Large deviations, weak convergence, and relative entropy*

Markus Fischer, University of Padua

Revised June 20, 2013

1 Introduction

Rare event probabilities and large deviations: basic example and definition in Section 2. Essential tools for large deviations analysis: weak convergence of probability measures (Section 3) and relative entropy (Section 4). Weak convergence especially useful in the Dupuis and Ellis [1997] approach – see lectures!

Table 1: Notation

$\mathcal{X}$	a topological space
$\mathcal{B}(\mathcal{X})$	σ-algebra of Borel sets over $\mathcal{X}$, i.e., smallest σ-algebra containing all open (closed) subsets of $\mathcal{X}$
$\mathcal{P}(\mathcal{X})$	space of probability measures on $\mathcal{B}(\mathcal{X})$, endowed with topology of weak convergence of probability measures
$\mathbf{M}_b(\mathcal{X})$	space of all bounded Borel measurable functions $\mathcal{X} \to \mathbb{R}$
$\mathbf{C}(\mathcal{X})$	space of all continuous functions $\mathcal{X} \to \mathbb{R}$
$\mathbf{C}_b(\mathcal{X})$	space of all bounded continuous functions $\mathcal{X} \to \mathbb{R}$
$\mathbf{C}_c(\mathcal{X})$	space of all continuous functions $\mathcal{X} \to \mathbb{R}$ with compact support
$\mathbf{C}^k(\mathbb{R}^d)$	space of all continuous functions $\mathbb{R}^d \to \mathbb{R}$ with continuous partial derivatives up to order k

*Preparatory lecture notes for a course on "Representations and weak convergence methods for the analysis and approximation of rare events" given by Prof. Paul Dupuis, Brown University, at the Doctoral School in Mathematics, University of Padua, May 2013.

$\mathbf{C}_c^k(\mathbb{R}^d)$	space of all continuous functions $\mathbb{R}^d \to \mathbb{R}$ with compact support and continuous partial derivatives up to order k
$\mathbb{T}$	a subset of $\mathbb{R}$, usually $[0,T]$ or $[0,\infty)$
$\mathbf{C}(\mathbb{T}:\mathcal{X})$	space of all continuous functions $\mathbb{T} \to \mathcal{X}$
$\mathbf{D}(\mathbb{T}:\mathcal{X})$	space of all càdlàg functions $\mathbb{T} \to \mathcal{X}$ (i.e., functions continuous from the right with limits from the left)
$\wedge$	minimum (as binary operator)
$\vee$	maximum (as binary operator)

2 Large deviations

A standard textbook on the theory of large deviations is Dembo and Zeitouni [1998]; also see Ellis [1985], Deuschel and Stroock [1989], Dupuis and Ellis [1997], den Hollander [2000], and the references therein. A foundational work in the theory is Varadhan [1966].

2.1 Coin tossing

Consider the following random experiments: Given a number $n \in \mathbb{N}$, toss n coins of the same type and count the number of coins that land heads up. Denote that (random) number by S_n. Then S_n/n is the empirical mean, here equal to the empirical probability, of getting heads. What can be said about S_n/n for n large?

To construct a mathematical model for the coin tossing experiments, let $X_1, X_2, \ldots$ be $\{0,1\}$-valued independent and identically distributed (i.i.d.) random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Thus each X_i has Bernoulli distribution with parameter $p \doteq \mathbf{P}(X_1 = 1)$. Interpret $X_i(\omega) = 1$ as saying that coin i at realization $\omega \in \Omega$ lands head up. Then $S_n = \sum_{i=1}^n X_i$. By the strong / weak law of large numbers (LLN),

$$\frac{S_n}{n} \overset{n\to\infty}{\longrightarrow} p \quad \text{with probability one / in probability.}$$

In particular, by the weak law of large numbers, for all $\varepsilon > 0$,

$$\mathbf{P}\left\{S_n/n - p \geq \varepsilon\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

More can be said about the asymptotic behavior of those *deviation* probabilities. Observe that S_n has binomial distribution with parameters (n,p),

that is,

$$\mathbf{P}\left\{S_n = k\right\} = \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k}, \quad k \in \{0, \ldots, n\}.$$

By Stirling's formula, asymptotically for large n,

$$\mathbf{P}\left\{S_n = k\right\} \simeq \frac{\sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi k}\, k^k e^{-k} \sqrt{2\pi(n-k)}(n-k)^{n-k} e^{-(n-k)}} p^k (1-p)^{n-k}.$$

Therefore, if $k \simeq n \cdot x$ for some $x \in (0,1)$, then

$$\begin{aligned}\log \mathbf{P}\left\{S_n = k\right\} \simeq -\frac{1}{2}\left(\log(2\pi) + \log(x) + \log(1-x) + \log(n)\right) \\ - n\, x \log\left(\frac{x}{p}\right) - n(1-x)\log\left(\frac{1-x}{1-p}\right),\end{aligned}$$

hence

$$\frac{1}{n}\log \mathbf{P}\left\{S_n = k\right\} \simeq -\left(x \log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right)\right).$$

The expression $x \log(\frac{x}{p}) + (1-x)\log(\frac{1-x}{1-p})$ gives the *relative entropy* of the Bernoulli distribution with parameter x w.r.t. the Bernoulli distribution with parameter p, which is minimal and zero if and only if $x = p$. The asymptotic equivalence

$$\frac{1}{n}\log \mathbf{P}\left\{S_n = k\right\} \simeq -\left(x \log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right)\right), \quad k \simeq n\, x,$$

shows that the probabilities of the events $\{S_n/n - p \geq \varepsilon\}$ converge to zero *exponentially fast* with rate (up to arbitrary small corrections)

$$-\left((p+\varepsilon)\log\left(\frac{p+\varepsilon}{p}\right) + (1-p-\varepsilon)\log\left(\frac{1-p-\varepsilon}{1-p}\right)\right),$$

which corresponds to $x = p + \varepsilon$, the rate of slowest convergence.

Events like $\{S_n/n - p \geq \varepsilon\}$ describe *large deviations* from the law of large numbers limit, in contrast to the *fluctuations* ("normal deviations") captured by the central limit theorem, which says here that the distribution of $\sqrt{n} \cdot (S_n/n - p)$ is asymptotically Gaussian with mean zero and variance $p(1-p)$.

2.2 The large deviation principle

The theory of large deviations will be developed in this course for random variables taking values in a Polish space. A *Polish space* is a separable topological space that is compatible with a complete metric. Examples of Polish spaces are

- $\mathbb{R}^d$ with the standard topology,
- any closed subset of $\mathbb{R}^d$ (or another Polish space) equipped with the induced topology,
- the space $\mathbf{C}(\mathbb{T} : \mathcal{X})$ of continuous functions, $\mathbb{T} \subseteq (-\infty, \infty)$ an interval, $\mathcal{X}$ a complete and separable metric space, equipped with the topology of uniform convergence on compact subsets of $\mathbb{T}$,
- the space $\mathbf{D}(\mathbb{T} : \mathcal{X})$ of càdlàg functions, $\mathbb{T} \subseteq (-\infty, \infty)$ an interval, a $\mathcal{X}$ a complete and separable metric space, equipped with the Skorohod topology [e.g. Billingsley, 1999, Chapter 3],
- the space $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{B}(\mathcal{X})$, $\mathcal{X}$ a Polish space, equipped with the weak convergence topology (cf. Section 3).

Let $(\xi^n)_{n\in\mathbb{N}}$ be a family of random variables with values in a Polish space $\mathcal{S}$. Let $I : \mathcal{S} \to [0, \infty]$ be a function with compact sublevel sets, i.e., $\{x \in \mathcal{S} : I(x) \leq c\}$ is compact for every $c \in [0, \infty)$. Such a function is lower semicontinuous and is called a (good) *rate function*.

Definition 2.1. The sequence $(\xi^n)_{n\in\mathbb{N}}$ satisfies the *large deviation principle* with rate function I iff for all $G \in \mathcal{B}(\mathcal{S})$,

$$\begin{aligned} -\inf_{x\in G^\circ} I(x) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\xi^n \in G\right\} \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\xi^n \in G\right\} \leq -\inf_{x\in \mathrm{cl}(G)} I(x). \end{aligned}$$

The large deviation principle is a distributional property: Writing $\mathbf{P}_n$ for $\mathrm{Law}(\xi^n)$, (ξ^n) satisfies the large deviation principle with rate function I if and only if for all $G \in \mathcal{B}(\mathcal{S})$,

$$-\inf_{x\in G^\circ} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}_n(G) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}_n(G) \leq -\inf_{x\in \mathrm{cl}(G)} I(x).$$

The large deviation principle gives a rough description of the asymptotic behavior of the probabilities of *rare events*. For simplicity, consider only I-continuity sets; a set $G \in \mathcal{B}(\mathcal{S})$ is called an *I-continuity set* if $\inf_{x \in \mathrm{cl}(G)} I(x) = \inf_{x \in G^\circ} I(x)$. For such G,

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left\{\xi^n \in G\right\} = -\inf_{x \in G} I(x) \doteq -I(G),$$

hence

$$\mathbf{P}\left\{\xi^n \in G\right\} = \mathbf{P}_n(G) = e^{-n(I(G)+o(1))}.$$

The probability of G w.r.t. the law of ξ^n therefore tends to zero exponentially fast as $n \to \infty$ whenever $I(G) > 0$. Thus, if $I(G) > 0$, then G is a rare event (w.r.t. $\mathbf{P}_n$ for large n).

Example 1 (Coin tossing). The sequence $\xi^n \doteq S_n/n$, $n \in \mathbb{N}$, satisfies the large deviation principle in $\mathcal{S} \doteq \mathbb{R}$ (or $\mathcal{S} \doteq [0,1]$) with rate function

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

If $p \in (0,1)$, then I is finite and continuous on $[0,1]$, and convex on $\mathbb{R}$.

A concept closely related to the large deviation principle is what is called Laplace principle. Let $(\xi^n)_{n \in \mathbb{N}}$ be a family of $\mathcal{S}$-valued random variables, $\mathbf{P}_n = \mathrm{Law}(\xi^n)$.

Definition 2.2. The sequence (ξ^n) satisfies the *Laplace principle* with rate function I iff for all $F \in \mathbf{C}_b(\mathcal{S})$,

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbf{E}\left[\exp\left(-n \cdot F(\xi^n)\right)\right] = \inf_{x \in \mathcal{S}} \left\{I(x) + F(x)\right\}.$$

In Definition 2.2 it is clearly equivalent to require that for all $F \in \mathbf{C}_b(\mathcal{S})$,

$$\lim_{n \to \infty} \frac{1}{n} \log \int_{\mathcal{S}} \exp\left(n \cdot F(x)\right) \mathbf{P}_n(dx) = \sup_{x \in \mathcal{S}} \left\{F(x) - I(x)\right\}.$$

If $\mathcal{S} = \mathbb{R}^d$ and if we took $F(x) = \theta \cdot x$ (such an F is clearly unbounded), then on the right-hand side above we would have the *Legendre transform* of I at $\theta \in \mathbb{R}^d$. Also notice the analogy with *Laplace's method* for asymptotically evaluating exponential integrals: $\lim_{n \to \infty} \frac{1}{n} \log \int_0^1 e^{nf(x)} dx = \max_{x \in [0,1]} f(x)$ for all $f \in \mathbf{C}([0,1])$.

Basic results [for instance Dembo and Zeitouni, 1998, Chapter 4]:

1. The (good) rate function of a large deviation principle is uniquely determined.

2. If I is the rate function of a large deviation principle, then $\inf_{x\in\mathcal{S}} I(x) = 0$ and $I(x^*) = 0$ for some $x^* \in \mathcal{S}$. If I has a unique minimizer, then the large deviation principle implies a corresponding law of large numbers.

3. The large deviation principle holds if and only if the Laplace principle holds, and the (good) rate function is the same.

4. Contraction principle: Let $\mathcal{Y}$ be a Polish space and $\psi : \mathcal{S} \to \mathcal{Y}$ be a measurable function. If (ξ^n) satisfies the large deviation principle with (good) rate function I and if ψ is continuous on $\{x \in \mathcal{S} : I(x) < \infty\}$, then $(\psi(\xi^n))$ satisfies the large deviation principle with (good) rate function $J(y) \doteq \inf_{x\in\psi^{-1}(y)} I(x)$.

2.3 Large deviations for empirical means

Let $X_1, X_2, \ldots$ be $\mathbb{R}$-valued i.i.d. random variables with common distribution μ such that $m \doteq \int x\mu(dx) = \mathbf{E}[X_1]$ is finite. As in the case of coin flipping, set $S_n \doteq \sum_{i=1}^n X_i$ and consider the asymptotic behavior of S_n/n, $n \in \mathbb{N}$. By the law of large numbers, $S_n/n \overset{n\to\infty}{\longrightarrow} m$ with probability one.

Let ϕ_μ be the *moment generating function* of μ (or $X_1, X_2, \ldots$), that is,

$$\phi_\mu(t) \doteq \int_{\mathbb{R}} e^{t\cdot x}\mu(dx) = \mathbf{E}\left[e^{t\cdot X_1}\right], \quad t \in \mathbb{R}.$$

Theorem 2.1 (Cramér)**.** *Suppose that μ is such that $\phi_\mu(t)$ is finite for all $t \in \mathbb{R}$. Then $(S_n/n)_{n\in\mathbb{N}}$ satisfies the large deviation principle with rate function I given by*

$$I(x) \doteq \sup_{t\in\mathbb{R}} \left\{t \cdot x - \log\left(\phi_\mu(t)\right)\right\}.$$

The rate function I in Theorem 2.1 is the Legendre transform of $\log \phi_\mu$, the logarithmic moment generating function or *cumulant generating function* of the common distribution μ. Two particular cases:

1. Bernoulli distribution: μ the Bernoulli distribution on $\{0, 1\}$ with parameter p. Then $\phi_\mu(t) = 1 - p + p \cdot e^t$ and

$$I(x) = x \log\left(\frac{x}{p}\right) + (1 - x) \log\left(\frac{1 - x}{1 - p}\right), \quad x \in [0, 1],$$

$I(x) = \infty$ for $x \in \mathbb{R} \setminus [0, 1]$, as in Example 1.

2. Normal distribution: μ the normal distribution with mean 0 and variance σ^2. Then $\phi_\mu(t) = e^{\sigma^2 t^2/2}$ and

$$I(x) = \frac{x^2}{2\sigma^2}, \quad x \in \mathbb{R}.$$

Some properties of $\log \phi_\mu$ and the associated rate function I from Theorem 2.1 (hypothesis $\phi_\mu(t) < \infty$ for all $t \in \mathbb{R}$):

- ϕ_μ is in $\mathbf{C}^\infty(\mathbb{R} : (0, \infty))$,
- $\log \phi_\mu$ is strictly convex,
- $I(x) = \infty$ for $x \notin [\operatorname{essinf} X_1, \operatorname{esssup} X_1] = \operatorname{Conv}(\operatorname{supp}(\mu))$,
- I is non-negative and convex,
- I is strictly convex and infinitly differentiable on $(\operatorname{essinf} X_1, \operatorname{esssup} X_1)$,
- $I(m) = 0$, $I'(m) = 0$, and $I''(m) = 1/\operatorname{var}(\mu)$.

Here we give a sketch of the proof of Theorem 2.1; for a complete proof of Cramér's theorem under weaker assumptions on ϕ_μ see Section 2.2 in Dembo and Zeitouni [1998, pp. 26-35].

Proof of Theorem 2.1 (sketch). May assume $m = \mathbf{E}[X_1] = 0$. Given $x > 0$, consider the probabilities of the events $\{S_n/n \geq x\}$. By Markov's inequality and since S_n is the sum of i.i.d. random variables with common distribution μ, we have for $t > 0$,

$$\mathbf{P}\left\{S_n \geq n\,x\right\} = \mathbf{P}\left\{e^{t\,S_n} \geq e^{t\,n\,x}\right\} \leq e^{-t\,n\,x}\,\mathbf{E}\left[e^{t\,S_n}\right] = e^{-t\,n\,x}(\phi_\mu(t))^n.$$

Since this inequality holds for any $t > 0$ and log is non-decreasing, we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{S_n/n \geq x\right\} \leq -\sup_{t>0}\left\{t \cdot x - \log(\phi_\mu(t))\right\}.$$

By Jensen's inequality, $\log \phi_\mu(t) \geq t \cdot \mathbf{E}[X_1] = t \cdot m = 0$. Since $\log \phi_\mu(0) = 0$, we have $I(m) = I(0) = 0$ and

$$t \cdot x - \log(\phi_\mu(t)) \leq t \cdot m - \log(\phi_\mu(t)) \leq I(m) = 0 \quad \text{for every } t \leq 0.$$

It follows that

$$\sup_{t>0}\left\{t \cdot x - \log(\phi_\mu(t))\right\} = \sup_{t\in\mathbb{R}}\left\{t \cdot x - \log(\phi_\mu(t))\right\} = I(x),$$

which yields the large deviation upper bound for closed sets of the form $[x, \infty)$, $x > 0$. A completely analogous argument gives the upper bound for closed sets of the form $(-\infty, x]$, $x < 0$. Let $G \subset \mathbb{R}$ be a closed set not containing zero. Then, thanks to the strict convexity of I and the fact that $I \geq 0$ and $I(0) = 0$, there are $x_+ > 0$, $x_- < 0$ such that $G \subseteq (-\infty, x_-] \cup [x_+, \infty)$ and $\inf_{x \in G} I(x) = \inf_{x \in (-\infty, x_-] \cup [x_+, \infty)} I(x)$. This establishes the large deviation upper bound.

To obtain the large deviation lower bound, we first consider open sets of the form $(x - \delta, x + \delta)$ for $x \in (\operatorname{essinf} X_1, \operatorname{esssup} X_1)$, $\delta > 0$. Fix such x, δ. Since ϕ_μ is everywhere finite by hypothesis, thus $\log \phi_\mu$ continuously differentiable and strictly convex on $\mathbb{R}$, there exists a unique solution $t_x \in \mathbb{R}$ to the equation

$$x = (\log \phi_\mu)'(t_x) = \frac{\phi'(t_x)}{\phi(t_x)} = \frac{\mathbf{E}\left[X_1 e^{t_x X_1}\right]}{\mathbf{E}\left[e^{t_x X_1}\right]},$$

and

$$I(x) = t_x \cdot x - \log \phi_\mu(t_x).$$

Define a probability measure $\tilde{\mu} \in \mathcal{P}(\mathbb{R})$ absolutely continuous with respect to μ according to

$$\frac{d\tilde{\mu}}{d\mu}(y) \doteq \exp\left(t_x \cdot y - \log \phi_\mu(t_x)\right) = \frac{e^{t_x \cdot y}}{\phi_\mu(t_x)}.$$

Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with common distribution $\tilde{\mu}$, and set $\tilde{S}_n \doteq \sum_{i=1}^n Y_i$, $n \in \mathbb{N}$. Then $\mathbf{E}[Y_i] = x$ and, for $\varepsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{S_n/n \in (x - \varepsilon, x + \varepsilon)\right\} \\
&= \int_{\left\{\boldsymbol{z} \in \mathbb{R}^n : \sum_{i=1}^n z_i \in (n(x-\varepsilon), n(x+\varepsilon))\right\}} \otimes^n \mu(d\boldsymbol{y}) \\
&\geq e^{-n(t_x(x-\varepsilon) \vee t_x(x+\varepsilon))} \int_{\left\{\boldsymbol{z} \in \mathbb{R}^n : \sum_{i=1}^n z_i \in (n(x-\varepsilon), n(x+\varepsilon))\right\}} e^{t_x \cdot \sum_{i=1}^n y_i} \otimes^n \mu(d\boldsymbol{y}) \\
&= e^{n \cdot \log \phi_\mu(t_x)} \cdot e^{-n(t_x(x-\varepsilon) \vee t_x(x+\varepsilon))} \cdot \mathbf{P}\left\{\tilde{S}_n/n \in (x - \varepsilon, x + \varepsilon)\right\}.
\end{aligned}$$

Since $Y_1, Y_2, \ldots$ are i.i.d. with $\mathbf{E}[Y_i] = x$, we have $\tilde{S}_n/n \to x$ in probability as $n \to \infty$ by the weak law of large numbers. It follows that

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left\{S_n/n \in (x - \varepsilon, x + \varepsilon)\right\} \geq \log \phi_\mu(t_x) - t_x(x - \varepsilon) \vee t_x(x + \varepsilon).$$

Since $\varepsilon > 0$ was arbitrary and $\log \phi_\mu(t_x) - t_x \cdot x = -I(x)$, we obtain

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left\{S_n/n \in (x - \delta, x + \delta)\right\} \geq -I(x).$$

If $G \subset \mathbb{R}$ is open such that $G \cap (\operatorname{essinf} X_1, \operatorname{esssup} X_1) \neq \emptyset$, then the preceding inequality implies that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{S_n/n \in G\right\} \geq \sup_{x\in G} -I(x) = -\inf_{x\in G} I(x).$$

□

The parameter t_x in the proof of Theorem 2.1 corresponds to an *absolutely continuous change of measure* with exponential density. Under the new measures, the random variable X_1 has expected value x instead of zero – the rare event $\{S_n/n \geq x\}$ becomes typical!

3 Weak convergence of probability measures

Here we collect some facts about weak convergence of probability measures. A standard reference on the topic is Billingsley [1968, 1999]; also see Chapter 3 in Ethier and Kurtz [1986] or Chapter 11 in Dudley [2002]. Let $\mathcal{X}$ be a Polish space. Let $\mathcal{B}(\mathcal{X})$ be the Borel σ-algebra over $\mathcal{X}$. Denote by $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{B}(\mathcal{X})$.

Definition 3.1. A sequence $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ is said to *converge weakly* to $\theta \in \mathcal{P}(\mathcal{X})$, in symbols $\theta_n \xrightarrow{w} \theta$, if

$$\int_{\mathcal{X}} f(x)\theta_n(dx) \overset{n\to\infty}{\longrightarrow} \int_{\mathcal{X}} f(x)\theta_n(dx) \text{ for all } f \in \mathbf{C}_b(\mathcal{X}).$$

Remark 3.1. In the terminology of functional analysis, the weak convergence of Definition 3.1 would be called weak-$*$ convergence. Denote by $\mathcal{M}_{sgn}(\mathcal{X})$ the space of finite signed measures on $\mathcal{B}(\mathcal{X})$. Then $\mathcal{M}_{sgn}(\mathcal{X})$ is a Banach space under the total variation norm ("strong convergence"), and $\mathcal{P}(\mathcal{X})$ is a closed convex subset of $\mathcal{M}_{sgn}(\mathcal{X})$ with respect to both the strong and the weak convergence topology. Moreover, $\mathcal{M}_{sgn}(\mathcal{X})$ can be identified with a subspace of the topological dual $\mathbf{C}_b(\mathcal{X})^*$ of the Banach space $\mathbf{C}_b(\mathcal{X})$ under the sup norm (if $\mathcal{X}$ is compact, then $\mathcal{M}_{sgn}(\mathcal{X}) \equiv \mathbf{C}_b(\mathcal{X})^*$; in general, $\mathbf{C}_b(\mathcal{X})$ is not reflexive, not even for $\mathcal{X} = [0,1]$). Thus $\mathcal{P}(\mathcal{X}) \subset \mathbf{C}_b(\mathcal{X})^*$, and the convergence of Definition 3.1 coincides with the weak-$*$ convergence on the dual space induced by the original space $\mathbf{C}_b(\mathcal{X})$.

Example 2 (Dirac measures). Let $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ be a convergent sequence with limit $x \in \mathcal{X}$. Then the sequence of Dirac measures $(\delta_{x_n})_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ converges weakly to the Dirac measure δ_x.

Example 3 (Normal distributions). Let $(m_n)_{n\in\mathbb{N}} \subset \mathbb{R}$, $(\sigma_n^2)_{n\in\mathbb{N}} \subset [0,\infty)$ be convergent sequences with limits m and σ^2, respectively. Then the sequence of normal distributions $(N(m_n,\sigma_n^2))_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R})$ converges weakly to $N(m,\sigma^2)$.

Example 4 (Product measures). Suppose $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$, $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{Y})$ are weakly convergent sequences with limits θ and μ, respectively, where $\mathcal{Y}$ is Polish, too. Then the sequence of product measures $(\theta_n \otimes \mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ converges weakly to $\theta\otimes\mu$.

Example 5 (Marginals vs. joint distribution). Let X, Y, Z be independent real-valued random variables defined on some probability space $(\Omega,\mathcal{F},\mathbf{P})$ such that $\mathrm{Law}(X) = N(0,1) = \mathrm{Law}(Y)$ (i.e., X, Y have standard normal distribution), and Z has Rademacher distribution, that is, $\mathbf{P}(Z=1) = 1/2 = \mathbf{P}(Z=-1)$. Define a sequence $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^2)$ by $\theta_{2n} \doteq \mathrm{Law}(X,Y)$, $\theta_{2n-1} \doteq \mathrm{Law}(X, Z\cdot X)$, $n\in\mathbb{N}$. Observe that the random variable $Z\cdot X$ has standard normal distribution $N(0,1)$. The marginal distributions of (θ_n) therefore converge weakly (they are constant and equal to $N(0,1)$), while the sequence (θ_n) itself does not converge (indeed, the joint distribution of X and $Z\cdot X$ is not even Gaussian). To prove weak convergence of probability measures on a product space it is therefore not enough to check convergence of the marginal distributions. This suffices only in the case of product measures; cf. Example 4.

The limit of a weakly convergent sequence in $\mathcal{P}(\mathcal{X})$ is unique. Weak convergence induces a topology on $\mathcal{P}(\mathcal{X})$; under this topology, $\mathcal{X}$ being a Polish space, $\mathcal{P}(\mathcal{X})$ is a Polish space, too. Let d be a complete metric compatible with the topology of $\mathcal{X}$; thus $(\mathcal{X},d)$ is a complete and separable metric space. There are different choices for a complete metric on $\mathcal{P}(\mathcal{X})$ that is compatible with the topology of weak convergence. Two common choices are the Prohorov metric and the bounded Lipschitz metric, respectively. The *Prohorov metric* on $\mathcal{P}(\mathcal{X})$ is defined by

$$\rho(\theta,\nu) \doteq \inf\left\{\varepsilon>0 : \theta(G) \le \nu(G^\varepsilon)+\varepsilon \text{ for all closed } G\subset\mathcal{X}\right\}, \qquad (3.1)$$

where $G^\varepsilon \doteq \{x\in\mathcal{X} : d(x,G)<\varepsilon\}$. Notice that ρ is indeed a metric. The *bounded Lipschitz metric* on $\mathcal{P}(\mathcal{X})$ is defined by

$$\tilde{\rho}(\theta,\nu) \doteq \sup\left\{\Big|\int f d\theta - \int f d\nu\Big| : f\in\mathbf{C}_b(\mathcal{X}) \text{ such that } \|f\|_{bL}\le 1\right\}, \quad (3.2)$$

where $\|f\|_{bL} \doteq \sup_{x\in\mathcal{X}} |f(x)| + \sup_{x,y\in\mathcal{X}:x\neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.

The following theorem gives a number of equivalent characterizations of weak convergence.

Theorem 3.1 ("Portemanteau theorem")**.** *Let $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$ be a sequence of probability measures, and let $\theta \in \mathcal{P}(\mathcal{X})$. Then the following are equivalent:*

(i) $\theta_n \xrightarrow{w} \theta$ *as* $n \to \infty$*;*

(ii) $\rho(\theta_n, \theta) \overset{n\to\infty}{\longrightarrow} 0$ *(Prohorov metric);*

(iii) $\tilde{\rho}(\theta_n, \theta) \overset{n\to\infty}{\longrightarrow} 0$ *(bounded Lipschitz metric);*

(iv) $\int f d\theta_n \overset{n\to\infty}{\longrightarrow} \int f d\theta$ *for all bounded Lipschitz continuous* $f\colon \mathcal{X} \to \mathbb{R}$*;*

(v) $\liminf_{n\to\infty} \theta_n(O) \geq \theta(O)$ *for all open* $O \subseteq \mathcal{X}$*;*

(vi) $\limsup_{n\to\infty} \theta_n(G) \leq \theta(G)$ *for all closed* $G \subseteq \mathcal{X}$*;*

(vii) $\lim_{n\to\infty} \theta_n(B) = \theta(B)$ *for all* $B \in \mathcal{B}(\mathcal{X})$ *such that* $\theta(\partial B) = 0$*, where* $\partial B \doteq \mathrm{cl}(B) \cap \mathrm{cl}(B^c)$ *denotes the boundary of the Borel set* B*;*

(viii) $\int f d\theta_n \overset{n\to\infty}{\longrightarrow} \int f d\theta$ *for all* $f \in \mathbf{M}_b(\mathcal{X})$ *such that* $\theta(U_f) = 0$ *where* $U_f \doteq \{x \in \mathcal{X} : f \text{ discontinuous at } x\}$*.*

Proof. For the equivalence of conditions (i), (ii), (iii), and (iv) see Theorem 11.3.3 in Dudley [2002, pp. 395-396].

(i) $\Rightarrow$ (v). Let $O \subseteq \mathcal{X}$ be open. If $f \in \mathbf{C}_b(\mathcal{X})$ is such that $0 \leq f \leq \mathbf{1}_O$, then

$$\liminf_{n\to\infty} \theta_n(O) \geq \int_{\mathcal{X}} f \, d\theta$$

since $\theta_n(O) \geq f$ for every $n \in \mathbb{N}$ and $\int f \, d\theta_n \to \int f \, d\theta$ as $n \to \infty$ by (i). Since O is open, we can find $(f_M)_{M\in\mathbb{N}} \in \mathbf{C}_b(\mathcal{X})$ such that $0 \leq f_M \leq \mathbf{1}_O$ and $f_M \nearrow \mathbf{1}_O$ pointwise as $M \to \infty$. Then $\int f_M \, d\theta \nearrow \theta(O)$ as $M \to \infty$ by monotone convergence. It follows that $\liminf_{n\to\infty} \theta_n(O) \geq \theta(O)$.

(v) $\Leftrightarrow$ (vi). This follows by taking complements (open/closed sets).

(v), (vi) $\Rightarrow$ (vii). Let $B \in \mathcal{B}(\mathcal{X})$. Clearly, $B^\circ \subseteq B \subseteq \mathrm{cl}(B)$ and B° open, $\mathrm{cl}(B)$ closed. Therefore by (v), (vi),

$$\begin{aligned}\theta(B^\circ) \leq \liminf_{n\to\infty} \theta_n(B^\circ) &\leq \liminf_{n\to\infty} \theta_n(B) \\ &\leq \limsup_{n\to\infty} \theta_n(B) \leq \limsup_{n\to\infty} \theta_n(\mathrm{cl}(B)) \leq \theta(\mathrm{cl}(B)).\end{aligned}$$

If $\theta(\partial B) = 0$, then $\theta(B^\circ) = \theta(\mathrm{cl}(B))$, hence $\lim_{n\to\infty} \theta_n(B) = \theta(B)$.

(vii) $\Rightarrow$ (viii). Let $f \in \mathbf{M}_b(\mathcal{X})$ be such that $\theta(U_f) = 0$, and let $A \doteq \{y \in \mathbb{R} : \theta(f^{-1}\{y\}) > 0\}$ be the set of atoms of $\theta \circ f^{-1}$. Since $\theta \circ f^{-1}$ is a finite measure, A is at most countable. Let $\varepsilon > 0$. Then there are $N \in \mathbb{N}$, $y_0, \ldots, y_N \in \mathbb{R} \setminus A$ such that

$$y_0 \leq -\|f\|_\infty < y_1 < \ldots < y_{N-1} < \|f\|_\infty \leq y_N, \qquad |y_i - y_{i-1}| \leq \varepsilon.$$

For $i \in \{1, \ldots, N\}$ set $B_i \doteq f^{-1}\{[y_{i-1}, y_i)\}$. Then

$$\theta(\partial B_i) \leq \theta(f^{-1}\{y_{i-1}\}) + \theta(f^{-1}\{y_i\}) + \theta\{U_f\} = 0.$$

Using (vii) we obtain

$$\limsup_{n\to\infty} \int f \, d\theta_n \leq \limsup_{n\to\infty} \sum_{i=1}^{N} \theta_n(B_i) \cdot y_i = \sum_{i=1}^{N} \theta(B_i) \cdot y_i \leq \varepsilon + \int f \, d\theta.$$

Since ε was arbitrary, it follows that $\limsup_{n\to\infty} \int f \, d\theta_n \leq \int f \, d\theta$. The same argument applied to $-f$ yields the inequality $\liminf_{n\to\infty} \int f \, d\theta_n \geq \int f \, d\theta$.

The implication (viii) $\Rightarrow$ (i) is immediate. □

From Definition 3.1 it is clear that weak convergence is preserved under continuous mappings. The mapping theorem for weak convergence requires continuity only with probability one with respect to the limit measure; this should be compared to characterization (vii) in Theorem 3.1.

Theorem 3.2 (Mapping theorem)**.** *Let $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$, $\theta \in \mathcal{P}(\mathcal{X})$. Let $\mathcal{Y}$ be a second Polish space, and let $\psi : \mathcal{X} \to \mathcal{Y}$ be a measurable mapping. If $\theta_n \xrightarrow{w} \theta$ and $\theta\{x \in \mathcal{X} : \psi$ discontinuous at $x\} = 0$, then $\psi \circ \theta_n \xrightarrow{w} \psi \circ \theta$.*

Proof. By part (v) of Theorem 3.1, it is enough to show that for every $O \subseteq \mathcal{Y}$ open,

$$\liminf_{n\to\infty} \theta_n\left(\psi^{-1}(O)\right) \leq \theta\left(\psi^{-1}(O)\right).$$

Let $O \subseteq \mathcal{Y}$ be open. Set $C \doteq \{x \in \mathcal{X} : \psi$ continuous at $x\}$. Then $\psi^{-1}(O) \cap C$ is contained in $(\psi^{-1}(O))^\circ$, the interior of $\psi^{-1}(O)$. Since $(\psi^{-1}(O))^\circ$ is open, $\theta(C) = 1$ and $\theta_n \xrightarrow{w} \theta$ by hypothesis, it follows from part (v) of Theorem 3.1 that

$$\theta\left(\psi^{-1}(O)\right) = \theta\left((\psi^{-1}(O))^\circ\right) \leq \liminf \theta_n\left((\psi^{-1}(O))^\circ\right) \leq \liminf_{n\to\infty} \theta_n\left(\psi^{-1}(O)\right).$$

□

The following generalization of Theorem 3.2 is sometimes useful.

Theorem 3.3 (Extended mapping theorem)**.** *Let $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$, $\theta \in \mathcal{P}(\mathcal{X})$. Let $\mathcal{Y}$ be a second Polish space, and let ψ_n, $n \in \mathbb{N}$, ψ be a measurable mappings $\mathcal{X} \to \mathcal{Y}$. If $\theta_n \xrightarrow{w} \theta$ and*

$$\theta\{x \in \mathcal{X} : \exists\, (x_n) \subset \mathcal{X} \text{ such that } x_n \to x \text{ while } \psi_n(x_n) \nrightarrow \psi(x)\} = 0,$$

then $\psi_n \circ \theta_n \xrightarrow{w} \psi \circ \theta$.

Proof. See Theorem 5.5 in Billingsley [1968, p. 34] or Theorem 4.27 in Kallenberg [2001, p. 76]. □

In the following version of Fatou's lemma the probability measures converge weakly, while the integrand is fixed (the latter condition can be weakened).

Lemma 3.1 (Fatou)**.** *Let $(\theta_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X})$, $\theta \in \mathcal{P}(\mathcal{X})$. Let $g\colon \mathcal{X} \to (-\infty, \infty]$ be a lower semicontinuous function. If $\theta_n \xrightarrow{w} \theta$, then*

$$\liminf_{n\to\infty} \int_{\mathcal{X}} g\, d\theta_n \geq \int_{\mathcal{X}} g\, d\theta.$$

Proof. See Theorem A.3.12 in Dupuis and Ellis [1997, p. 307]. □

A sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$-valued random variables (possibly defined on different probability spaces) is said to *converge in distribution* to some $\mathcal{X}$-valued random variable X if the respective laws converge weakly.

There is a version of the dominated convergence theorem in connection with convergence in distribution for real-valued (or $\mathbb{R}^d$-valued) random variables.

Theorem 3.4 (Dominated convergence)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}$-valued random variables. Suppose that (X_n) converges in distribution to X for some random variable X. If $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable, then*

$$\mathbf{E}[|X|] < \infty \qquad \text{and} \qquad \mathbf{E}_n[X_n] \overset{n\to\infty}{\longrightarrow} \mathbf{E}[X].$$

Proof. For $M \in \mathbb{N}$, the functions $x \mapsto -M \vee (x \wedge M)$ and $x \mapsto |x| \wedge M$ are in $\mathbf{C}_b(\mathbb{R})$. Since (X_n) converges in distribution to X by hypothesis, this implies

$$\mathbf{E}_n[|X_n| \wedge M] \overset{n\to\infty}{\longrightarrow} \mathbf{E}[|X| \wedge M], \tag{3.3a}$$

$$\mathbf{E}_n[-M \vee (X_n \wedge M)] \overset{n\to\infty}{\longrightarrow} \mathbf{E}[-M \vee (X \wedge M)]. \tag{3.3b}$$

By hypothesis, $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable, that is,

$$\lim_{M\to\infty}\sup_{n\in\mathbb{N}}\mathbf{E}_n\left[|X_n|\cdot\mathbf{1}_{\{|X_n|\geq M\}}\right]=0.$$

This implies, in particular, that $\sup_{n\in\mathbb{N}}\mathbf{E}_n\left[|X_n|\right]<\infty$. By (3.3a), for $M\in\mathbb{N}$ we can choose $n_M\in\mathbb{N}$ such that $|\,\mathbf{E}_{n_M}\left[|X_{n_M}|\wedge M\right]-\mathbf{E}\left[|X|\wedge M\right]|\leq 1$. It follows that

$$\begin{aligned}&\sup_{M\in\mathbb{N}}\mathbf{E}\left[|X|\wedge M\right]\\&\leq\sup_{M\in\mathbb{N}}\left\{|\mathbf{E}_{n_M}\left[|X_{n_M}|\wedge M\right]-\mathbf{E}\left[|X|\wedge M\right]|+\sup_{n\in\mathbb{N}}\mathbf{E}_n\left[|X_n|\wedge M\right]\right\}\\&\leq 1+\sup_{n\in\mathbb{N}}\mathbf{E}_n\left[|X_n|\right]<\infty.\end{aligned}$$

This shows $\mathbf{E}[|X|]<\infty$, that is, X is integrable, since $\mathbf{E}\left[|X|\wedge M\right]\nearrow\mathbf{E}[|X|]$ as $M\to\infty$ by monotone convergence. Now for any $n\in\mathbb{N}$, any $M\in\mathbb{N}$,

$$\begin{aligned}|\mathbf{E}\left[X\right]-\mathbf{E}_n\left[X_n\right]|\leq{}&|\mathbf{E}\left[-M\vee(X\wedge M)\right]-\mathbf{E}_n\left[-M\vee(X_n\wedge M)\right]|\\&+\mathbf{E}\left[|X|\cdot\mathbf{1}_{\{|X|\geq M\}}\right]+\mathbf{E}_n\left[|X_n|\cdot\mathbf{1}_{\{|X_n|\geq M\}}\right]\end{aligned}$$

Let $\varepsilon>0$. By the integrability of X and the uniform integrability of (X_n) one finds $M_\varepsilon\in\mathbb{N}$ such that

$$\mathbf{E}\left[|X|\cdot\mathbf{1}_{\{|X|\geq M\}}\right]+\sup_{k\in\mathbb{N}}\mathbf{E}_k\left[|X_k|\cdot\mathbf{1}_{\{|X_k|\geq M\}}\right]\leq\frac{\varepsilon}{2}.$$

By (3.3a), we can choose $n_\varepsilon=n(M_\varepsilon)$ such that

$$|\mathbf{E}\left[-M_\varepsilon\vee(X\wedge M_\varepsilon)\right]-\mathbf{E}_{n_\varepsilon}\left[-M_\varepsilon\vee(X_{n_\varepsilon}\wedge M_\varepsilon)\right]|\leq\frac{\varepsilon}{2}.$$

It follows that $|\mathbf{E}\left[X\right]-\mathbf{E}_{n_\varepsilon}\left[X_{n_\varepsilon}\right]|\leq\varepsilon$, which establishes the desired convergence since ε was arbitrary. □

Suppose we have a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables that converges in distribution to some random variable X; thus $\mathrm{Law}(X_n)\xrightarrow{w}\mathrm{Law}(X)$. If the relation (in particular, joint distribution) between the $X_1,X_2,\ldots$ is irrelevant, one may work with random variables that converge almost surely.

Theorem 3.5 (Skorohod representation)**.** *Let $(\theta_n)_{n\in\mathbb{N}}\subset\mathcal{P}(\mathcal{X})$. If $\theta_n\xrightarrow{w}\theta$ for some $\theta\in\mathcal{P}(\mathcal{X})$, then there exists a probability space $(\Omega,\mathcal{F},\mathbf{P})$ carrying $\mathcal{X}$-valued random variables X_n, $n\in\mathbb{N}$, and X such that $\mathbf{P}\circ X_n^{-1}=\theta_n$ for every $n\in\mathbb{N}$, $\mathbf{P}\circ X^{-1}=\theta$, and $X_n\to X$ as $n\to\infty$ $\mathbf{P}$-almost surely.*

Proof (for $\mathcal{X}$ finite). Assume that $\mathcal{X} \equiv \{1, \ldots, N\}$ for some $N \in \mathbb{N}$. Set $\Omega \doteq [0,1]$, $\mathcal{F} \doteq \mathcal{B}([0,1])$, and let $\mathbf{P}$ be Lebesgue measure on $\mathcal{B}([0,1])$. Set

$$X(\omega) \doteq \sum_{k=1}^{N} k \cdot \mathbf{1}_{(\theta\{i<k\}, \theta\{i \leq k\}]}(\omega),$$

$$X_n(\omega) \doteq \sum_{k=1}^{N} k \cdot \mathbf{1}_{(\theta_n\{i<k\}, \theta_n\{i \leq k\}]}(\omega).$$

The random variables thus defined have the desired properties since, for every $k \in \{1, \ldots, N\}$, $\theta_n\{i < k\} \to \theta\{i < k\}$, $\theta_n\{i \leq k\} \to \theta\{i \leq k\}$ as $n \to \infty$. For a general Polish space $\mathcal{X}$ (actually, completeness not needed) use countable partitions and approximation argument; cf. the proof of Theorem 4.30 in Kallenberg [2001, p. 79] or of Theorem 11.7.2 in Dudley [2002, pp. 415-417]. □

A standard method for proving that a sequence $(a_n)_{n\in\mathbb{N}}$ of elements of a complete metric space $\mathcal{S}$ converges to a unique limit $a \in \mathcal{S}$ is to proceed as follows. First show that $(a_n)_{n\in\mathbb{N}}$ is relatively compact (i.e., $\mathrm{cl}(\{a_n : n \in \mathbb{N}\})$ is compact in $\mathcal{S}$). Then, taking any convergent subsequence $(a_{n(j)})_{j\in\mathbb{N}}$ with limit $\tilde{a}$, show that $\tilde{a} = a$. This establishes $a_n \to a$ as $n \in \mathbb{N}$. Relative compactness in $\mathcal{P}(\mathcal{X})$ with the topology of weak convergence when $\mathcal{X}$ is Polish is equivalent to uniform exhaustibility by compact sets or "tightness."

Definition 3.2. Let I be a non-empty set. A family $(\theta_i)_{i\in I} \subset \mathcal{P}(\mathcal{X})$ is called *tight* (or *uniformly tight*) if for any $\varepsilon > 0$ there is a compact set $K_\varepsilon \subset \mathcal{X}$ such that

$$\inf_{i\in I} \theta_i(K_\varepsilon) \geq 1 - \varepsilon.$$

Theorem 3.6 (Prohorov)**.** *Let I be a non-empty set, and let $(\theta_i)_{i\in I} \subset \mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is Polish. Then $(\theta_i)_{i\in I} \subset \mathcal{P}(\mathcal{X})$ is tight if and only if $(\theta_i)_{i\in I}$ is relatively compact in $\mathcal{P}(\mathcal{X})$ with respect to the topology of weak convergence.*

Proof. See, for instance, Section 1.5 in Billingsley [1999, pp. 57-65]. □

Depending on the structure of the underlying space $\mathcal{X}$, conditions for tightness or relative compactness can be derived. Let us consider here the case $\mathcal{X} = \boldsymbol{C}([0,\infty), \mathbb{R}^d)$ with the topology of uniform convergence on compact time intervals. With this choice, $\mathcal{X}$ is the canonical path space for ($\mathbb{R}^d$-valued) continuous processes. Let X be the canonical process on $\boldsymbol{C}([0,\infty), \mathbb{R}^d)$, that is, $X(t,\omega) \doteq \omega(t)$ for $t \geq 0$, $\omega \in \boldsymbol{C}([0,\infty), \mathbb{R}^d)$.

Theorem 3.7. *Let I be a non-empty set, and let $(\theta_i)_{i\in I} \subset \mathcal{P}(\boldsymbol{C}([0,\infty),\mathbb{R}^d))$. Then $(\theta_i)_{i\in I}$ is relatively compact if and only if the following two conditions hold:*

(i) $(\theta_i \circ (X(0))^{-1}$ *is tight in* $\mathcal{P}(\mathbb{R}^d)$*, and*

(ii) for every $\varepsilon > 0$*, every* $T \in \mathbb{N}$ *there is* $\delta > 0$ *such that*

$$\sup_{i\in I} \theta_i \left(\left\{ \omega \in \boldsymbol{C}([0,\infty),\mathbb{R}^d) : \boldsymbol{w}_T(\omega,\delta) > \varepsilon \right\} \right) \leq \varepsilon,$$

where $\boldsymbol{w}_T(\omega,\delta) \doteq \sup_{s,t\in[0,T]:|t-s|\leq\delta} |\omega(t)-\omega(s)|$ *is the* modulus of continuity *of* ω *with size* δ *over the time interval* $[0,T]$*.*

Proof. See, for instance, Theorem 2.7.3 in Billingsley [1999, pp. 82-83]; the extension from a compact time interval to $[0,\infty)$ is straightforward.[1] □

Theorem 3.7 should be compared to the Arzelà-Ascoli criterion for relative compactness in $\boldsymbol{C}([0,\infty),\mathbb{R}^d)$. The next theorem gives a sufficient condition for relative compactness (or tightness) in $\mathcal{P}(\boldsymbol{C}([0,\infty),\mathbb{R}^d))$; the result should be compared to the Kolmogorov-Chentsov continuity theorem.

Theorem 3.8 (Kolmogorov's sufficient condition)**.** *Let I be a non-empty set, and let $(\theta_i)_{i\in I} \subset \mathcal{P}(\boldsymbol{C}([0,\infty),\mathbb{R}^d))$. Suppose that*

(i) $(\theta_i \circ (X(0))^{-1}$ *is tight in* $\mathcal{P}(\mathbb{R}^d)$*, and*

(ii) there are strictly positive numbers C, α, β such that for all $t,s \in [0,\infty)$, all $i \in I$,

$$\mathbf{E}_{\theta_i}\left[|X(s)-X(t)|^\alpha\right] \leq C|t-s|^{1+\beta}.$$

Then $(\theta_i)_{i\in I}$ is relatively compact in $\mathcal{P}(\boldsymbol{C}([0,\infty),\mathbb{R}^d))$.

Proof. See, for instance, Corollary 16.5 in Kallenberg [2001, p. 313]. □

Tightness of a family $(\theta_i)_{i\in I} \subset \mathcal{P}(\mathcal{S})$ can often be established by showing that $\sup_{i\in I} G(\theta_i) < \infty$ for an appropriate function G. This works if G is a tightness function in the sense of the definition below.

[1] The issue is more delicate when passing from the Skorohod space $\mathbf{D}([0,T],\mathcal{X})$ to the Skorohod space $\mathbf{D}([0,\infty),\mathcal{X})$; cf. Chapter 3 in Billingsley [1999].

Definition 3.3. A measurable function $g\colon \mathcal{S} \to [0,\infty]$ is called a *tightness function* on $\mathcal{S}$ if g has relatively compact sublevel sets, that is, for every $M \in \mathbb{R}$, $\mathrm{cl}\{x \in \mathcal{S} : g(x) \le M\}$ is compact in $\mathcal{S}$.

A way of constructing tightness functions on $\mathcal{P}(\mathcal{S})$ is provided by the following result.

Theorem 3.9. *Suppose g is a tightness function on $\mathcal{S}$. Define a function G on $\mathcal{P}(\mathcal{S})$ by*

$$G(\theta) \doteq \int_{\mathcal{S}} g(x)\theta(dx).$$

Then G is a tightness function on $\mathcal{S}$.

Proof. The function G is well-defined with values in $[0,\infty]$. Let $M \in [0,\infty)$, and set $A \doteq \{\theta \in \mathcal{P}(\mathcal{S}) : G(\theta) \le M\}$. By Prohorov's theorem it is enough to show that $A = A(M)$ is tight. Let $\varepsilon > 0$. Set $K_\varepsilon \doteq \mathrm{cl}\{x \in \mathcal{S} : g(x) \le M/\varepsilon\}$. Then K_ε is compact since g is a tightness function and for $\theta \in A$,

$$M \ge G(\theta) = \int_{\mathcal{S}} g\, d\theta \ge \int_{\mathcal{S}\setminus K_\varepsilon} g\, d\theta \ge \frac{M}{\varepsilon} \cdot \theta(\mathcal{S} \setminus K_\varepsilon).$$

It follows that $\inf_{\theta\in A} \theta(\mathcal{S} \setminus K_\varepsilon) \le \varepsilon$, hence $\sup_{\theta\in A} \theta(K_\varepsilon) \ge 1 - \varepsilon$. □

4 Relative entropy

Here we collect properties of relative entropy for probability measures on Polish spaces. We will mostly refer to Dupuis and Ellis [1997]; also see the classical works by Kullback and Leibler [1951] and Kullback [1959], where relative entropy is introduced as an information measure and called *directed divergence*. Let $\mathcal{S}$, $\mathcal{X}$, $\mathcal{Y}$ be Polish spaces.

Definition 4.1. Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. The *relative entropy* of μ with respect to ν is given by

$$R(\mu\|\nu) = \begin{cases} \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\nu}(x)\right) \mu(dx) & \text{if } \mu \ll \nu, \\ \infty & \text{else.} \end{cases}$$

Relative entropy is well-defined as a function $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to [0,\infty]$. Indeed, if $\mu \ll \nu$, then a density $f \doteq \frac{d\mu}{d\nu}$ exists by the Radon-Nikodym theorem with f uniquely determined ν-almost surely. In this case,

$$R(\mu\|\nu) = \int_{\mathcal{S}} f(x) \log\left(f(x)\right) \nu(dx).$$

Clearly, $\lim_{x\to 0+} x\log(x) = 0$. Since $\int f d\nu = 1$ and $x\log(x) \geq x - 1$ for all $x \geq 0$ with equality if and only if $x = 1$, it follows that $R(\mu\|\nu) \geq 0$ with $R(\mu\|\nu) = 0$ if and only if $\mu = \nu$. Relative entropy can actually be defined for σ-finite measures on an arbitrary measurable space.

Lemma 4.1 (Basic properties)**.** *Properties of relative entropy $R(.\,\|.\,)$ for probability measures on a Polish space $\mathcal{S}$.*

(a) Relative entropy is a non-negative, convex, lower semicontinuous function $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to [0, \infty]$.

(b) For $\nu \in \mathcal{P}(\mathcal{S})$, $R(.\,\|\nu)$ is strictly convex on $\{\mu \in \mathcal{P}(\mathcal{S} : R(\mu\|\nu) < \infty\}$.

(c) For $\nu \in \mathcal{P}(\mathcal{S})$, $R(.\,\|\nu)$ has compact sublevel sets.

(d) Let $\Pi_{\mathcal{S}}$ denote the set of finite measurable partitions of $\mathcal{S}$. Then for all $\mu, \nu \in \mathcal{P}(\mathcal{S})$,

$$R(\mu\|\nu) = \sup_{\pi\in\Pi_{\mathcal{S}}} \sum_{A\in\pi} \mu(A) \log\left(\frac{\mu(A)}{\nu(A)}\right),$$

where $x\log(x/y) = 0$ if $x = 0$, $x\log(x/y) = \infty$ if $x > 0$ and $y = 0$.

(e) For every $A \in \mathcal{B}(\mathcal{S})$, any $\mu, \nu \in \mathcal{P}(\mathcal{S})$,

$$R(\mu\|\nu) \geq \mu(A) \log\left(\frac{\mu(A)}{\nu(A)}\right) - 1.$$

Proof. See Lemma 1.4.3, parts (b), (c), (g), in Dupuis and Ellis [1997, pp. 29-30]. □

Lemma 4.2 (Contraction property)**.** *Let $\psi\colon \mathcal{Y} \to \mathcal{X}$ be a Borel measurable mapping. Let $\eta \in \mathcal{P}(\mathcal{X})$, $\gamma_0 \in \mathcal{P}(\mathcal{Y})$. Then*

$$R\big(\eta\|\gamma_0 \circ \psi^{-1}\big) = \inf_{\gamma\in\mathcal{P}(\mathcal{Y}):\gamma\circ\psi^{-1}=\eta} R\big(\gamma\|\gamma_0\big), \tag{4.1}$$

where $\inf\emptyset = \infty$ by convention.

Proof (sketch). Inequality "$\leq$" analogous to proof of Lemma E.2.1 in Dupuis and Ellis [1997, p. 366]. For the opposite inequality, check that the probability measure γ defined by $\gamma(dy) \doteq \frac{d\eta}{d\gamma_0\circ\psi^{-1}}(\psi(y))\gamma_0(dy)$ attains the infimum whenever that infimum is finite.[2] □

[2]For more details and an application see, for instance, M. Fischer, On the form of the large deviation rate function for the empirical measures of weakly interacting systems, `arXiv:1208.0472 [math.PR]`.

Lemma 4.2 yields the invariance property of relative entropy established in Lemma E.2.1 [Dupuis and Ellis, 1997, p. 366] in the case when ψ is a bijective bi-measurable mapping; also cf. Theorem 4.1 in Kullback and Leibler [1951], where the inequality that is implied by Lemma 4.2 is derived.

Lemma 4.3 (Chain rule)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ be Polish spaces. Let $\alpha, \beta \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and denote their marginal distributions on $\mathcal{X}$ by α_1 and β_1, respectively. Let $\alpha(.|.)$, $\beta(.|.)$ be stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ such that for all $A \in \mathcal{B}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{Y})$,*

$$\alpha(A \times B) = \int_A \alpha(B|x)\alpha_1(dx), \qquad \beta(A \times B) = \int_A \beta(B|x)\beta_1(dx).$$

Then the mapping $x \mapsto R\left(\alpha(.|x) \| \beta(.|x)\right)$ is measurable and

$$R(\alpha\|\beta) = R\left(\alpha_1\|\beta_1\right) + \int_{\mathcal{X}} R\left(\alpha(.|x)\|\beta(.|x)\right) \alpha_1(dx).$$

In particular, if α, β are product measures, then

$$R(\alpha_1 \otimes \alpha_2 \| \beta_1 \otimes \beta_2) = R\left(\alpha_1\|\beta_1\right) + R\left(\alpha_2\|\beta_2\right).$$

Proof. Appendix C.3 in Dupuis and Ellis [1997, pp. 332-334] □

The variational representation for Laplace functionals given in Lemma 4.4 below is the starting point for the weak convergence approach to large deviations; its statement should be compared to Definition 2.2, the definition of the Laplace principle.

Lemma 4.4 (Laplace functionals)**.** *Let $\nu \in \mathcal{P}(\mathcal{S})$. Then for all $g \in \mathbf{M}_b(\mathcal{S})$,*

$$-\log \int_{\mathcal{S}} \exp\left(-g(x)\right) \nu(dx) = \inf_{\mu \in \mathcal{P}(\mathcal{S})} \left\{ R(\mu\|\nu) + \int_{\mathcal{S}} g(x)\mu(dx) \right\},$$

Infimum in variational formula above is attained at $\mu^* \in \mathcal{P}(\mathcal{S})$ given by

$$\frac{d\mu^*}{d\nu}(x) \doteq \frac{\exp\left(-g(x)\right)}{\int_{\mathcal{S}} \exp\left(-g(y)\right) \nu(dy)}, \quad x \in \mathcal{S}.$$

Proof. Let $g \in \mathbf{M}_b(\mathcal{S})$, and define μ^* through its density with respect to ν as above. Notice that μ^*, ν are mutually absolutely continuous. Let $\mu \in \mathcal{P}(\mathcal{S})$ be such that $R(\mu\|\nu) < \infty$. Then μ is absolutely continuous with respect to

ν with density $\frac{d\mu}{d\nu}$, but also absolutely continuous with respect to μ^* with density $\frac{d\mu}{d\mu^*} = \frac{d\mu}{d\nu} \cdot \frac{d\nu}{d\mu^*}$, where $\frac{d\nu}{d\mu^*} = \frac{e^g}{\int e^g d\mu^*}$. It follows that

$$\begin{aligned} R(\mu\|\nu) + \int_{\mathcal{S}} g\, d\mu &= \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\nu}\right) d\mu + \int_{\mathcal{S}} g\, d\mu \\ &= \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\mu^*}\right) d\mu + \int_{\mathcal{S}} \log\left(\frac{d\mu^*}{d\nu}\right) d\mu + \int_{\mathcal{S}} g\, d\mu \\ &= R(\mu\|\mu^*) - \log \int_{\mathcal{S}} e^{-g}\, d\nu. \end{aligned}$$

This yields the assertion since $R(\mu\|\mu^*) \geq 0$ with $R(\mu\|\mu^*) = 0$ if and only if $\mu = \mu^*$. □

Lemma 4.4 also allows to derive the Donsker-Varadhan variational formula for relative entropy itself.

Lemma 4.5 (Donsker-Varadhan)**.** *Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. Then*

$$R(\mu\|\nu) = \sup_{g \in \mathbf{M}_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g(x)\mu(dx) - \log \int_{\mathcal{S}} \exp\left(g(x)\right) \nu(dx) \right\}.$$

Proof. Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. By Lemma 4.4, for every $g \in \mathbf{M}_b(\mathcal{S})$

$$R(\mu\|\nu) \geq -\int_{\mathcal{S}} g\, d\mu - \log \int_{\mathcal{S}} e^{-g}\, d\nu,$$

hence

$$\begin{aligned} R(\mu\|\nu) &\geq \sup_{g \in \mathbf{M}_b(\mathcal{S})} \left\{ -\int_{\mathcal{S}} g\, d\mu - \log \int_{\mathcal{S}} e^{-g}\, d\nu \right\} \\ &= \sup_{g \in \mathbf{M}_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g\, d\mu - \log \int_{\mathcal{S}} e^{g}\, d\nu \right\}. \end{aligned}$$

For $g \in \mathbf{M}_b(\mathcal{S})$ set $J(g) \doteq \int_{\mathcal{S}} g\, d\mu - \log \int_{\mathcal{S}} e^g d\nu$. Thus $R(\mu\|\nu) \geq \sup_{g \in \mathbf{M}_b} J(g)$. To obtain equality, it is enough to find a sequence $(g_M)_{M \in \mathbb{N}} \subset \mathbf{M}_b(\mathcal{S})$ such that $\limsup_{M\to\infty} J(g_M) = R(\mu\|\nu)$. We distinguish two cases.

First case: μ is not absolutely continuous with respect to ν. Then $R(\mu\|\nu) = \infty$ and there exists $A \in \mathcal{B}(\mathcal{S})$ such that $\mu(A) > 0$ while $\nu(A) = 0$. Choose such a set A and set $g_M \doteq M \cdot \mathbf{1}_A$. Then, for every $M \in \mathbb{N}$, $g_M = 0$ ν-almost surely, thus $\int e^{g_M} d\nu = \int e^0 d\nu = 1$, hence $\log \int e^{g_M} d\nu = 0$. It follows that

$$\limsup_{M\to\infty} J(g_M) = \limsup_{M\to\infty} \int_{\mathcal{S}} g_M\, d\mu = \limsup_{M\to\infty} M \cdot \mu(A) = \infty.$$

Second case: μ is absolutely continuous with respect to ν. Then we can choose a measurable function $f : \mathcal{S} \to [0, \infty)$ such that f is a density for μ with respect to ν (f a version of the Radon-Nikodym derivative $d\mu/d\nu$), and $R(\mu\|\nu) = \int f \cdot \log(f) d\nu$, where the value of the integral is in $[0, \infty]$. Set

$$g_M(x) \doteq \log(f(x)) \cdot \mathbf{1}_{[1/M,M]}(f(x)) - M \cdot \mathbf{1}_{\{0\}}(f(x)), \quad x \in \mathcal{S}.$$

Then

$$\begin{aligned}
&\lim_{M\to\infty} \int_{\mathcal{S}} g_M \, d\mu \\
&= \lim_{M\to\infty} \int_{\mathcal{S}} f \cdot \log(f) \cdot \mathbf{1}_{[1/M,M]}(f) \, d\nu \\
&= \lim_{M\to\infty} \left(\int_{\mathcal{S}} \big(f \cdot \log(f) + \mathbf{1}_{(0,\infty)}(f)\big) \cdot \mathbf{1}_{[1/M,M]}(f) \, d\nu - \int_{\mathcal{S}} \mathbf{1}_{[1/M,M]}(f) \, d\nu \right) \\
&= \int_{\mathcal{S}} f \cdot \log(f) \, d\nu + \nu\{f > 0\} - \nu\{f > 0\} \\
&= R(\mu\|\nu)
\end{aligned}$$

by dominated convergence and monotone convergence since $t \cdot \log(t) \geq -1$ for every $t \geq 0$ and $t \cdot \log(t) = 0$ if $t = 0$, hence, for every $x \in \mathcal{S}$, $(f(x) \cdot \log(f(x)) + \mathbf{1}_{(0,\infty)}(f(x))) \cdot \mathbf{1}_{[1/M,M]}(f(x)) \nearrow f(x) \cdot \log(f(x)) + \mathbf{1}_{(0,\infty)}(f(x))$ as $M \to \infty$. On the other hand, again using dominated and monotone convergence, respectively,

$$\begin{aligned}
&\lim_{M\to\infty} \log \int_{\mathcal{S}} e^{g_M} \, d\nu \\
&= \log \left(\lim_{M\to\infty} \int_{\mathcal{S}} \big(f \cdot \mathbf{1}_{[1/M,M]}(f) + \mathbf{1}_{(0,1/M)\cup(M,\infty)}(f) + e^{-M} \cdot \mathbf{1}_{\{0\}}(f)\big) \, d\nu \right) \\
&= \log \int_{\mathcal{S}} f \, d\nu = \log(1) = 0.
\end{aligned}$$

It follows that $\limsup_{M\to\infty} J(g_M) = \lim_{M\to\infty} \int_{\mathcal{S}} g_M \, d\mu = R(\mu\|\nu)$. □

Remark 4.1. If the state space $\mathcal{S}$ is Polish as we assume, then the supremum in the Donsker-Varadhan formula of Lemma 4.5 can be restricted to bounded and continuous functions, that is,

$$\begin{aligned}
&\sup_{g\in\mathbf{M}_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g(x)\mu(dx) - \log \int_{\mathcal{S}} \exp\big(g(x)\big) \, \nu(dx) \right\} \\
&= \sup_{g\in\mathbf{C}_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g(x)\mu(dx) - \log \int_{\mathcal{S}} \exp\big(g(x)\big) \, \nu(dx) \right\};
\end{aligned}$$

see the proof of formula (C.1) in Dupuis and Ellis [1997, pp. 329-330].

Remark 4.2. Lemmata 4.4 and 4.5 imply a relationship of convex duality between Laplace functionals and relative entropy. Let $\nu \in \mathcal{P}(\mathcal{S})$. Then

$$R(\mu\|\nu) = \sup_{g \in \mathbf{M}_b(\mathcal{X})} \left\{ \int_{\mathcal{S}} g \, d\mu - \log \int_{\mathcal{S}} e^g \, d\nu \right\}, \quad \mu \in \mathcal{P}(\mathcal{S}),$$

$$\log \int_{\mathcal{S}} e^g \, d\nu = \sup_{\mu \in \mathcal{P}(\mathcal{S})} \left\{ \int_{\mathcal{S}} g \, d\mu - R(\mu\|\nu) \right\}, \quad g \in \mathbf{M}_b(\mathcal{S}),$$

that is, the functions $\mu \mapsto R(\mu\|\nu)$ and $g \mapsto \log \int e^g d\nu$ are convex conjugates.

References

P. Billingsley. *Convergence of Probability Measures*. Wiley series in Probability and Statistics. John Wiley & Sons, New York, 1968.

P. Billingsley. *Convergence of Probability Measures*. Wiley series in Probability and Statistics. John Wiley & Sons, New York, 2nd edition, 1999.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics*. Springer, New York, 2nd edition, 1998.

F. den Hollander. *Large Deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.

J.-D. Deuschel and D. W. Stroock. *Large Deviations*. Academic Press, Boston, 1989.

R. Dudley. *Real Analysis and Probability*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, 2002.

P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1997.

R. S. Ellis. *Entropy, Large Deviations and Statistical Mechanics*, volume 271 of *Grundlehren der mathematischen Wissenschaften*. Springer, New York, 1985.

S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1986.

O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer, New York, 2nd edition, 2001.

S. Kullback. *Information Theory and Statistics*. John Wiley and Sons, Inc., New York, 1959.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.

S. R. S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.*, 19:261–286, 1966.