \tag{53}$$

The right hand side of equation (53) is the the Legendre-Fenchel transform of I. Furthermore, since any $\lambda \in \mathcal{X}^*$ is an element of $C_b(\mathcal{X})$ we can apply Varadhan's Lemma to the function $f(x) = \langle \lambda, x\rangle$. Then, we get the following corollary.

Corollary 4.1. *Let* $\mathcal{X}$ *be a regular topological space and* (μ_n) *satisfy the large deviation principle with good rate function* I*. If* $\lambda \in \mathcal{X}^*$ *and there exists* $\alpha > 1$ *such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log\left[\int_{\mathcal{X}} e^{\alpha n\langle \lambda, x\rangle} \, \mathrm{d}\mu_n\right] < \infty,$$

then

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\int_{\mathcal{X}} e^{n\langle \lambda, x\rangle} \, \mathrm{d}\mu_n\right] = \sup_{x\in\mathcal{X}}\{\langle \lambda, x\rangle - I(x)\} = I^*(x). \tag{54}$$

If the limit in equation (54) exists and is finite, then we define

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{n} \Lambda_{\mu_n}(n\lambda) = \lim_{n\to\infty} \frac{1}{n} \log\left[\int_{\mathcal{X}} e^{n\langle \lambda, x\rangle} \, \mathrm{d}\mu_n(x)\right]. \tag{55}$$

Note, that if $\Lambda(\lambda)$ is finite for every $\lambda \in \mathcal{X}^*$, then we may consider Λ as a function from $\mathcal{X}^*$ to $\mathbb{R}$. The next theorem, which is a well known result from convex analysis can be applied to Λ to identify the rate function.

Theorem 4.4 (Biconjugate Theorem, see e.g. [8, Theorem 2.22])**.** *Let* $f : \mathcal{X} \to (-\infty, \infty]$ *not be identically* ∞*, then* $f = f^{**}$ *if and only if* f *is convex and lower semicontinuous.*

If I is a good convex rate function, then the biconjugate Theorem implies that $I = I^{**}$. If the conditions of Varadhan's Lemma hold for every $\lambda \in \mathcal{X}^*$, then it follows from Corollary 4.1 that $I^{**} = \Lambda^*$. This proves the following theorem.

Theorem 4.5. *Let* $\mathcal{X}$ *be a regular topological space and* (μ_n) *satisfy the large deviation principle with good convex rate function* I*. If* $\Lambda(\lambda)$ *exists and is finite for every* $\lambda \in \mathcal{X}^*$*, then*

$$I(x) = \Lambda^*(x).$$

There are many variations and results which are similar to Theorem 4.5 and that can be used to identify the rate function with Λ^* under different conditions. However, the main idea underlying these results is the variational formula given in Varadhan's Lemma and the biconjugate Theorem. The next theorem, which we state without proof, can be used to identify the rate function associated with a weak large deviation principle. We will use this to identify the rate function in Cramér's Theorem.

Theorem 4.6 ([24, Theorem 4.5.14])**.** *Let* (μ_n) *satisfy a weak large deviation principle with convex rate function* I *on* $\mathcal{X}$*, and the limits*

$$\Lambda_t(\lambda) := \lim_{n\to\infty} \frac{1}{n} \Lambda_{\mu_n}(\lambda n t) \in [-\infty, \infty]$$

exist for every $\lambda \in \mathcal{X}^*$ and are lower semicontinuous on $\mathbb{R}$ with respect to t. If it holds for every $\lambda \in \mathcal{X}^*$ and $\alpha \in \mathbb{R}$ that

$$\inf\{I(x) : x \in \mathcal{X},\ \langle \lambda, x\rangle - \alpha > 0\} \leq \inf_{s>\alpha} \Lambda_\lambda^*(s), \tag{56}$$

then

$$I(x) = \Lambda^*(x).$$

4.3 Cramér's Theorem

In this section we present one of the most well known results from the theory of large deviations Cramér's Theorem. We will prove a general version of Cramér's Theorem for random variables taking values in Polish spaces. Given a sequence (X_i) of i.i.d. real valued random variables, we can form the empirical means

$$\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The strong law of large numbers implies that $\mathbf{S}_n \to \mathbb{E}[X_i]$ almost surely if the expected value of the random variables X_i is finite. The classical version Cramér's Theorem establishes a large deviation principle for the distributions of the random variables $\mathbf{S}_n$ based on the criteria that the random variables have finite moment generating functions.

Theorem 4.7 (Cramér's Theorem)**.** *Let (X_i) be a sequence of i.i.d. real valued random variables with distribution μ, and let μ_n denote the distribution of the empirical mean $\mathbf{S}_n$. If $M_X(s) < \infty$ for every $s \in \mathbb{R}$, then the sequence (μ_n) satisfies the large deviation principle with rate function*

$$I = \Lambda_X^*(x) = \sup_{s\in\mathbb{R}}\{sx - \Lambda_X(s)\}.$$

Furthermore, for every $\alpha \in \mathbb{R}$ it holds that

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\mu_n[\alpha, \infty)\right] = -\inf_{s\geq\alpha} \Lambda^*(s).$$

Cramér's Theorem can be extended to $\mathbb{R}^d$ for any $d \in \mathbb{N}$ and to more general (even infinite-dimensional) topological linear spaces. The main goal of this section is to prove a version of Cramér's Theorem that holds under the following assumption.

Assumption 1. *The space $\mathcal{X}$ is a locally convex Hausdorff topological linear space and $\mathcal{E} \subset \mathcal{X}$ a closed convex subset satisfying*

1. *$\mu(\mathcal{E}) = 1$.*
2. *$\mathcal{E}$ is Polish in the subspace topology.*

Throughout this section we assume that the criteria in Assumption 1 are satisfied.

Assumption 1 is equivalent to part (a) of [24, Assumption 6.1.2], however we drop part (b) of [24, Assumption 6.1.2]: *$\overline{\text{co}}\, K$ is compact whenever $K \subset \mathcal{E}$ is compact.* The reason for dropping part (b) is that in a completely metrizable locally convex space this always holds by Theorem A.9 (see [5, Thm 5.35]). Therefore part (b) of Assumption 6.1.2. in [24] is redundant as it follows from part (a).

Empirical Means in Topological Linear Spaces

Consider a sequence (X_i) if i.i.d. random variables with distribution μ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the topological linear space $\mathcal{X}$. In this section we study the large deviations of the distribution of the empirical means $\mathbf{S}_n : \Omega \to \mathcal{X}$, defined by

$$\mathbf{S}_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} X_i(\omega).$$

The distribution of the empirical mean is given by $\mu_n = \mathbb{P} \circ \mathbf{S}_n{}^{-1}$. Now consider the random vector $\mathbf{X}_n : \Omega \to \mathcal{X}^n$ defined by

$$\mathbf{X}_n = (X_1, \cdots, X_n),$$

which is a random variable with distribution $\mu^n = \mu \otimes \cdots \otimes \mu$ on $\left(\mathcal{X}^n, \mathcal{B}^n(\mathcal{X})\right)$. Then we may express the empirical means as $\mathbf{S}_n = S_n \circ \mathbf{X}_n$, where $S_n : \mathcal{X}^n \to \mathcal{X}$ is the map defined by

$$S_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Since $\mathcal{X}$ is a topological linear space it follows that the map $S_n : \mathcal{X}^n \to \mathcal{X}$ is continuous and therefore it is Borel-measurable. Thus, by Lemma A.7 it follows that the distribution μ_n of $\mathbf{S}_n$ equals the distribution of the random variable S_n, i.e.

$$\mu_n = \mathbb{P} \circ \mathbf{S}_n{}^{-1} = \mu^n \circ S_n{}^{-1},$$

and we will use both representations. A useful property of the empirical means is that it is possible to express them as convex combinations on the form

$$\mathbf{S}_{m+n} = \frac{m}{m+n} \mathbf{S}_m + \frac{n}{m+n} \frac{1}{n} \sum_{i=m+1}^{m+n} X_i.$$

This type of decomposition will be used often and therefore we introduce the notation

$$\mathbf{S}_k^m := \frac{1}{k-m} \sum_{i=m+1}^{k} X_i,$$

using this we can rewrite the previous equation as

$$\mathbf{S}_{m+n} = \frac{m}{m+n} \mathbf{S}_m + \frac{n}{m+n} \mathbf{S}_{m+n}^m. \tag{57}$$

Subadditivity of μ_n

Lemma 4.6. *Let $A \subset \mathcal{E}$ be Borel measurable and convex and $m, n \in \mathbb{N}$, then*

$$\mu_m(A)\mu_n(A) \leq \mu_{m+n}(A).$$

Proof. Let $A \subset \mathcal{E}$ be a Borel-measurable convex set. For any $m, n \in \mathbb{N}$, equation (57) shows that we may express $\mathbf{S}_{m+n}$ as a convex combination of $\mathbf{S}_m$ and $\mathbf{S}_{m+n}^m$. The set A is convex, thus if

$$y \in \mathbf{S}_m{}^{-1}(A) \cap \mathbf{S}_{m+n}^m{}^{-1}(A),$$

then it follows from the convexity of A and equation (57) that

$$\mathbf{S}_{m+n}\big(\mathbf{S}_m{}^{-1}(A) \cap \mathbf{S}^m_{m+n}{}^{-1}(A)\big) \subset A.$$

Which shows that

$$\{\mathbf{S}_m \in A\} \cap \{\mathbf{S}^m_{m+n} \in A\} \subset \{\mathbf{S}_{m+n} \in A\}. \tag{58}$$

Since the random variables X_i are independent, it follows that $\mathbf{S}_m$ and $\mathbf{S}^m_{m+n}$ are independent, and that

$$\begin{aligned}\mathbb{P}\big(\{\mathbf{S}_m \in A\} \cap \{\mathbf{S}^m_{m+n} \in A\}\big) &= \mathbb{P}\big(\mathbf{S}_m \in A\big)\mathbb{P}\big(\mathbf{S}^m_{m+n} \in A\big)\\ &= \mu^m(S_m \in A)\mu^n(S_n \in A)\\ &= \mu_m(A)\mu_n(A).\end{aligned}$$

Combining this with equation (58) yields the inequality

$$\mu_m(A)\mu_n(A) \leq \mathbb{P}(\mathbf{S}_{m+n} \in A) = \mu_{m+n}(A).$$

■

Corollary 4.2. *Let $A \subset \mathcal{E}$ be Borel measurable and convex, then the function*

$$f(n) := -\log\big[\mu_n(A)\big]$$

is sub-additive.

Proof. The measures μ_n are probability measures, hence $0 \leq \mu_n(A) \leq 1$ for every $n \in \mathbb{N}$. Using Lemma 4.6 and taking logarithms on both sides yields, the inquality

$$0 \leq -\log\big[\mu_m(A)\big] - \log\big[\mu_n(A)\big] \leq -\log\big[\mu_{m+n}(A)\big].$$

Which shows that $f(n) = -\log\big[\mu_n(A)\big]$ is sub-additive.

■

We have showed that (μ_n) are sub-additive, but in order to apply Lemma 4.4 we must show the following Lemma.

Lemma 4.7. *Let $A \subset \mathcal{E}$ be open and assume that there exists $m \in \mathbb{N}$ such that $\mu_m(A) > 0$. Then there exists $N \in \mathbb{N}$ such that*

$$\mu_n(A) > 0, \quad \textit{for every; } n \geq N.$$

Proof. Let $\mu_m(A) > 0$. We are going to show that there exists a point $z \in A$ such that all open neighborhoods of z have positive μ_m-measure. Assume for contradiction that no such point exists, then there exists an open neighborhood U_x of every point in A such that $\mu_m(U_x) = 0$. Since the topology on $\mathcal{E}$ is metrizable it follows from Urysohn's Metrization Theorem that it is second countable, hence Lindelöf[12]. Thus, every open cover of A has a countable subcover, which means that there exists a countable subcollection $\{U_{x_i}\}$ that covers A. Hence,

$$0 < \mu_m(A) \leq \mu\left[\bigcup U_{x_i}\right] \leq \sum \mu_m(U_{x_i}),$$

which implies that $\mu_m(U_{x_i}) > 0$ for some $i \in \mathbb{N}$ which is a contradiction.

[12] A topological space is Lindelöf if every open cover has a countable subcover.

Now, let $z \in A$ be a point satisfying $\mu_m(U) > 0$ for every open neighborhood and consider the function $f : [0,1] \times \mathcal{E} \times \mathcal{E} \to \mathcal{E}$ defined by

$$f(t,x,y) = tx + (1-t)y.$$

Since $\mathcal{X}$ is a topological linear space and $\mathcal{E}$ inherits its topology from $\mathcal{X}$ it follows that the function f is continuous with respect to the product topology on $[0,1] \times \mathcal{E} \times \mathcal{E}$. Furthermore,

$$f(0,z,y) = z \in A, \quad \text{for every } y \in \mathcal{E},$$

and the continuity of f implies that $f^{-1}(A)$ is open. Hence, for every $y \in \mathcal{E}$ there exists an open neighborhood $[0,\varepsilon_y) \times U_y \times V_y$ of $(0,z,y)$ such that

$$f\big([0,\varepsilon_y) \times U_y \times V_y\big) = (1-t)U_y + tV_y \subset A, \quad \text{for every } t \in [0,\varepsilon_y). \tag{59}$$

The sets $\{V_y\}_{y\in K}$ form an open cover of K, which is compact thus, there exists a finite subcover of K. Let $\{V_{y_i}\}_{i=1}^n$ be a finite subcover and define

$$U := \bigcap_{i=1}^n U_{y_i}, \quad \varepsilon := \min_{1\le i\le n} \varepsilon_{y_i}.$$

Then U is an open set, and by equation (59) it follows that

$$(1-t)U + tK \subset A, \quad \text{for every } t \in [0,\varepsilon).$$

Since $\mathcal{X}$ is a locally convex space there is a basis consisting of open convex sets, thus we may find an open and convex subset $C \subset \mathcal{X}$ such that $z \in C \subset U$ and

$$(1-t)C + tK \subset A, \quad \text{for every } t \in [0,\varepsilon). \tag{60}$$

Hence,

$$\mu_n\big((1-t)C + tK\big) \le \mu_n(A), \quad \text{for every } t \in [0,\varepsilon). \tag{61}$$

Let $N = \lceil \frac{1}{\varepsilon} \rceil$ and $n \ge N$, then there exists $q, r \in \mathbb{N}$ such that $1 \le r \le m$ and $n = mq + r$. Note that the decomposition always leaves the term r nonzero, which we will use later. Furthermore we can decompose $\mathbf{S}_n$ as

$$\begin{aligned}\mathbf{S}_n &= \frac{1}{n}\left(mq\mathbf{S}_{mq} + r\mathbf{S}_n^{mq}\right) = \frac{1}{n}\left((n-r)\mathbf{S}_{mq} + r\mathbf{S}_n^{mq}\right)\\ &= \left(1 - \frac{r}{n}\right)\mathbf{S}_{mq} + \frac{r}{n}\mathbf{S}_n^{mq}.\end{aligned}$$

Then $r/n < \varepsilon$ and

$$\left\{\mathbf{S}_{mq} \in C\right\} \cap \left\{\mathbf{S}_n^{mq} \in K\right\} \subset \left\{\mathbf{S}_n \in \left(1 - \frac{r}{n}\right)C + \frac{r}{n}K\right\}.$$

Thus, it follows from the independence of $\mathbf{S}_{mq}$ and $\mathbf{S}_n^{mq}$ that

$$\mu_{mq}(C)\mu_r(K) \le \mu_n\left(\mathbf{S}_n \left(1 - \frac{r}{n}\right)C + \frac{r}{n}K\right),$$

which combined with equation (61) yields the inequality

$$\mu_{mq}(C)\mu_r(K) \le \mu_n(A). \tag{62}$$

Thus, all it remains to show is that both factors on the left-hand side of equation (62) are nonzero. Consider the finite collection $\{\mu_1, \dots, \mu_m\} \subset \mathbf{M}_1(\mathcal{X})$. Any finite collection in a topological space is compact and thus Prokhorov's Theorem (3.15) implies that $\{\mu_1, \dots, \mu_m\}$ is tight, i.e. there exists a compact set $K \subset \mathcal{E}$ such that

$$\mu_i(K) > 0, \quad i = 1, \dots, m. \tag{63}$$

Since, $1 \leq r \leq m$ it follows from equation (63) that $\mu_r(K) > 0$. Next, note that C is an open convex subset of A, and therefore Lemma 4.6 implies that $\mu_{mq}(C) \geq \mu_m(C)^q$.

Finally, C is a neighborhood of z, hence of positive μ_m-measure, which shows that

$$0 < \mu_m(C)^q \mu_r(K) \leq \mu_n(A).$$

■

Cramér's Theorem in Polish spaces

We are now ready to prove that the empirical means satisfy a weak large deviation principle, and that a version of Cramér's Theorem also holds under Assumption 1.

Lemma 4.8 ([24, Lemma 6.1.7 & 6.1.8])**.** *The sequence (μ_n) of distributions of the empirical means satisfy a weak large deviation principle with a convex rate function I satisfying*

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = -\inf_{x\in A} I(x), \tag{64}$$

for every convex and open $A \subset \mathcal{X}$.

Proof. The space $\mathcal{X}$ is a locally convex topological linear space and therefore there exists a basis $\mathcal{C}$ consisting of open convex subsets. Let $A \in \mathcal{C}$, then $A \cap \mathcal{E}$ is convex and

$$\mu_n(A) = \mu_n(A \cap \mathcal{E}), \quad \text{for every } n \in \mathbb{N},$$

since $\mu_n(\mathcal{E}) = \mu_n(\mathcal{X}) = 1$ for every n, Hence, we may without loss of generality assume that $A \subset \mathcal{E}$ is convex and open. Now, we are going to show that the limit $L(A)$ exists for every basis element. If $\mu_n(A) = 0$ for every $n \in \mathbb{N}$, then $L(A) = \infty$, and the limit trivially exists. Otherwise, there exists $m \in \mathbb{N}$ such that $\mu_m(A) > 0$ and by Lemma 4.7 it follows that $\mu_n(A)$ satisfy the assumptions of Lemma 4.6. Hence, Corollary 4.2 implies that $f(n) = -\log[\mu_n(A)]$ is sub-additive, and therefore by Lemma 4.4 the limit $L(A)$ exists and is given by

$$L(A) = \lim_{n\to\infty} -\frac{1}{n} \log\big[\mu_n(A)\big] = -\sup_{n\geq N} \frac{1}{n} \log\big[\mu_n(A)\big].$$

Thus, $L(A)$ exists for every $A \in \mathcal{C}$, and the topological basis existence theorem, Theorem 4.1, yields the existence of a weak large deviation principle for (μ_n) with rate function I given by

$$I(x) := \sup\{L(A) \ : \ x \in A, \ A \in \mathcal{C}\}.$$

Next we show that the rate function I is convex. In order to apply Theorem 4.2 we must prove that L satisfies the condition

$$L\left(\frac{A_1 + A_2}{2}\right) \leq \frac{1}{2}\big(L(A_1) + L(A_2)\big), \quad \text{for every } A_1, A_2 \in \mathcal{C}. \tag{65}$$

Let $A_1, A_2 \in \mathcal{C}$, and assume without loss of generality that $A_1, A_2 \subset \mathcal{E}$. Note that we may decompose $\mathbf{S}_{2n}$ as

$$\frac{1}{2}\left(\mathbf{S}_n + \mathbf{S}_{2n}^n\right) = \mathbf{S}_{2n}.$$

By convexity it follows that

$$\left\{\mathbf{S}_n \in A_1\right\} \cap \left\{\mathbf{S}_{2n}^n \in A_2\right\} \subset \left\{\mathbf{S}_{2n} \in \frac{A_1 + A_2}{2}\right\},$$

and by the independence of $\mathbf{S}_n$ and $\mathbf{S}_{2n}^n$ combined with the inclusion above we get the inequality

$$\mu_n(A_1)\mu_n(A_2) \leq \mu_{2n}\left(\frac{A_1 + A_2}{2}\right).$$

Taking logarithms on both sides yields

$$\log\left[\mu_n(A_1)\right]\log\left[\mu_n(A_2)\right] \leq \log\left[\mu_{2n}\left(\frac{A_1 + A_2}{2}\right)\right] \leq 0.$$

By multiplying both sides by -1 and taking limits as $n \to \infty$ shows that equation (65) holds. The convexity of I now follows from Theorem 4.2.

We will now prove that the identity in equation (64) holds for every open and convex set $A \subset \mathcal{X}$. Without loss of generality we may assume that $A \subset \mathcal{E}$. If $L(A) = \infty$, then (64) trivially holds, otherwise if $L(A) < \infty$ then

$$L(A) = -\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq \inf_{x \in A} I(x) \tag{66}$$

by the large deviation principle lower bound. We will now show that the reverse inequality also holds. Let $\varepsilon > 0$, then it follows from (66) that there exists $N \in \mathbb{N}$ such that

$$-\frac{1}{n}\log\left[\mu_n(A)\right] < L(A) + \varepsilon, \quad \text{for every } n \geq N.$$

Since $\mathcal{X}$ is Polish it follows from Ulam's Theorem that every finite measure is tight. Hence for any $\delta > 0$ there exists a compact set $K \subset A$ such that

$$\mu_N(A \setminus K) < \delta.$$

Choose $\delta = (1 - e^{-\varepsilon})\mu_N(A)$, then

$$e^{-\varepsilon}\mu_N(A) = \mu_N(A) - \delta < \mu_N(K),$$

and

$$-\log\left[\mu_N(K)\right] < -\log\left[\mu_N(A)\right] + \varepsilon.$$

Which shows that

$$-\log\left[\mu_N(K)\right] < -\log\left[\mu_N(A)\right] + \varepsilon < L(A) + 2\varepsilon. \tag{67}$$

Thus the identity (64) holds for compact sets. We will now show that it holds for arbitrary convex open sets $A \subset \mathcal{X}$. The space $\mathcal{E}$ is metrizable and therefore regular by Urysohn's metrization theorem, which implies that for every $x \in A$ there exists an open set U_x with $x \in U_x$ and $\overline{U_x} \subset A$. Furthermore, since $\mathcal{E}$ is locally convex the sets U_x can be taken to

be convex. Since K is compact it may be covered by a finite subcollection $\{U_i\}_{i=1}^k$ of these sets. Define

$$\tilde{K} = \bigcup_{i=1}^{k} \left[\overline{U}_i \cap \overline{\text{co}}(K)\right],$$

and note that $\tilde{K} \subset \overline{\text{co}}(K)$ is a union of compact sets (since $\overline{\text{co}}(K)$ is compact whenever K is compact in completely metrizable locally convex space). Let $C = \overline{\text{co}}(\tilde{K})$, then C is a compact, closed, and convex subset of $\mathcal{E}$ that satisfies the set inclusion

$$K \subset C \subset \overline{\text{co}}(K) \subset A.$$

Therefore

$$-\log\left[\mu_n(A)\right] \leq -\log\left[\mu_N(C)\right] \leq -\log\left[\mu_N(K)\right] < L(A) + 2\varepsilon.$$

Since C is convex and closed it follows from Corollary 4.2 and the properties of subadditivity that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log\left[\mu_n(C)\right] &= \liminf_{n\to\infty} -\frac{1}{n} \log\left[\mu_n(C)\right] = \liminf_{n\to\infty} -\frac{N}{nN} \log\left[\mu_n(C)\right] \\
&\leq \liminf_{n\to\infty} -\frac{1}{nN} \log\left[\mu_{nN}(C)\right] \leq \lim_{n\to\infty} -\frac{1}{N} \log\left[\mu_N(C)\right] \\
&= -\frac{1}{N} \log\left[\mu_N(C)\right] < L(A) + 2\varepsilon.
\end{aligned} \tag{68}$$

And since C is compact and $C \subset A$ the large deviation upper bound gives

$$\inf_{x\in A} I(x) \leq \inf_{x\in C} \leq -\limsup_{n\to\infty} \frac{1}{n} \log\left[\mu_n(C)\right] < L(A) + 2\varepsilon.$$

Since this holds for every $\varepsilon > 0$ it follows that

$$\inf_{x\in A} I(x) \leq L(A),$$

Combined with equation (66) this shows that

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

■

We are now going to show that the rate function is given by the Legendre-Fenchel transform of the moment logarithmic moment generating, just as in the one-dimensional case.

Theorem 4.8 (Weak Cramér's Theorem)**.** *The sequence (μ_n) of distributions of the empirical means satisfy a weak large deviation principle with a convex rate function $I = \Lambda^*$, and*

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = -\inf_{x\in A} \Lambda^*(x), \tag{69}$$

for every convex and open $A \subset \mathcal{X}$.

Proof. It follows from the previous Lemma that (μ_n) satisfy a weak large deviation principle with convex rate function I. We are now going to use Theorem 4.6 to identify the

rate function. We start by noting that since the variables X_i are independent it follows that

$$\begin{aligned}\frac{1}{n}\Lambda_{\mu_n}(\lambda nt) &= \frac{1}{n}\log \mathbb{E}\left[e^{nt\langle \lambda, \mathbf{S}_n\rangle}\right] \\ &= \frac{1}{n}\log \mathbb{E}\left[e^{t(\langle \lambda, X_1\rangle + \cdots + \langle \lambda, X_n\rangle)}\right] \\ &= \frac{1}{n}\log\left(\mathbb{E}\left[e^{t\langle \lambda, X_1\rangle}\right]\cdots \mathbb{E}\left[e^{t\langle \lambda, X_n\rangle}\right]\right) \\ &= \frac{1}{n}\log\left[\int_{\mathcal{X}} e^{\langle \lambda, x\rangle}\, \mathrm{d}\mu\right] = \Lambda_{\mu}(\lambda t).\end{aligned}$$

Hence,

$$\Lambda_{\lambda}(t) = \lim_{n\to\infty}\frac{1}{n}\Lambda_{\mu_n}(\lambda nt) = \Lambda_{\mu}(\lambda t),$$

and the limit exists for every $\lambda \in \mathcal{X}^*$ and $t \in \mathbb{R}$. Furthermore, it follows from Lemma 4.5 that $\Lambda_\lambda(t)$ is lower semicontinuous. It remains to show that the inequality in equation (56) is satisfied, i.e. that

$$\inf\{I(x) : x \in \mathcal{X},\ \langle \lambda, x\rangle - \alpha > 0\} \leq \inf_{s>\alpha} \Lambda_\lambda^*(s)$$

Since every element $\lambda \in \mathcal{X}^*$ is bounded it holds that the random variables $\langle \lambda, \mathbf{S}_n\rangle$ are bounded. Furthermore, by linearity of λ we get that

$$\langle \lambda, \mathbf{S}_n\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle \lambda, X_i\rangle.$$

This is the empirical mean of a sum of i.i.d. bounded real valued random variables and therefore the logarithmic moment generating functions of $\langle \lambda, X_i\rangle$ are finite and Cramér's Theorem for real valued random variables holds. Especially, we get that

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{P}\big(\langle \lambda, \mathbf{S}_n\rangle \in [\alpha, \infty)\big) = \lim_{n\to\infty}\frac{1}{n}\log\big[\mu_n\{x : \langle \lambda, x\rangle - \alpha \geq 0\}\big] = -\inf_{s\geq\alpha}\Lambda_\lambda^*(s). \quad (70)$$

Note that level sets $\{x : \langle \lambda, x\rangle - \alpha > 0\}$ are convex and open subsets of $\mathcal{X}$, thus by Lemma 4.8 it holds that

$$\lim_{n\to\infty}\frac{1}{n}\log\mu_n\big(\{x : \langle \lambda, x\rangle - \alpha > 0\}\big) = -\inf\{I(x) : \langle \lambda, x\rangle - \alpha > 0\}. \quad (71)$$

And since $\{x : \langle \lambda, x\rangle - \alpha > 0\} \subset \{x : \langle \lambda, x\rangle - \alpha \geq 0\}$ we get that

$$\log\mu_n\big(\{x : \langle \lambda, x\rangle - \alpha > 0\}\big) \leq \log\mu_n\big(\{x : \langle \lambda, x\rangle - \alpha \geq 0\}\big),$$

which combined with equation (70) and equation (71) yields the inequality

$$-\inf_{s>\alpha}\Lambda_\lambda^*(s) \leq -\inf\{I(x) : x \in \mathcal{X},\ \langle \lambda, x\rangle - \alpha > 0\}.$$

This is equivalent to the condition on the rate function in Theorem 4.6, and it follows that (μ_n) satisfy a weak large deviation principle with rate Λ_μ^*.

■

Even if Theorem 4.8 can be applied in much more general spaces than Cramér's Theorem in $\mathbb{R}$ it only proves that (μ_n) satisfy a weak large deviation principle. However, by Lemma

4.2, the full large deviation principle can be attained by showing that the sequence (μ_n) is exponentially tight. This the case for the distributions of the empirical means in $\mathbb{R}^d$ if the logarithmic moment generating function is bounded (see e.g. [24, pp. 38–39]. Thus, by an application of Lemma 4.2 we get the following corollary of the Weak version of Cramér's Theorem.

Corollary 4.3 (Cramér's Theorem in $\mathbb{R}^d$)**.** *Let (X_i) be a sequence of i.i.d. real valued random variables with finite logarithmic moment generating functions. Then the distributions (μ_n) of the empirical means $(\mathbf{S}_n)$ satisfy the large deviation principle with good rate function Λ^*_μ.*

4.4 Transformations and Large Deviations

A natural question to ask is how the large deviation principle behaves under transformations. When $f : \mathcal{X} \to \mathcal{Y}$ is a continuous map and (μ_n) satisfy a large deviation principle on $\mathcal{X}$, then it can be shown that a large deviation principle also holds on $\mathcal{Y}$.

Theorem 4.9 (Contraction Principle)**.** *Let $\mathcal{X}, \mathcal{Y}$ be Hausdorff topological space and $f : \mathcal{X} \to \mathcal{Y}$ be continuous. If I is a good rate function on $\mathcal{X}$, then J defined by*

$$J(y) := \inf\{I(x) \ : \ x \in \mathcal{X}, \ f(x) = y\},$$

is a good rate function on $\mathcal{Y}$. Furthermore if (μ_n) satisfy the large deviation principle with rate I on $\mathcal{X}$, then $(\mu_n \circ f^{-1})$ satisfy the large deviation principle on $\mathcal{Y}$ with rate J.

The contraction principle can be very useful to derive new large deviations from existing ones and we will see an example of this later when we introduce Sanov's Theorem for the empirical distributions of the IS estimator. Even when the map f is not continuous the contraction provides a good guess for how the rate function might look on $\mathcal{Y}$ if it exists.

Another very useful tool that can be used to transform existing large deviations results to new spaces is based on projective limits.[13] We start with the definition of projective limits of topological spaces and some of their basic properties which can be found in [16, §4.4] and [24, Appendix B.1].

Definition 4.9. Let $(\mathcal{X}_\alpha, \mathcal{T}_\alpha)_{\alpha \in I}$ a collection of topological spaces indexed over a partially ordered set I, and $(p_{\alpha\beta})_{\alpha,\beta \in I}$ a collection of maps satisfying

1. $p_{\alpha\beta} : \mathcal{X}_\beta \to \mathcal{X}_\alpha$.
2. $p_{\alpha\gamma} = p_{\alpha\beta} \circ p_{\beta\gamma}$ if $\alpha \leq \beta \leq \gamma$.

Then $(\mathcal{X}_\alpha, p_{\alpha\beta})$ is called a *projective system.*

Definition 4.10. Let $(\mathcal{X}_\alpha, p_{\alpha\beta})$ be a projective system, then the *projective limit* of $(\mathcal{X}_\alpha)$ is the space

$$\varprojlim \mathcal{X}_\alpha := \left\{ \mathbf{x} \in \prod \mathcal{X}_\alpha \ : \ x_\alpha = p_{\alpha\beta}(x_\beta) \text{ for every } \alpha \leq \beta \right\},$$

equipped with subspace topology inherited from the product space $\prod \mathcal{X}_\alpha$.

The following theorem summarizes some important facts about projective limits.

Theorem 4.10 ([16, §4.4 Proposition 9])**.** *Let $(\mathcal{X}_\alpha, p_{\alpha\beta})$ be a projective system, then*

[13] Projective limits are also known as inverse limits (see e.g. [16]) but in the large deviations literature the term projective limit is most commonly used.

1. *The restriction of the projection maps* $p_\alpha : \prod \mathcal{X}_\alpha \to \mathcal{X}$ *to* $\varprojlim \mathcal{X}$ *are continuous maps.*

2. *If for every* $\alpha \in I$ *there is basis,* $\mathcal{U}_\alpha$*, of* $\mathcal{X}_\alpha$*, then the collection*

$$\mathcal{U} = \{p_\alpha^{-1}(B_\alpha) \ : \ \alpha \in I, \ B_\alpha \in \mathcal{U}_\alpha\}$$

is a basis for the topology on $\varprojlim \mathcal{X}$*.*

3. *If* $\mathcal{X}_\alpha$ *are all Hausdorff, then* $\varprojlim \mathcal{X}$ *is a closed subset of* $\prod \mathcal{X}_\alpha$ *(thus Hausdorff).*

4. *If* $A \subset \prod \mathcal{X}_\alpha$ *is closed, then*

$$A = \varprojlim p_\alpha(A) = \varprojlim \overline{p_\alpha(A)}.$$

The usefulness of projective limits to theory of large deviations is that large deviations results can be transformed from the spaces $\mathcal{X}_\alpha$ to the projective limit $\varprojlim \mathcal{X}_\alpha$. One of the main advantages of the projective limits methods is that it is possible to go from finite dimensional large deviation principles to an infinite-dimensional large deviation principle. Dawson and Gärtner introduced the method of projective limits in [23]. These ideas have been expanded upon by de Acosta in [3] and [1]. One major difference in de Acosta's work is that he shows that it possible drop the assumption that $\mathcal{X}$ is the projective limit of the projective system. Instead, the following assumption is used in [3] and [1].

Assumption 2. *Assume that* $(\mathcal{X}_\alpha, p_{\alpha\beta})$ *is a projective system and that there exists a set* $\mathcal{X}$ *and a collection of maps* $\{p_\alpha\}$ *that satisfy:*

1. $p_\alpha : \mathcal{X} \to \mathcal{X}_\alpha$ *are surjective maps.*

2. $p_\alpha = p_{\alpha\beta} \circ p_\beta$ *for every* $\alpha \leq \beta$*.*

3. *The collection* $\{p_\alpha\}$ *separates points in* $\mathcal{X}$*.*

Further assume that $\mathcal{X}$ *is equipped with the weak topology generated by the maps* $\{p_\alpha\}$*, and that* $\mathcal{B}$ *is a* σ*-algebra of subsets of* $\mathcal{X}$ *that satisfy:*

1. *Every compact subset of* $\mathcal{X}$ *is in* $\mathcal{B}$*.*

2. *There exists a basis* $\mathcal{U} \subset \mathcal{B}$ *for the topology on* $\mathcal{X}$*.*

Note that since the collection $\{p_\alpha\}$ is separating it follows that $\mathcal{X}$ is a Hausdorff topological space in the weak topology generated by $\{p_\alpha\}$.

Theorem 4.11 ([3, Theorem 4])**.** *Let* $\mathcal{X}$*,*$(\mathcal{X}_\alpha, p_{\alpha\beta})$ *and* $\mathcal{B} = \sigma(p_\alpha)$ *satisfy assumption 2 and* (μ_n) *be a sequence of probability measures on* $(\mathcal{X}, \mathcal{B})$*. If*

1. $\left(\mu_n \circ {p_\alpha}^{-1}\right)$ *satisfy the large deviation principle on* $\mathcal{X}_\alpha$ *with rate function* I_α *for every* α*, and*

2. *there exists a function* $I : \mathcal{X} \to [0, \infty]$ *such that* $\{I \leq M\}$ *is compact for every* $M \geq 0$ *and*

$$I_\alpha(z) = \inf\{I(x) : z = p_\alpha(x)\}.$$

Then (μ_n) *satisfy the large deviation principle on* $\mathcal{X}$ *with rate* I*. Furthermore,* I *is a good rate function and*

$$I(x) = \sup_\alpha \left\{I_\alpha\big(p_\alpha(x)\big)\right\}.$$

Theorem 4.11 may seem very abstract, but it is a very powerful tool that lets us lift large deviation principles from lower dimensions to higher dimensions. The general idea for us will be the following. We are interested in the large deviations of the distributions $(\boldsymbol{\mu}_n) \in \mathbf{M}_1\big(\mathbf{M}_1(\mathcal{X})\big)$ of either the empirical distributions $\mathbf{L}_n$ or $\mathbf{I}_n$. Then we can use Cramér's Theorem for large deviations in the finite dimensional spaces $\mathbb{R}^d$ and by defining a suitable projective system show that that the measures $(\boldsymbol{\mu}_n)$ also satisfy a large system principle, and we will show how this can be done in the next section.

4.5 Sanov's Theorem

In this section we study the large deviations of the empirical distributions. Sanov's Theorem asserts that $\mathbf{L}_n$ satisfy a large deviation principle and just like in the case of Cramér's Theorem there are many variations of Sanov's Theorem.

Sanov's Theorem in the Topology of Weak Convergence

Let $\mathcal{X}$ be a Polish space, then we know from Theorem 3.14 that $\mathbf{M}_1(\mathcal{X})$ is Polish. Furthermore, if (X_i) is a sequence of i.i.d. $\mathcal{X}$-valued random variables with distribution $\mu \in \mathbf{M}_1(\mathcal{X})$. Then the random variables δ_{X_i} are $\mathcal{B}_w$-measurable by Lemma 3.10 and Lemma 3.9. Since the space $\mathbf{M}(\mathcal{X})$ is a is a topological linear space in the topology of weak convergence it follows that the empirical distributions

$$\mathbf{L}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

are $\mathcal{B}_w$-measurable. Let $\boldsymbol{\mu}$ denote the distribution of the random variables δ_{X_i}, and note that

$$\boldsymbol{\mu}\big(\mathbf{M}_1(\mathcal{X})\big) = \mathbb{P}\big(\delta_{X_i} \in \mathbf{M}_1(\mathcal{X})\big) = 1.$$

Thus, the conditions of Assumption 1 are satisfied and the Weak Cramér's Theorem in Polish spaces (Theorem 4.8) can be applied to the distributions $(\boldsymbol{\mu}_n)$ of $(\mathbf{L}_n)$. This shows that $(\boldsymbol{\mu}_n)$ satisfy a weak large deviation principle on $\mathbf{M}_1(\mathcal{X})$. We state this as a Lemma below.

Lemma 4.9. *The distributions $(\boldsymbol{\mu}_n)$ of $(\mathbf{L}_n)$ satisfy a weak large deviation principle on $\mathbf{M}(\mathcal{X})$ with convex rate function*

$$I(\nu) = \Lambda^*_{\boldsymbol{\mu}}(\nu) = \sup_{f \in C_b(\mathcal{X})} \{\langle f, \nu\rangle - \Lambda_{\boldsymbol{\mu}}(f)\}.$$

Note that

$$\begin{aligned}
\Lambda(f) &= \sup_{f \in C_b(\mathcal{X})} \big\{\langle f, \nu\rangle - \Lambda_{\boldsymbol{\mu}}(f)\big\} \\
&= \sup_{f \in C_b(\mathcal{X})} \Big\{\langle f, \nu\rangle - \log \mathbb{E}\big[e^{\langle f, \delta_{X_i}\rangle}\big]\Big\} \\
&= \sup_{f \in C_b(\mathcal{X})} \Big\{\langle f, \nu\rangle - \log \mathbb{E}\big[e^{f(X_i)}\big]\Big\} \\
&= \sup_{f \in C_b(\mathcal{X})} \left\{\langle f, \nu\rangle - \log\left[\int_{\mathcal{X}} e^f \,\mathrm{d}\mu\right]\right\},
\end{aligned}$$

where the supremum is taken over $C_b(\mathcal{X})$. The reason for this is that $C_b(\mathcal{X})$ is separating and $\langle f, \cdot\rangle$ is linear on $\mathbf{M}(\mathcal{X})$, hence, the topological dual of $\mathbf{M}(\mathcal{X})$ in the weak topology is $C_b(\mathcal{X})$.

A full large deviation principle can be achieved by showing that the distributions $(\boldsymbol{\mu}_n)$ are exponentially tight. This is a well-known result which we state below. A proof can be found in [24, Lemma 6.2.6].

Lemma 4.10. *The sequence of distributions $(\boldsymbol{\mu}_n)$ of the empirical distributions $(\mathbf{L}_n)$ are exponentially tight.*

The final component that will be used in the proof of Sanov's Theorem is the identification of the rate function $\Lambda^*_{\boldsymbol{\mu}}$ with the relative entropy. The following lemma can be found in [25, Lemma 3.2.12]

Lemma 4.11. *Let $\boldsymbol{\mu}$ be the distribution of δ_{X_i}, then*

$$\mathbf{R}(\nu|\mu) = \Lambda^*_{\boldsymbol{\mu}}(\nu).$$

Theorem 4.12 (Sanov's Theorem)**.** *Let $\mathcal{X}$ be a Polish space and X_i a sequence of $\mathcal{X}$-valued random variables with distribution μ. Then the distributions $(\boldsymbol{\mu}_n)$ of the empirical distributions $(\mathbf{L}_n)$ satisfy the large deviations principle with good rate function $I(\nu) = \mathbf{R}(\nu|\mu)$.*

Proof. By Lemma 4.9 the distributions $(\boldsymbol{\mu}_n)$ satisfy a weak large deviation principle with convex rate function $I(\nu) = \Lambda_{\boldsymbol{\mu}}(\nu)$. Next, we use the fact that the distributions $(\boldsymbol{\mu}_n)$ are exponentially tight (Lemma 4.10), hence, by Lemma 4.2, they satisfy the full large deviation principle with rate I. Finally, by Lemma 4.11, the rate functions equals $\mathbf{R}(\cdot|\mu)$.

■

Sanov's Theorem in the τ-topology

Sanov's Theorem also holds on $(\mathbf{M}_1(\mathcal{X}), \mathcal{B}_\Psi)$ and then the condition that $\mathcal{X}$ is Polish can be dropped. In [2], it is showed that Sanov's theorem holds in the τ-topology whenever $\mathcal{X}$ is a measurable space and the X_i's are $\mathcal{X}$-valued random variables. In what follows we show the main ideas from [2] and how they can be used to derive Sanov's Theorem in the τ-topology when combined with the projective limits results from earlier. The main idea is to define a projective system such that $\boldsymbol{\mu}_n \circ {p_\alpha}^{-1}$ satisfy a finite-dimensional large deviation principle and then apply Theorem 4.11 to get the large deviation principle of $(\boldsymbol{\mu}_n)$ in $\mathbf{M}_1(\mathcal{X})$.

Let $\mathcal{F}$ be the collection of all finite subsets of $B(\mathcal{X})$. Then $\mathcal{F}$ is a directed set when ordered by inclusion, and we can define the projective system $\left(\mathbb{R}^F, (\Pi_{FG})_{F,G\in\mathcal{F}}\right)$. When $F \subset G$ the map $\Pi_{FG} : \mathbb{R}^G \to \mathbb{R}^F$ is defined to take a function $\Phi : G \to \mathbb{R}$ to its restriction on F, i.e.

$$\Pi_{FG}(\Phi) := \Phi|_F.$$

Next, for each $F \in \mathcal{F}$ we define the the projections $\Pi_F : \mathbf{M}(\mathcal{X}) \mapsto \mathbb{R}^F$ as

$$\Pi_F(\nu) := \Phi_\nu|_F,$$

where $\Phi_\nu|F \in \mathbb{R}^F$, denotes the restriction to F of the map $\Phi_\nu : B(\mathcal{X}) \to \mathbb{R}$, given by

$$\Phi_\nu(f) = \int_{\mathcal{X}} f \, \mathrm{d}\nu \,.$$

It can be shown that the weak topology induced by the maps $\{\Pi_F\}_{F\in\mathcal{F}}$ is equivalent to weak topology generated by Ψ_f (i.e. the τ-topology).

Now, consider the distributions of the projection maps $\boldsymbol{\mu}_n \circ \Pi_F{}^{-1} = \mathbb{P} \circ \mathbf{L}_n{}^{-1} \circ \Pi_F{}^{-1}$ on $\mathbb{R}^F$. By definition, $\boldsymbol{\mu}_n \circ \Pi_F{}^{-1}$ is the distribution of the random variables $\mathbf{S}_n : \Omega \to \mathbb{R}^F$ given by $\mathbf{S}_n = \Pi_F \circ \mathbf{L}_n$. Furthermore, we can express $\mathbf{S}_n$ as

$$\mathbf{S}_n = \frac{1}{n} \sum_{j=1}^{n} \mathbf{F}_j, \quad \mathbf{F}_j = \Phi_{\delta_{X_j}}|_F.$$

The functions $\mathbf{F}_j : \Omega \to \mathbb{R}^F$ are random variables. Hence, $\mathbf{S}_n$ can be considered as the mean of collection of random variable of $\mathbb{R}^F$-valued random variables. Furthermore, any finite dimensional real linear space is linearly isometric to $\mathbb{R}^d$, where the d is the dimension of the linear space. Hence Cramér's Theorem in $\mathbb{R}^d$ (Corollary 4.3) implies that the distributions of $\mathbf{S}_n$ satisfy the large deviation principle, and the rate function is given by

$$I_F(z) = \sup_{w \in \mathbb{R}^F} \left\{ \sum_{f \in F}^{n} z(f) w(f) - \log \left[\int_{\mathcal{X}} e^{\sum_{f \in F}^{n} z(f) f(x)} \, \mathrm{d}\mu(x) \right] \right\}.$$

This shows that the first condition of Theorem 4.11 is satisfied. Part of the second condition is covered by the following lemma from [2].

Lemma 4.12 ([2, Lemma 2.2]). *The rate function $I_F : \mathbb{R}^{\mathbb{F}} \to [0, \infty]$ satisfies*

$$I_F(z) = \inf \{ \mathbf{R}(\nu|\mu) : z = \Pi_F(\nu) \}.$$

In light of this result, all conditions of Theorem 4.11 are satisfied if the relative entropy has compact level sets in the τ-topology, which we know is true (see Corollary 3.2). Thus, the following version of Sanov's Theorem follows.

Theorem 4.13 (Sanov's Theorem in the τ-topology). *Let $\mathcal{X}$ be a measurable space and (X_i) a sequence of $\mathcal{X}$-valued random variables with distribution μ. Then the distributions $(\boldsymbol{\mu}_n)$ of the empirical distributions $(\mathbf{L}_n)$ satisfy the large deviations principle on $(\mathbf{M}_1(\mathcal{X}), \mathcal{B}_\Psi)$ with good rate function $I(\nu) = \mathbf{R}(\nu|\mu)$.*

Sanov's Theorem can be used to analyze the convergence rate of empirical distributions of the CMC-estimator and we will show examples of this in the next chapter. However, it cannot be applied directly to analyze the convergence rate to the IS estimator. Using the weak convergence approach to large deviations Hult and Nyquist proved in [33] that the empirical measures of the IS estimator with importance function f satisfy a Laplace principle.

Theorem 4.14 (Hult & Nyquist [33]). *Let X, Y be $\mathcal{X}$ valued random variables with distributions $\mu \ll \pi$, and $f : \mathcal{X} \to [0, \infty)$. If there exists a function $U : \mathcal{X} \to [0, \infty]$ satisfying*

1. *$\int_{\mathcal{X}} e^U \, \mathrm{d}\pi < \infty$, and*
2. *$\int_{\mathcal{X}} e^{tf(x)\rho(x)} \, \mathrm{d}\pi < \infty$, for every $t > 0$.*

Then the sequence $\mathbf{I}_n(f)$ satisfy the Laplace principle:

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\left[e^{-n\Phi(\mathbf{I}_n(f))} \right] = - \inf_{\nu \in \mathbf{M}_+(\mathcal{X})} \{ \Phi(\nu) + I(\nu) \}, \quad \text{for every } \Phi \in C_b\big(\mathbf{M}_+(\mathcal{X})\big),$$

with rate function

$$I(\nu) = \inf \left\{ R(\sigma, \pi) \ : \ \sigma \in \mathbf{M}_1(\mathcal{X}), \ \mathbf{R}(\sigma, \pi) < \infty, \ \nu = \textstyle\int_{\mathcal{X}} f w \, \mathrm{d}\sigma \right\},$$

in the τ-topology.

4.6 Applications to Monte Carlo Estimators

We are now going to look at how the large deviation principle can be applied to estimate the sample size required in CMC. Let us start by considering the case where we want the CMC estimator to achieve relative precision ε with confidence level $1-\alpha$, i.e.

$$\mathbb{P}\Big(|\theta_n - \theta| < \varepsilon|\theta|\Big) > 1-\alpha. \tag{72}$$

We write

$$R_\varepsilon := B\left(\theta, \varepsilon|\theta|\right),$$

to denote the open balls in $\mathbb{R}$ corresponding to the relative precision ε. Let $A_\varepsilon = R_\varepsilon^c$, then the identity in equation (72) is equivalent to

$$\mathbb{P}(\theta_n \in A_\varepsilon) \leq \alpha.$$

Let μ_n denote the distribution of the CMC estimator θ_n. We want to find $n \in \mathbb{N}$ such that the $\mu_n(A_\varepsilon) \leq \alpha$. Assume that the sequence of measures (μ_n) satisfies a large deviation principle with rate function I. Then, since the sets A_ε are open, the large deviation upper bound would imply that

$$\limsup_{n\to 0} \frac{1}{n} \log\left[\mu_n(A_\varepsilon)\right] \leq -\inf_{A_\varepsilon} I.$$

For large n we can interpret this as

$$\mu_n(A_\varepsilon) \lessapprox e^{-nI(A_\varepsilon)}.$$

Thus, to achieve the desired precision with confidence $1-\alpha$ we want

$$e^{-nI(A_\varepsilon)} \leq \alpha.$$

Note that $I \geq 0$ and $\alpha \in (0,1)$, thus by taking logarithms of both sides we get

$$nI(A_\varepsilon) \geq \log(\alpha)$$

Rearranging we get the identity

$$n \gtrapprox \frac{\log(\alpha)}{I(A_\varepsilon)}. \tag{73}$$

Example 5. Consider the case when want to compute $\theta = \mathbb{E}[X]$ for $X \sim N(\theta, \sigma^2)$. Then the rate function is given by

$$I(x) = \frac{(x-\theta)^2}{2\sigma^2}.$$

and

$$I(A_\varepsilon) = \inf\{I(x) : |x-\theta| > \varepsilon|\theta|\} = \frac{\varepsilon^2\theta^2}{2\sigma^2}$$

Thus, by equation (73), we want

$$n \gtrapprox \frac{\log(\alpha)2\sigma^2}{\varepsilon^2\theta^2}.$$

Note that this is very similar to the bound we derived in equation (12) in chapter 2, where we derived the following estimate from the confidence intervals of the CMC-estimator.

$$n \gtrapprox \frac{z_{1-\alpha/2}^2\sigma^2}{\varepsilon^2\theta^2}.$$

One may go one step further and ask how well does the empirical distributions $\mathbf{L}_n$ approximate the measure μ. Given a function $\Phi : \mathbf{M}_1(\mathcal{X}) \to \mathbb{R}$ one may define the sets

$$A_\varepsilon = \{\nu \in \mathbf{M}_1(\mathcal{X}) : |\Phi(\nu) - \Phi(\mu)| \geq \varepsilon\Phi(\mu)\}.$$

Sanov's Theorem can then be applied similarly to how Cramér's Theorem was used above.

$$\mathbb{P}(\mathbf{L}_n \in A_\varepsilon) \lessapprox e^{-nI(A_\varepsilon)}.$$

Thus, we would like to have

$$n \gtrapprox \frac{\log(\alpha)}{I(A_\varepsilon)},$$

just like before, but where

$$I(A_\varepsilon) = \inf\{\mathbf{R}(\nu|\mu) : |\Phi(\nu) - \Phi(\mu)| \geq \varepsilon\Phi(\mu)\}.$$

Section 4 of the article by Hult & Nyquist [33] contains several examples of how Sanov's Theorem and Theorem 4.14 can be applied to evaluate the performance CMC and IS estimators. For more on the application of large deviations to importance sampling and rare events the book [17] is a good introduction. Another recent book on large and moderate deviations with applications to rare events, based on the weak convergence approach, that covers many more applications of the large deviations principle is [18].

5 Conclusion

We have shown that many known results in the theory of weak convergence of probability measures also hold in the space of nonnegative finite measures. Especially, we have shown that empirical distributions of the IS estimator converge weakly. Furthermore, we introduced the τ-topology and compared it with the topology of weak convergence. Many of these results were required in chapter 4 on large deviations.

All of the results in chapter 4 are known and due to others. We closely followed the approach of Dembo & Zeitouni [24], Deuschel & Strock [25] and de Acosta [2], [3], where convexity in topological linear spaces play an important role. We also introduced projective limits and showed how Sanov's Theorem can be deduced from Cramér's Theorem for finite dimensional spaces. We believe that Sanov's Theorem for the empirical distributions of the IS estimators (Theorem 4.14) can be proved using similar methods and that the condition that $\mathcal{X}$ is Polish can be replaced with the much weaker condition that $\mathcal{X}$ is a measurable space. Just like in the case of Sanov's Theorem in the τ-topology.

The proof of Theorem 4.14 in [33] uses the weak convergence approach to large deviations. We have not discussed the weak convergence approach, but it is extensively covered in the research monograph [30]. An interesting question is whether these results can be extended to the weighted importance sampling estimator J_n. The case of the weighted importance sampling estimator is much more difficult, because independence is lost. Another challenge is that the estimator is biased.

Finally, we showed how Cramér's and Sanov's Theorem can be applied to approximate required sample size to achieve a desired precision using the Monte Carlo method. There are many clever way to define sets like we did with A_ε in Chapter 5 that can be used gain insight into the performance of the CMC and IS estimators. Several examples are given in section 4 of [33]. We compared this with the required sample sizes that can be derived from the confidence intervals of the CMC estimator, but there are other results which can be used to estimate the required sample sizes to achieve a desired precision. A recent result of this nature is given [19], where the relative entropy appear in their formula for the required sample size. It would be interesting to study whether some of these bounds on the sample size can be attained using the large deviation bounds.

A Preliminaries

A.1 Measure Theory

We use $\mathcal{A}$ to denote an algebra of subsets of $\mathcal{X}$ and $\mathcal{B}$ to denote a σ-algebra of subsets of $\mathcal{X}$. Given a collection $\mathcal{C} \subset 2^{\mathcal{X}}$ we use $\sigma(\mathcal{C})$ to denote the smallest σ-algebra containing $\mathcal{C}$ and $\mathcal{A}(\mathcal{C})$ to denote the smallest algebra containing $\mathcal{C}$. The inverse image of a σ-algebra is always a σ-algebra and it leads to the following.

Lemma A.1 ([5, Lemma 4.23]). *Let $f : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{C} \subset 2^{\mathcal{Y}}$, then*

$$\sigma\left(f^{-1}(\mathcal{C})\right) = f^{-1}\big(\sigma(\mathcal{C})\big).$$

The lemma above implies the following useful result.

Lemma A.2 ([5, Corollary 4.24]). *Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ be measurable spaces. If $\mathcal{B}_{\mathcal{Y}} = \sigma(\mathcal{C})$, then $f : \mathcal{X} \to \mathcal{Y}$ is measurable if and only if*

$$f^{-1}(C) \in \mathcal{B}_{\mathcal{X}}, \quad \text{for every } C \in \mathcal{C}.$$

Given a nonempty collection of functions $\mathcal{F}$ from $\mathcal{X}$ to a measurable space $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, the σ-algebra generated by $\mathcal{F}$ is defined as the smallest σ-algebra which makes every function in $\mathcal{F}$ measurable. It is given by

$$\sigma(f : f \in \mathcal{F}) := \sigma\left(f^{-1}(\mathcal{B}_{\mathcal{Y}}) : f \in \mathcal{F}\right).$$

Whenever $\mathcal{X}$ is a topological space we use $\mathcal{B}_{\mathcal{X}}$ to denote the Borel σ-algebra. Recall that the product σ-algebra on the product space $\mathcal{X}$ of the measurable spaces $\{(\mathcal{X}_\alpha, \mathcal{B}_\alpha\}$ is defined as

$$\bigotimes \mathcal{B}_\alpha := \sigma\left(\{p_\alpha^{-1}(A_\alpha) \ : \ A_\alpha \in B_\alpha\}\right),$$

where p_α denotes the projection map onto the α-th coordinate. From this definition and the definition of the product topology it is clear that for every collection $(\mathcal{X}_\alpha)$ of topological spaces

$$\bigotimes \mathcal{B}(\mathcal{X}_\alpha) \subset \mathcal{B}\left(\prod \mathcal{X}_\alpha\right).$$

For countable collections the following useful results hold (see e.g. [31, Proposition 1.4-5])

Theorem A.1. *Let I be a countable collection, and $(\mathcal{X}_i, \mathcal{B}_i)$ a collection of measurable spaces, then*

$$\bigotimes_{i \in I} B_i = \sigma\left(\left\{\prod A_i \ : \ A_i \in \mathcal{B}_i\right\}\right).$$

Furthermore if $\mathcal{X}_i$ are separable and second countable topological spaces, then

$$\bigotimes \mathcal{B}(\mathcal{X}_i) = \mathcal{B}\left(\prod X_i\right).$$

Definition A.1. A collection, $\mathcal{P}$ of subsets of $\mathcal{X}$ is called a π-system if $\mathcal{P}$ is closed under finite intersections.

Definition A.2. A collection, $\mathcal{D}$ of subsets of $\mathcal{X}$ is called a λ-system if it satisfies

1. $\mathcal{X} \in \mathcal{D}$.

2. $B \setminus A \in \mathcal{D}$ whenever $A, B \in \mathcal{D}$ and $A \subset B$.

3. $\lim_n A_n \in \mathcal{D}$ whenever (A_i) is an increasing sequence of sets in $\mathcal{D}$.

The following result is well known see e.g. [35, Theorem 1.1].

Theorem A.2 (Dynkins $\pi - \lambda$ Theorem)**.** *Let $\mathcal{D}$ be a λ-system and $\mathcal{P}$ be a π-system on $\mathcal{X}$. If $\mathcal{P} \subset \mathcal{D}$, then $\sigma(\mathcal{P}) \subset \mathcal{D}$.*

A.2 Probability Theory

In general the *$k-$th moment* of a real valued random variable is defined as $\mathbb{E}[|X^k|]$. Since probability spaces are finite measure spaces, it follows from Hölder's inequality that if a random variable has finite moment for some positive integer n, then it has finite moment for every positive integer $k \leq n$. For the 2-nd moment we especially get the following probabilistic *Cauchy-Schwartz inequality*

$$\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[|X^2|]}.$$

The *moment generating function* of a real valued random variable X with distribution μ, is given by

$$M_X(s) := \int_{\mathcal{X}} e^{sx} \, \mathrm{d}\mu(x) \,,$$

and is defined for every real s for which $M_X(s)$ is finite. It is clear that $M_X(0) = 1$ for every real valued random variable X, hence $M_X(s)$ is always well defined at 0. We say that the moment generating function of X exists whenever there exists an open interval of the form $(-h, h)$ where $M_X(s)$ is finite. The moment generating function is log-convex which means that $\log M_X$ is convex.

Inequalities

In this section we present three inequalities which will be used throughout the text. We start with *Markov's inequality* which provides an upper bound for the probability that the absolute value of a random variable is greater than $t > 0$.

Lemma A.3 (Markov's Inequality)**.** *Let X be a real valued random variable, then for every $t > 0$ it holds that*

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

The proof follows immediately after noting that

$$\mathbb{P}(|X| \geq t) = \int_{|X| \geq t} 1 \, \mathrm{d}\mathbb{P} \leq \frac{1}{t} \int_{\Omega} |X| \, \mathrm{d}\mathbb{P} \,.$$

By applying Markov's inequality to the function $(X - \mathbb{E}[X])^2$ the following very similar result follows.

Lemma A.4 (Chebychev's inequality)**.** *Let X be a real valued random variable with finite variance, then for every $t > 0$ it holds that*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{V}[X]}{t^2}$$

Chebychev's inequality differs from Markov's inequality in that it describes the probability that a random variable X deviates from its expected value.

Lemma A.5 (Chernoff's Bound)**.** *Let X be a real valued random variable and assume that the moment generating function $M_X(s)$ exists in some open interval $(-h, h)$, then for every $t > 0$ and $s \in (-h, h)$ it holds that*

$$\mathbb{P}(X \geq t) \leq M_X(s)e^{-st}.$$

Independence

Throughout this section we assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we use $\mathcal{F}_i$ to denote sub $\sigma-$algebras of $\mathcal{F}$. Let $\{F_i\}$ be a collection of sub σ-algebras of $\mathcal{F}$, then $\{F_i\}$ are said to be *independent* provided for every finite subcollection $\{\mathcal{F}_{i_1}, \dots \mathcal{F}_{i_n}\}$ it holds that

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_n}),$$

whenever $A_{i_k} \in F$ Similarly, we say that the events $A_1, \dots, A_n \in \mathcal{F}$ are independent if

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n).$$

Given a random variable $X : \Omega \to \mathcal{X}$ the *σ-algebra generated by X* is defined as the smallest σ-algebra containing $X^{-1}(\mathcal{B}_{\mathcal{X}})$ and denoted by $\sigma(X)$. A collection of random variables $\{X_i\}$ are *independent* if the σ-algebras $\{\sigma(X_i)\}$ are independent. Independence of sets is equivalent to independence of their indicator functions.

Lemma A.6. *The sets $A_1, \dots A_n$ are independent if and only if the indicator functions $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n}$ are independent.*

Lemma A.7. *Let $X_1, \dots, X_n$ be random variables taking values in the measurable spaces $(\mathcal{X}_1, \mathcal{B}_1), \dots, (\mathcal{X}_n, \mathcal{B}_n)$, then the random vector $\mathbf{X}_n : \Omega :\to \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ defined by $\mathbf{X}_n = (X_1, \dots, X_n)$ has distribution $\mu_1 \otimes \cdots \otimes \mu_n$ if and only if $X_1, \dots, X_n$ are independent.*

The following Lemma show some useful properties of independent random variables.

Lemma A.8. *Let $X_1, \dots, X_n$ be independent random variables taking values in the measurable spaces $(\mathcal{X}_1, \mathcal{B}_1), \dots, (\mathcal{X}_n, \mathcal{B}_n)$ and f be an integrable function on $(\mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \mu_1 \otimes \cdots \otimes \mu_n)$ then*

$$\begin{aligned}\mathbb{E}[f(X_1, \dots, X_n) &= \int_{\mathcal{X}_1 \times \cdots \mathcal{X}_n} f(x_1, \dots, x_n)\, \mathrm{d}\mu_1 \otimes \cdots \otimes \mathrm{d}\mu_n \\ &= \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_n} f(x_1, \dots, x_n \, \mathrm{d}\mu_n(x_n) \cdots \mathrm{d}\mu_1(x_1)\,.\end{aligned}$$

Using Lemma A.8 it is easy to derive that for independent real valued random variables

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n],$$

and

$$\mathbb{V}[X_1 + \cdots + X_n] = \mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n].$$

Limit Theorems

We use $S_n : \Omega \to \mathcal{X}$ to denote the random variable given by

$$S_n(\omega) := \frac{1}{n} \sum_{i=1} X_i(\omega).$$

Theorem A.3 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. real valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation $\mathbb{E}[X]$, then $S_n \xrightarrow{\mathbb{P}} \mathbb{E}[X]$, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\left(\left\{\omega \in \Omega \ : |S_n(\omega) - \mathbb{E}[X]| \geq \varepsilon\right\}\right) = 0.$$

Theorem A.4 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. real valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation $\mathbb{E}[X]$, then $S_n \to \mathbb{E}[X]$ almost surely, i.e.*

$$\mathbb{P}\left(\left\{\omega \in \Omega \ : \lim_{n\to\infty} S_n \to \mathbb{E}[X]\right\}\right) = 1.$$

Theorem A.5 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. real valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation $\mathbb{E}[X]$, and finite variance $\sigma^2 = \mathbb{V}[X] > 0$, then*

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right) \implies \mathcal{N}(0,1).$$

A.3 Topological Preliminaries

Nets and Convergence in Topological Spaces

This section briefly reviews some basic definitions about convergence and nets in general topological spaces. All spaces we shall be concerned with in this text are assumed to be Hausdorff (T2). There are many great textbooks on general topology and functional analysis which treat these topics in detail. See for example Willard [53], Kelley [36], and Aliprantis & Charalambos [5].

Let $(\mathcal{X}, \mathcal{T})$ be a topological space and $x \in \mathcal{X}$, recall that a collection, $\mathcal{N}(x)$, of neighbourhoods of x is called a *local basis* at x (also known as neighbourhoods basis) if for every neighbourhood U of x there exists an element $N \in \mathcal{N}(x)$ such that $N \subset U$. A topological space $\mathcal{X}$ is said to be *first countable* if every point $x \in \mathcal{X}$ has a countable local basis. In first countable spaces, sequences preserve many topological properties which are generally of interest. For instance, a point x lies in the closure $\overline{A}$ of a subset of a first countable space iff there exists a sequence in A converging to x. It follows from this that a map $f : \mathcal{X} \to \mathcal{Y}$ between two first countable spaces sequences is continuous if and only if $f(x_n) \to f(\lim x_n)$ for every convergent sequence (x_n) in $\mathcal{X}$ (see e.g. Willard [53, Theorem 10.4 and Corollary 10.5]).

In Hausdorff spaces which are not first countable nets provide a generalization of limits that preserves all topological invariants of interest to us when introducing different topologies on $\mathbf{M}_1(\mathcal{X})$. In order to define nets we need must start with the definition of a directed set.

Definition A.3. A *directed set*, is a nonempty set I with a binary relation $\leq$ that satisfy

1. $\alpha \leq \alpha$ for every $\alpha \in I$.
2. If $\alpha, \beta, \gamma \in I$ satisfy $\alpha \leq \beta$ and $\beta \leq \gamma$, then $\alpha \leq \gamma$.
3. If $\alpha, \beta \in I$, then there exists $\gamma \in I$ such that $\alpha \leq \gamma$ and $\beta \leq \gamma$.

If we would remove the third criteria in the definition of a directed set we would instead have a *partial ordering* ($\leq$) on I. Nets are an extension of sequences in general topological spaces with directed sets as index sets.

Definition A.4. A *net* in a topological space $\mathcal{X}$ is a function $P : I \to \mathcal{X}$ with domain I a directed set.

Just like with sequences we use the notation x_α to denote $P(\alpha) \in \mathcal{X}$ for $\alpha \in I$, and we write (x_α) to denote the net $P : I \to \mathcal{X}$. The definition of convergence of nets is similar to the definition of convergence for sequences in topological spaces.

Definition A.5. A net (x_α) over the index set I *converges* to $x \in \mathcal{X}$ if for every neighbourhood U of x there exists $\alpha_0 \in I$ such that $x_\alpha \in U$ whenever $\alpha \geq \alpha_0$.

In Theorem A.6 below we summarize three of the main properties of nets which make them useful generalizations of sequences for general topological spaces. Proofs of the statements can be found in most topology textbooks (see e.g. [53, Thm. 11.7-8]).

Theorem A.6. *Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces, then*

1. *If $A \subset \mathcal{X}$, then $x \in \overline{A}$ if and only if there exists a net (x_α) converging to $\mathcal{X}$.*
2. *A function $f : \mathcal{X} \to \mathcal{Y}$ is continuous if and only if the net $f(x_\alpha) \to f(x)$ in $\mathcal{Y}$ for every net (x_α)converging to $x \in \mathcal{X}$.*
3. *If $\mathcal{X}$ is Hausdorff and a net converges, then the limit is unique.*

We may also define the limit superior and limit inferior for nets of real numbers.

Definition A.6. Let (x_α) be a net in $\mathbb{R}$, then we define the *limit inferior* and *limit superior* of $(x_\alpha)_{\alpha \in I}$ as

$$\liminf_\alpha x_\alpha := \sup_{\alpha \in I} \left(\inf_{\beta \geq \alpha} x_\beta \right),$$
$$\limsup_\alpha x_\alpha := \inf_{\alpha \in I} \left(\sup_{\beta \geq \alpha} x_\beta \right).$$

A.4 Topological Linear Spaces

Weak Topologies in Linear Spaces

Throughout this section we use $\mathcal{X}$ to denote be a (real) linear space. We say that $\mathcal{X}$ is a *topological linear space* if it is equipped with a topology $\mathcal{T}$ such that that vector addition and and scalar multiplication are continuous maps. All vector spaces of interest to us will be Hausdorff, i.e. for any two distinct points in $\mathcal{X}$ there exists disjoint open neighbourhoods. In a Hausdorff space singletons are closed, and in topological linear spaces the converse also

holds, i.e. if $\{x\}$ is closed for every $x \in \mathcal{X}$, then $\mathcal{X}$ is Hausdorff (see e.g. [45, Theorem 1.12]). A topological linear space is *locally convex* if there exists a local base at 0 in which every member is convex. I.e. $\mathcal{X}$ is locally convex iff every neighbourhood of 0 has subset that is a convex neighbourhood of 0. We shall now turn our attention to different topologies on linear spaces and their dual spaces.

A subset $\mathcal{S} \subset \mathcal{T}$ is a *sub-basis* for the topology $\mathcal{T}$ on $\mathcal{X}$ if every open set, U, can be expressed as a union of finite intersections of elements of $\mathcal{S}$. The topology generated by a sub-basis $\mathcal{S}$ is the same as the topology generated by the basis

$$\mathcal{U} = \{S_1 \cap \cdots \cap S_n \ : \ n \in \mathbb{N},\ S_1, \ldots, S_n \in \mathcal{S}\},$$

i.e. the collection of all finite intersections of subsets of $\mathcal{S}$. Thus, a set U is open in the topology generated by $\mathcal{S}$ if and only if for every $x \in U$ there exists a finite collection $S_1, \ldots, S_n \in \mathcal{S}$ such that

$$x \in S_1 \cap \cdots \cap S_n \subset U.$$

A sub-basis is in many ways more natural than a basis for generating a topology from a collection of subsets of a space. Any collection $\mathcal{S} \subset 2^{\mathcal{X}}$ is a sub-basis for a topology on $\mathcal{X}$, which is the unique weakest topology on $\mathcal{X}$ containing $\mathcal{S}$. Let $\mathcal{F}$ be a collection of functions with domain $\mathcal{X}$ and each function $f \in \mathcal{F}$ taking values in a topological space $(\mathcal{X}_f, \mathcal{T}_f)$. Then, a sub-basis for the weakest topology such that every map $f \in \mathcal{F}$ is continuous is given by

$$\mathcal{S}_{\mathcal{F}} = \{f^{-1}(E) \ : \ f \in \mathcal{F},\ E \in \mathcal{T}_f\}. \tag{74}$$

If $\mathcal{U}_f$ is a basis for the topology $\mathcal{T}_f$ we may restrict the sets E in equation (74) to be taken from $\mathcal{U}_f$ instead. The *weak topology generated by* $\mathcal{F}$ on $\mathcal{X}$ is the topology $\mathcal{T}$ which have $\mathcal{S}_{\mathcal{F}}$ as sub-basis, it is the weakest topology such that every map $f \in \mathcal{F}$ is continuous. Let $\mathcal{X}^*$ denote the dual space of the locallly convex linear space $\mathcal{X}$, i.e. the space of continuous real valued linear functionals with domain $\mathcal{X}$. There is a canonical isomorphism of $\mathcal{X}$ into a subspace of its double dual V^{**} where each element $x \in \mathcal{X}$ is mapped into the evaluation map $\hat{x} \in \mathcal{X}^{**}$ defined by

$$\hat{x}(\Lambda) := \Lambda(x), \quad \text{for every } \Lambda \in \mathcal{X}^*.$$

An application of the Hahn-Banach Theorem can be used to show that the mapping $x \mapsto \hat{x}$ is norm-preserving and that $\mathcal{X}$ is linearly isomorphic to its image $\hat{\mathcal{X}} \subset \mathcal{X}^{**}$ under this map. For a proof of this fact the reader is referred to Theorem 10, Chapter 3 of [14] and the surrounding discussion. The *weak* topology* on the dual $\mathcal{X}^*$ is then weak topology generated by the collection $\hat{\mathcal{X}} \subset \mathcal{X}^{**}$. It follows from the definition of the weak topology that a set $U \subset \mathcal{X}^*$ is open in the weak* topology iff for every $\Lambda \in U$ there exists a finite collection of elements $\hat{x}_i$ and open intervals $(a_i, b_i) \subset \mathbb{R}$ such that

$$\Lambda \in \hat{x}_1^{-1}(a_1, b_1) \cap \cdots \cap \hat{x}_n^{-1}(a_n, b_n) \subset U.$$

The next lemma follows directly from the definition of the weak* topology and provides a usefula critetion for a set being open in the weak* topology. Some authors use it as the definition of the weak* topology.

Lemma A.9. *A set $U \subset \mathcal{X}^*$ is open in the weak* topology iff for every $\Lambda \in U$ there exists a finite collection of points $x_1, \ldots, x_n \in \mathcal{X}$ and $\varepsilon > 0$ such that*

$$\{\Lambda' \in \mathcal{X}^* \ : \ |\Lambda(x_i) - \Lambda'(x_i)| < \varepsilon, \text{ for every } i = 1, \ldots, n\} \subset U.$$

Equivalently put, for every $\Lambda \in \mathcal{X}^*$ the collection of all sets of the form above constitute a neighbourhood basis at Λ. From the definition of the weak* topology it is immediate that a net (Λ_α) in $\mathcal{X}^*$ converges to $\Lambda \in \mathcal{X}^*$ if and only if the net $\Lambda_\alpha(x)$ converges pointwise to $\Lambda(x)$ for every $x \in \mathcal{X}$. This is of such importance that we state it in the theorem below.

Theorem A.7. *Let $\mathcal{X}$ be a topological vector space, then a net (Λ_α) in the dual $\mathcal{X}^*$ converges weakly* to $\Lambda \in \mathcal{X}^*$ iff*

$$\lim_\alpha \Lambda_\alpha(x) = \Lambda(x), \quad \text{for every } x \in \mathcal{X}.$$

Weak Compactness

Theorem A.8 (Eberlain-Smulian Theorem (see e.g. [29, §V.6.1])**.** *Let $\mathcal{X}$ be a Banach space and $K \subset \mathcal{X}$, then the following three statements are equivalent:*

1. *K is weakly sequentially compact.*
2. *If $A = \{x_n : n \in \mathbb{N}\} \subset K$, then A has a weak limit point in K.*
3. *The weak closure of K is weakly compact.*

Convexity

In this section we review some basic facts about convex sets in topological linear spaces all of which can be found in chapter 5 of Aliprantis and Border [5].

Definition A.7. A subset $A \subset \mathcal{X}$ is convex if

$$\{tx + (1-t)y \ : \ t \in [0,1]\} \subset A, \quad \text{for every } x, y \in A.$$

Equivalently a set A is convex if and only if it holds that

$$\sum_{i=1}^n t_i x_i \in A,$$

for every finite collection of points $x_1, \ldots, x_n \in A$ and nonnegative real numbers $t_1, \ldots, t_n$ which satisfy $t_1 + \ldots, t_n = 1$. The following Lemma captures important properties of convex sets.

Lemma A.10 ([5, Lemma 5.27 & 5.28])**.** *Let $\{A_i\}$ be a, possibly uncountable, collection of convex sets, then*

1. *$C_1 + C_2$ is convex.*
2. *tC is convex for every $t \in \mathbb{R}$.*
3. *$\bigcap_i A_i$ is convex.*
4. *A° and $\overline{A}$ are convex and satisfy*

$$tA^\circ + (1-t)\overline{A} \subset \overline{A}, \quad t \in (0,1].$$

5. *If $A^\circ \neq \emptyset$, then $\overline{A^\circ} = \overline{A}$, and $\overline{A}^\circ = A^\circ$.*

Definition A.8. The *convex hull* of $A \subset \mathcal{X}$, written $\mathrm{co}(A)$ is the smallest convex set containing A.

Since the intersection of arbitrary collections of convex sets are closed it follows that the convex hull of A is the intersection of all convex subsets of $\mathcal{X}$ containing A. The convex hull of A can also be expressed as

$$\mathrm{co}(A) = \left\{ \sum_{i=1}^{n} t_i x_i \ : \ x_i \in A,\ t_i \in [0,1],\ \sum_{i=1}^{n} t_i = 1,\ n \in \mathbb{N} \right\}.$$

Definition A.9. The *closed convex hull* of $A \subset \mathcal{X}$, written $\overline{\mathrm{co}}(A)$ is the smallest closed convex subset of $\mathcal{X}$ containing A.

The closed convex hull of A is equal to the closure of the convex hull of A, i.e. $\overline{\mathrm{co}}(A) = \overline{\mathrm{co}(A)}$.

Theorem A.9 ([5, Thm 5.35])**.** *Let $\mathcal{X}$ be a locally convex completely metrizable topological linear space and $K \subset \mathcal{X}$ compact, then $\overline{\mathrm{co}}(K)$ is compact.*

Theorem A.10. *A convex subset of a locally convex space is weakly closed if and only if it is strongly closed.*

References

[1] A. de Acosta. "Projective Systems in Large Deviation Theory II: Some Applications". In: *Probability in Banach Spaces, 9*. Ed. by Jørgen Hoffmann-Jørgensen, James Kuelbs, and Michael B. Marcus. Boston, MA: Birkhäuser Boston, 1994, pp. 241–250. ISBN: 978-1-4612-0253-0.

[2] Alejandro de Acosta. "On large deviations of empirical measures in the τ-topology". In: *Journal of Applied Probability* 31.A (1994), pp. 41–47.

[3] Alejandro de Acosta. "Exponential tightness and projective systems in large deviation theory". In: *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (1997), pp. 143–156.

[4] AD Alexandroff. "Additive set-functions in abstract spaces". In: *Rec. Math.[Mat. Sbornik] NS* 13.2-3 (1943), pp. 169–238.

[5] Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. eng. 3rd ed. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2006. ISBN: 3540295860.

[6] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*. Vol. 57. Springer, 2007.

[7] Søren Asmussen and Mogens Steffensen. *Risk and insurance*. Springer, 2020.

[8] Viorel. Barbu and Teodor. Precupanu. *Convexity and Optimization in Banach Spaces*. eng. 4th ed. 2012. Springer Monographs in Mathematics. Dordrecht: Springer Netherlands, 2012. ISBN: 94-007-2247-8.

[9] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*. Vol. 5. Athena Scientific, 1996.

[10] P Billingsley. *Probability and measure*. 2nd ed. Wiley series in probability and mathematical statistics. New York [etc.]: Wiley, 1995.

[11] Patrick Billingsley. "The Invariance Principle for Dependent Random Variables". In: *Transactions of the American Mathematical Society* 83.1 (1956), pp. 250–268. ISSN: 00029947. URL: http://www.jstor.org/stable/1992915 (visited on 11/02/2023).

[12] Patrick Billingsley. *Convergence of probability measures*. eng. 2. ed. Wiley series in probability and statistics, Probability and statistics section. New York ; Wiley, 1999. ISBN: 0-471-19745-9.

[13] Vladimir I Bogachev. *Measure Theory*. eng. 1. Aufl. Vol. 1. Berlin, Heidelberg: Springer-Verlag, 2007. ISBN: 9783540345138.

[14] Béla. Bollobás. *Linear analysis : an introductory course*. eng. 2. ed. Cambridge mathematical textbooks. Cambridge: Cambridge Univ. Press, 1999. ISBN: 0-521-65577-3.

[15] Erwin Bolthausen and Uwe Schmock. "On the maximum entropy principle for uniformly ergodic Markov chains". In: *Stochastic Processes and their applications* 33.1 (1989), pp. 1–27.

[16] N Bourbaki. *General Topology: Chapters 1-4*. eng. 1st ed. Vol. 18. Elements of Mathematics. Berlin, Heidelberg: Springer Berlin / Heidelberg, 1988. ISBN: 9783540193746.

[17] James. Bucklew. *Introduction to Rare Event Simulation*. eng. 1st ed. 2004. Springer Series in Statistics. New York, NY: Springer New York, 2004. ISBN: 1-4757-4078-6.

[18] Amarjit Budhiraja and Paul Dupuis. *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods*. eng. 1st ed. 2019. Vol. 94. Probability Theory and Stochastic Modelling. New York, NY: Springer, 2019. ISBN: 9781493995776.

[19] Sourav Chatterjee and Persi Diaconis. "The sample size required in importance sampling". In: *The Annals of Applied Probability* 28.2 (2018), pp. 1099–1135.

[20] Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*. Vol. 4. Springer, 2020.
[21] John B. Conway. *A Course in Functional Analysis*. eng. 2nd ed. 2007. Graduate Texts in Mathematics, 96. New York, NY: Springer New York, 2007. ISBN: 1-4757-4383-1.
[22] Harald Cramér. "Sur un nouveau theoreme-limite de la theorie des probabilities". In: *Scientifiques et Industrielles* 736 (1938), pp. 5–23.
[23] Donald A Dawson and Jürgen Gärtner. "Large deviations from the McKean-Vlasov limit for weakly interacting diffusions". In: *Stochastics: An International Journal of Probability and Stochastic Processes* 20.4 (1987), pp. 247–308. DOI: 10.1080/17442508708833446.
[24] Amir. Dembo and Ofer. Zeitouni. *Large Deviations Techniques and Applications*. eng. 2nd ed. 2010. Stochastic Modelling and Applied Probability, 38. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. ISBN: 3-642-03311-3.
[25] Jean-Dominique Deuschel and Daniel W. Stroock. *Large deviations*. eng. Rev. ed. Pure and applied mathematics (1949) 137. Boston: Academic Press, 1989. ISBN: 0-12-213150-9.
[26] M. D. Donsker and S. R. S. Varadhan. "Asymptotic evaluation of certain markov process expectations for large time, I". eng. In: *Communications on pure and applied mathematics* 28.1 (1975), pp. 1–47. ISSN: 0010-3640.
[27] Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*. Vol. 1. 2. Springer, 2001.
[28] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511755347.
[29] Nelson Dunford and Jacob T Schwartz. *Linear operators, part 1: general theory*. Vol. 10. John Wiley & Sons, 1988.
[30] Paul Dupuis and Richard S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations: Dupuis/A Weak*. eng. John Wiley & Sons, Inc, 1997. ISBN: 9781118165904.
[31] Gerald B Folland. *Real analysis: modern techniques and their applications*. Vol. 40. John Wiley & Sons, 1999.
[32] Avner Friedman. *Foundations of modern analysis*. Courier Corporation, 1982.
[33] Henrik Hult and Pierre Nyquist. "Large deviations for weighted empirical measures arising in importance sampling". In: *Stochastic Processes and their Applications* 126.1 (2016), pp. 138–170.
[34] Olav Kallenberg. *Foundations of modern probability*. Second. Probability and its Applications (New York). Springer-Verlag, New York, 2002.
[35] Olav Kallenberg. *Foundations of Modern Probability*. Third. Probability Theory and Stochastic Modeling 99. Springer-Verlag, New York, 2021.
[36] John L Kelley. *General topology*. reprint ed. Courier Dover Publications, 2017 (1955).
[37] Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*. John Wiley & Sons, 2013.
[38] Oscar E. Lanford. "Entropy and equilibrium states in classical statistical mechanics". In: *Statistical Mechanics and Mathematical Problems*. Ed. by A. Lenard. Berlin, Heidelberg: Springer Berlin Heidelberg, 1973, pp. 1–113. ISBN: 978-3-540-38468-7.
[39] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*. Vol. 75. Springer, 2001.
[40] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces : fractals and rectifiability*. eng. Cambridge studies in advanced mathematics 44. Cambridge: Cambridge Univ. Press, 1995. ISBN: 0-521-46576-1.
[41] K. R. Parthasarathy. *Probability measures on metric spaces*. eng. Probability and Mathematical Statistics. New York, New York ; Academic Press, Inc., 1967. ISBN: 1-4832-2525-9.

[42] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.
[43] HL Royden and PM Fitzpatrick. *Real analysis 4th Edition*. Printice-Hall Inc, Boston, 2010.
[44] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*. 3rd ed. John Wiley & Sons, 2016.
[45] Walter Rudin. *Functional analysis*. eng. 2. ed. International series in pure and applied mathematics (New York). New York: McGraw-Hill, 1991. ISBN: 0070542368.
[46] D. Ruelle. "Correlation Functionals". eng. In: *Journal of mathematical physics* 6.2 (1965), pp. 201–220. ISSN: 0022-2488.
[47] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*. 2nd ed. Vol. 17. Cambridge university press, 2023.
[48] Daniel W Stroock. *Probability theory: an analytic view*. Cambridge university press, 2010.
[49] Flemming Topsøe. *Topology and measure*. eng. 1st ed. 1970. Lecture notes in mathematics ; 133. Berlin ; Springer-Verlag, 1970. ISBN: 3-540-36284-3.
[50] Veeravalli S Varadarajan. "On the convergence of sample probability distributions". In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 19.1/2 (1958), pp. 23–26.
[51] Veeravalli S Varadarajan. "Weak convergence of measures on separable metric spaces". In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 19.1/2 (1958), pp. 15–22.
[52] S. R. S. Varadhan. "Asymptotic probabilities and differential equations". In: *Communications on Pure and Applied Mathematics* 19.3 (1966), pp. 261–286. DOI: https://doi.org/10.1002/cpa.3160190303. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160190303. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160190303.
[53] Stephen Willard. *General topology*. eng. Addison-Wesley series in mathematics. Reading, Mass, 1970. ISBN: 99-0049174-2.