Chapter 2
Probability Distances and Probability Metrics: Definitions

The goals of this chapter are to:

- Provide examples of metrics in probability theory;
- Introduce formally the notions of a probability metric and a probability distance;
- Consider the general setting of random variables (RVs) defined on a given probability space $(\Omega, \mathcal{A}, \Pr)$ that can take values in a separable metric space U in order to allow for a unified treatment of problems involving random elements of a general nature;
- Consider the alternative setting of probability distances on the space of probability measures $\mathcal{P}_2$ defined on the σ-algebras of Borel subsets of $U^2 = U \times U$, where U is a separable metric space;
- Examine the equivalence of the notion of a probability distance on the space of probability measures $\mathcal{P}_2$ and on the space of joint distributions $\mathcal{L}\mathfrak{X}_2$ generated by pairs of RVs (X, Y) taking values in a separable metric space U.

S.T. Rachev et al., *The Methods of Distances in the Theory of Probability and Statistics*,
DOI 10.1007/978-1-4614-4869-3_2,

Notation introduced in this chapter:

Notation	Description
EN	Engineer's metric
$\mathfrak{X}^p$	Space of real-valued random variables with $E\|X\|^p < \infty$
$\boldsymbol{\rho}$	Uniform (Kolmogorov) metric
$\mathfrak{X} = \mathfrak{X}(\mathbb{R})$	Space of real-valued random variables
L	Lévy metric
$\boldsymbol{\kappa}$	Kantorovich metric
$\boldsymbol{\theta}_p$	L_p-metric between distribution functions
K, **K***	Ky Fan metrics
$\mathcal{L}_p$	L_p-metric between random variables
$\mathbf{MOM}_p$	Metric between pth moments
(S, ρ)	Metric space with metric ρ
$\mathbb{R}^n$	n-dimensional vector space
$r(C_1, C_2)$	Hausdorff metric (semimetric between sets)
$s(F, G)$	Skorokhod metric
$\mathbb{K} = \mathbb{K}_\rho$	Parameter of a distance space
$\mathcal{H}$	Class of Orlicz's functions
ρ_H	Birnbaum–Orlicz distance
Kr	Kruglov distance
(U, d)	Separable metric space with metric d
s.m.s.	Separable metric space
U^k	k-fold Cartesian product of U
$\mathcal{B}_k = \mathcal{B}_k(U)$	Borel σ-algebra on U^k
$\mathcal{P}_k = \mathcal{P}_k(U)$	Space of probability laws on $\mathcal{B}_k$
$T_{\alpha,\beta,\dots,\gamma}P$	Marginal of $P \in \mathcal{P}_k$ on coordinates $\alpha, \beta, \dots, \gamma$
$\Pr_X$	Distribution of X
μ	Probability semidistance
$\mathfrak{X} := \mathfrak{X}(U)$	Set of U-valued RVs
$\mathcal{L}\mathfrak{X}_2 := \mathcal{L}\mathfrak{X}_2(U)$	Space of $\Pr_{X,Y}$, $X, Y \in \mathfrak{X}(U)$
u.m.	Universally measurable
u.m.s.m.s.	Universally measurable separable metric space

2.1 Introduction

Generally speaking, a functional that measures the distance between random quantities is called a *probability metric*.[1] In this chapter, we provide different examples of probability metrics and discuss an application of the Kolmogorov

[1] Mostafaei and Kordnourie (2011) is a more recent general publication on probability metrics and their applications.

metric in mathematical statistics. Then we proceed with the axiomatic construction of probability metrics on the space of probability measures defined on the twofold Cartesian product of a separable metric space U. This definition induces by restriction a probability metric on the space of joint distributions of random elements defined on a probability space $(\Omega, \mathcal{A}, \Pr)$ and taking values in the space U. Finally, we demonstrate that under some fairly general conditions, the two constructions are essentially the same.

2.2 Examples of Metrics in Probability Theory

Below is a list of various metrics commonly found in probability and statistics.

1. *Engineer's metric*:

$$\mathbf{EN}(X,Y) := |\mathbb{E}(X) - \mathbb{E}(Y)|, \quad X,Y \in \mathfrak{X}^1, \tag{2.2.1}$$

   where $\mathfrak{X}^p$ is the space of all real-valued RVs) with $\mathbb{E}|X|^p < \infty$.
2. *Uniform (or Kolmogorov) metric*:

$$\boldsymbol{\rho}(X,Y) := \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\}, \quad X,Y \in \mathfrak{X} = \mathfrak{X}(\mathbb{R}), \tag{2.2.2}$$

   where F_X is the distribution function (DF) of X, $\mathbb{R} = (-\infty, +\infty)$, and $\mathfrak{X}$ is the space of all real-valued RVs.
3. *Lévy metric*:

$$\mathbf{L}(X,Y) := \inf\{\varepsilon > 0 : F_X(x-\varepsilon) - \varepsilon \le F_Y(x) \le F_X(x+\varepsilon) + \varepsilon, \quad \forall x \in \mathbb{R}\}. \tag{2.2.3}$$

Remark 2.2.1. We see that $\boldsymbol{\rho}$ and $\mathbf{L}$ may actually be considered metrics on the space of all distribution functions. However, this cannot be done for **EN** simply because $\mathbf{EN}(X,Y) = 0$ does not imply the coincidence of F_X and F_Y, while $\boldsymbol{\rho}(X,Y) = 0 \iff \mathbf{L}(X,Y) = 0 \iff F_X = F_Y$. The Lévy metric metrizes weak convergence (convergence in distribution) in the space $\mathcal{F}$, whereas $\boldsymbol{\rho}$ is often applied in the central limit theorem (CLT).[2]

4. *Kantorovich metric*:

$$\boldsymbol{\kappa}(X,Y) = \int_{\mathbb{R}} |F_X(x) - F_Y(x)| \mathrm{d}x, \qquad X,Y \in \mathfrak{X}^1.$$

[2] See Hennequin and Tortrat (1965).

5. *L_p-metrics between distribution functions*:

$$\boldsymbol{\theta}_p(X,Y) := \left(\int_{-\infty}^{\infty} |F_X(t) - F_Y(t)|^p \mathrm{d}t\right)^{1/p}, \quad p \geq 1, \quad X, Y \in \mathfrak{X}^1. \quad (2.2.4)$$

Remark 2.2.2. Clearly, $\boldsymbol{\kappa} = \boldsymbol{\theta}_1$. Moreover, we can extend the definition of $\boldsymbol{\theta}_p$ when $p = \infty$ by setting $\boldsymbol{\theta}_\infty = \boldsymbol{\rho}$. One reason for this extension is the following dual representation for $1 \leq p \leq \infty$:

$$\theta_p(X,Y) = \sup_{f \in \mathcal{F}_p} |Ef(X) - Ef(Y)|, \quad X, Y \in \mathfrak{X}^1,$$

where $\mathcal{F}_p$ is the class of all measurable functions f with $\|f\|_q < 1$. Here, $\|f\|_q (1/p + 1/q = 1)$ is defined, as usual, by[3]

$$\|f\|_q := \begin{cases} \left(\int |f|^q\right)^{1/q}, \ 1 \leq q < \infty, \\ \operatorname*{ess\,sup}_{\mathbb{R}} |f|, \quad q = \infty. \end{cases}$$

6. *Ky Fan metrics*:

$$\mathbf{K}(X,Y) := \inf\{\varepsilon > 0 : \Pr(|X - Y| > \varepsilon) < \varepsilon\}, \qquad X, Y \in \mathfrak{X}, \qquad (2.2.5)$$

and

$$\mathbf{K}^*(X,Y) := E\frac{|X - Y|}{1 + |X - Y|}. \qquad (2.2.6)$$

Both metrics metrize convergence in probability on $\mathfrak{X} = \mathfrak{X}(\mathbb{R})$, the space of real RVs.[4]

7. *L_p-metric*:

$$\mathcal{L}_p(X,Y) := \{E|X - Y|^p\}^{1/p}, \quad p \geq 1, \quad X, Y \in \mathfrak{X}^p. \qquad (2.2.7)$$

Remark 2.2.3. Define

$$m^p(X) := \{E|X|^p\}^{1/p}, \quad p > 1, \quad X \in \mathfrak{X}^p. \qquad (2.2.8)$$

and

$$\mathbf{MOM}_p(X,Y) := |m^p(X) - m^p(Y)|, \quad p \geq 1, \quad X, Y \in \mathfrak{X}^p. \qquad (2.2.9)$$

[3]The proof of this representation is given by (Dudley, 2002, p. 333) for the case $p = 1$.

[4]See Lukacs (1968, Chap. 3) and Dudley (1976, Theorem 3.5).

Then we have, for $X_0, X_1, \ldots \in \mathfrak{X}^p$,

$$\mathcal{L}_p(X_n, X_0) \to 0 \iff \begin{cases} \mathbf{K}(X_n, X_0) \to 0, \\ \mathbf{MOM}_p(X_n, X_0) \to 0 \end{cases} \tag{2.2.10}$$

[see, e.g., Lukacs (1968, Chap. 3)].

Other probability metrics in common use include the discrepancy metric, the Hellinger distance, the relative entropy metric, the separation distance metric, the χ^2-distance, and the f-divergence metric. These probability metrics are summarized in Gibbs and Su (2002).

All of the aforementioned (semi-)metrics on subsets of $\mathfrak{X}$ may be divided into three main groups: primary, simple, and compound (semi-)metrics. A metric μ is *primary* if $\mu(X, Y) = 0$ implies that certain moment characteristics of X and Y agree. As examples, we have **EN** (2.2.1) and $\mathbf{MOM}_p$ (2.2.9). For these metrics

$$\begin{aligned} \mathbf{EN}(X, Y) = 0 &\iff EX = EY, \\ \mathbf{MOM}_p(X, Y) = 0 &\iff m^p(X) = m^p(Y). \end{aligned} \tag{2.2.11}$$

A metric μ is *simple* if

$$\mu(X, Y) = 0 \iff F_X = F_Y. \tag{2.2.12}$$

Examples are $\boldsymbol{\rho}$ (2.2.2), **L** (2.2.3), and $\boldsymbol{\theta}_p$ (2.2.4). The third group, the *compound* (semi-)metrics, has the property

$$\mu(X, Y) = 0 \iff \Pr(X = Y) = 1. \tag{2.2.13}$$

Some examples are **K** (2.2.5), $\mathbf{K}^*$ (2.2.6), and $\mathcal{L}_p$ (2.2.7).

Later on, precise definitions of these classes will be given as well as a study of the relationships between them. Now we will begin with a common definition of probability metric that will include the types mentioned previously.

2.3 Kolmogorov Metric: A Property and an Application

In this section, we consider a paradoxical property of the Kolmogorov metric and an application in the area of mathematical statistics.

Consider the metric space $\mathfrak{F}$ of all one-dimensional distributions metrized by the Kolmogorov distance

$$\boldsymbol{\rho}(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|, \tag{2.3.1}$$

which we define now in terms of the elements of $\mathfrak{F}$ rather than in terms of RVs as in the definition in (2.2.2). Denote by $B(F, r)$ an open ball of radius $r > 0$ centered

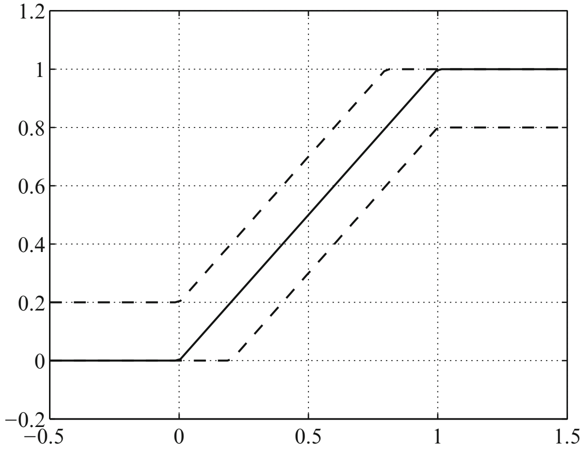

Fig. 2.1 The ball $B(F_o, \delta_\alpha)$. The *solid line* is the center of the ball and the *dashed line* represents the boundary of the ball

at F in the metric space $\mathfrak{F}$ with $\boldsymbol{\rho}$-distance and let F_o be a continuous distribution function (DF). The following result holds.

Theorem 2.3.1. *For any $r > 0$ there exists a continuous DF F_r such that*

$$B(F_r, r) \subset B(F_o, r) \tag{2.3.2}$$

and

$$B(F_r, r) \neq B(F_o, r).$$

Proof. Let us show that there are F_o and F_a such that (2.3.2) holds. Without loss of generality we may choose

$$F_o(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

For a given (but fixed) n define δ_α such that (2.3.1) is true.

Figure 2.1 provides an illustration of the ball $B(F_o, \delta_\alpha)$. The boundary of the ball is shown by means of a dashed line, the center of the ball is the solid line, and the radius δ_α equals 0.2.

Consider now F_a defined in the following way:

$$F_a(x) = \begin{cases} 0, & x < \delta_\alpha/2, \\ 2x - \delta_\alpha, & \delta_\alpha/2 \le x < \delta_\alpha, \\ x, & \delta_\alpha \le x < 1 - \delta_\alpha, \\ 2x - (1 - \delta_\alpha), & 1 - \delta_\alpha \le x < 1 - \delta_\alpha/2, \\ 1, & x \ge 1 - \delta_\alpha/2. \end{cases}$$

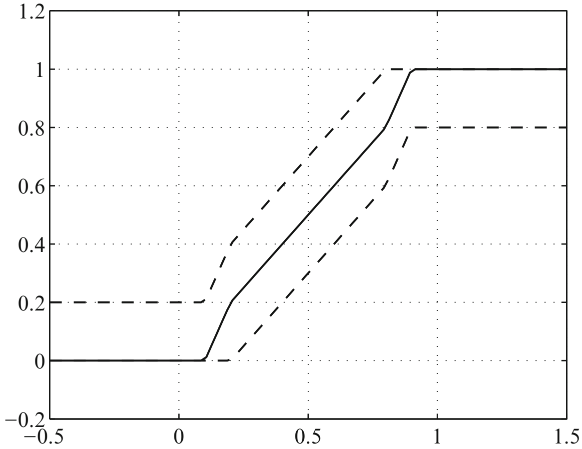

Fig. 2.2 The ball $B(F_a, \delta_\alpha)$. The *solid line* is the center of the ball and the *dashed line* represents the boundary of the ball

An illustration is given in Fig. 2.2. Comparing Figs. 2.1 and 2.2, we can see that

$$B(F_a, \delta_\alpha) \subset B(F_o, \delta_\alpha)$$

and

$$B(F_a, \delta_\alpha) \neq B(F_o, \delta_\alpha). \qquad \square$$

We demonstrate that this property leads to biasedness of the Kolmogorov goodness-of-fit tests. Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) RVs (observations) with (unknown) DF F. Based on the observations, one needs to test the hypothesis

$$H_o : \ F = F_o,$$

where F_o is a fixed DF.

Definition 2.3.1. For a specific alternative hypothesis, a test is said to be unbiased if the probability of rejecting the null hypothesis

(a) Is greater than or equal to the significance level α when the alternative is true and
(b) Is less than or equal to the significance level when the null hypothesis is true.

A test is said to be biased for an alternative hypothesis if it is not unbiased for this alternative.

Let d be a distance in the space of all probability distributions on the real line. Below we consider a test with the following properties:

1. We reject the null hypothesis H_o if

$$d(G_n, F_o) > \delta_\alpha,$$

where G_n is an empirical DF constructed on the basis of the observations $X_1, \dots, X_n$ and δ_α satisfies

$$\Pr\{d(G_n, F_o) > \delta_\alpha\} \le \alpha. \tag{2.3.3}$$

2. The test is distribution free, i.e.,

$$\Pr_F\{d(G_n, F_o) > \delta_\alpha\}$$

does not depend on continuous DF F.

We refer to such tests as *distance-based tests*.

Theorem 2.3.2. *Suppose that for some* $\alpha > 0$ *there exists a continuous DF* F_a *such that*

$$B(F_a, \delta_\alpha) \subset B(F_o, \delta_\alpha) \tag{2.3.4}$$

and

$$\Pr_{F_o}\{G_n \in B(F_o, \delta_\alpha) \setminus B(F_a, \delta_\alpha)\} > 0. \tag{2.3.5}$$

Then the distance-based test is biased for the alternative F_a.

Proof. Let $X_1, \dots, X_n$ be a sample from F_a and G_n be the corresponding empirical DF. Then

$$\Pr_{F_o}\{G_n \in B(F_o, \delta_\alpha)\} \ge 1 - \alpha.$$

In view of (2.3.4) and (2.3.5), we have

$$\Pr_{F_o}\{G_n \in B(F_o, \delta_\alpha)\} > 1 - \alpha,$$

that is,

$$\Pr_{F_o}\{d(G_n, F_o) > \delta_\alpha\} < \alpha.$$ □

Now let us consider the Kolmogorov goodness-of-fit test. Clearly, it is a distance-based test for the distance

$$d(F, G) = \boldsymbol{\rho}(F, G).$$

From Theorem 2.3.1 it follows that (2.3.4) holds. The relation (2.3.5) is almost obvious. From Theorem 2.3.2 it follows that the Kolmogorov goodness-of-fit test is biased.

Remark 2.3.1. The biasedness of the Kolmogorov goodness-of-fit test is a known fact.[5] The same property holds for the Cramér–von Mises goodness-of-fit test.[6]

[5] See Massey (1950) and Thompson (1979).

[6] See Thompson (1966).

2.4 Metric and Semimetric Spaces, Distance, and Semidistance Spaces

Let us begin by recalling the notions of metric and semimetric spaces. Generalizations of these notions will be needed in the theory of probability metrics (TPM).

Definition 2.4.1. A set $S := (S, \rho)$ is said to be a *metric* space with the metric ρ if ρ is a mapping from the product $S \times S$ to $[0, \infty)$ having the following properties for each $x, y, z \in S$:

(1) *Identity property*: $\rho(x, y) = 0 \iff x = y$;
(2) *Symmetry*: $\rho(x, y) = \rho(y, x)$;
(3) *Triangle inequality*: $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$.

Here are some well-known examples of metric spaces:

(a) *The n-dimensional vector space* $\mathbb{R}^n$ endowed with the metric $\rho(x, y) := \|x - y\|_p$, where

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\min(1,1/p)}, \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad 0 < \rho < \infty,$$

$$\|x\|_\infty := \sup_{1 \leq i \leq n} |x_i|.$$

(b) *The Hausdorff metric between closed sets*

$$r(C_1, C_2) = \max\left\{\sup_{x_1 \in C_1} \inf_{x_2 \in C_2} \rho(x_1, x_2), \sup_{x_2 \in C_2} \inf_{x_1 \in C_1} \rho(x_1, x_2)\right\},$$

where the C_i are closed sets in a bounded metric space (S, ρ).[7]

(c) *The H-metric*. Let $D(\mathbb{R})$ be the space of all bounded functions $f : \mathbb{R} \to \mathbb{R}$, continuous from the right and having limits from the left, $f(x-) = \lim_{t \uparrow x} f(t)$. For any $f \in D(\mathbb{R})$ define the graph Γ_f as the union of the sets $\{(x, y) : x \in \mathbb{R}, y = f(x)\}$ and $\{(x, y) : x \in \mathbb{R}, y = f(x-)\}$. The H-metric $H(f, g)$ in $D(\mathbb{R})$ is defined by the Hausdorff distance between the corresponding graphs, $H(f, g) := r(\Gamma_f, \Gamma_g)$. Note that in the space $\mathcal{F}(\mathbb{R})$ of distribution functions, H metrizes the same convergence as the *Skorokhod metric*:

[7]See Hausdorff (1949).

$$s(F,G) = \inf\Big\{\varepsilon > 0 : \text{there exists a strictly increasing continuous}$$
$$\text{function } \lambda : \mathbb{R} \to \mathbb{R} \text{ such that } \lambda(\mathbb{R}) = \mathbb{R},\ \sup_{t\in\mathbb{R}} |\lambda(t) - t| < \varepsilon,$$
$$\text{and } \sup_{t\in\mathbb{R}} |F(\lambda(t)) - G(t)| < \varepsilon\Big\}.$$

Moreover, H-convergence in $\mathcal{F}$ implies convergence in distributions (the weak convergence). Clearly, $\boldsymbol{\rho}$-convergence [see (2.2.2)] implies H-convergence.[8]

If the identity property in Definition 2.4.1 is weakened by changing property (1) to

$$x = y \Rightarrow \rho(x,y) = 0, \tag{1*}$$

then S is said to be a *semimetric space* (or *pseudometric space*) and ρ a *semimetric* (or *pseudometric*) in S. For example, the Hausdorff metric r is only semimetric in the space of all Borel subsets of a bounded metric space (S,ρ).

Obviously, in the space of real numbers, **EN** [see (2.2.1)] is the usual uniform metric on the real line $\mathbb{R}$ [i.e., $\mathbf{EN}(a,b) := |a-b|$, $a,b \in \mathbb{R}$]. For $p \geq 0$, define $\mathcal{F}^p$ as the space of all distribution functions F with $\int_{-\infty}^0 F(x)^p \mathrm{d}x + \int_0^\infty (1 - F(x))^p \mathrm{d}x < \infty$. The distribution function space $\mathcal{F} = \mathcal{F}^0$ can be considered a metric space with metrics $\boldsymbol{\rho}$ and $\mathbf{L}$, while $\boldsymbol{\theta}_p (1 \leq p < \infty)$ is a metric in $\mathcal{F}^p$. The Ky Fan metrics [see (2.2.5), (2.2.6)], resp. $\mathcal{L}_p$-metric [see (2.2.7)], may be viewed as semimetrics in $\mathfrak{X}$ (resp. $\mathfrak{X}^1$) as well as metrics in the space of all Pr-equivalence classes

$$\widetilde{X} := \{Y \in \mathfrak{X} : \Pr(Y = X) = 1\}, \quad \forall X \in \mathfrak{X} \text{ [resp. } \mathfrak{X}^p]. \tag{2.4.1}$$

EN, $\mathbf{MOM}_p$, $\boldsymbol{\theta}_p$, and $\mathcal{L}_p$ can take infinite values in $\mathfrak{X}$, so we will assume, in the next generalization of the notion of metric, that ρ may take infinite values; at the same time, we will also extend the notion of triangle inequality.

Definition 2.4.2. The set **S** is called a *distance space* with distance ρ and parameter $\mathbb{K} = \mathbb{K}_\rho$ if ρ is a function from $S \times S$ to $[0,\infty]$, $\mathbb{K} \geq 1$, and for each $x,y,z \in S$ the identity property (1) and the symmetry property (2) hold, as does the following version of the triangle inequality: (3*) (*Triangle inequality with parameter* $\mathbb{K}$)

$$\rho(x,y) \leq \mathbb{K}[\rho(x,z) + \rho(z,y)]. \tag{2.4.2}$$

If, in addition, the identity property (1) is changed to (1*), then S is called a semidistance space and ρ is called a semidistance (with parameter $\mathbb{K}_\rho$).

Here and in what follows we will distinguish the notions *metric* and *distance*, using *metric* only in the case of *distance with parameter* $\mathbb{K} = 1$, *taking finite or infinite values*.

[8] A more detailed analysis of the metric H will be given in Sect. 4.2.

Remark 2.4.1. It is not difficult to check that each distance ρ generates a topology in S with a basis of open sets $B(a,r) := \{x \in S; \rho(x,a) < r\}, \in S, r > 0$. We know, of course, that every metric space is normal and that every separable metric space has a countable basis. In much the same way, it is easily shown that the same is true for distance space. Hence, by Urysohn's metrization theorem,[9] every separable distance space is metrizable.

Actually, distance spaces have been used in functional analysis for a long time, as shown by the following examples.

Example 2.4.1. Let $\mathcal{H}$ be the class of all nondecreasing continuous functions H from $[0,\infty)$ onto $[0,\infty)$, which vanish at the origin and satisfy Orlicz's condition

$$K_H := \sup_{t>0} \frac{H(2t)}{H(t)} < \infty. \tag{2.4.3}$$

Then $\widetilde{\rho} := H(\rho)$ is a distance in S for each metric ρ in S and $\mathbb{K}_{\widetilde{\rho}} = K_H$.

Example 2.4.2. The *Birnbaum–Orlicz space* L^H $(H \in \mathcal{H})$ consists of all integrable functions on $[0,1]$ endowed with the Birnbaum–Orlicz distance[10]

$$\rho_H(f_1, f_2) := \int_0^1 H(|f_1(x) - f_2(x)|)\mathrm{d}x. \tag{2.4.4}$$

Obviously, $\mathbb{K}_{\rho_H} = K_H$.

Example 2.4.3. Similarly to (2.4.4), Kruglov (1973) introduced the following distance in the space of distribution functions:

$$\mathbf{Kr}(F,G) = \int \phi(F(x) - G(x))\mathrm{d}x, \tag{2.4.5}$$

where the function ϕ satisfies the following conditions:

(a) ϕ is even and strictly increasing on $[0,\infty)$, $\phi(0) = 0$;
(b) For any x and y and some fixed $A \geq 1$

$$\phi(x+y) \leq A(\phi(x) + \phi(y)). \tag{2.4.6}$$

Obviously, $\mathbb{K}_{\mathbf{Kr}} = A$.

[9] See Dunford and Schwartz (1988, Theorem 1.6.19).

[10] Birnbaum and Orliz (1931) and Dunford and Schwartz (1988, p. 400)

2.5 Definitions of Probability Distance and Probability Metric

Let U be a separable metric space (s.m.s.) with metric d, $U^k = \underbrace{U \times \cdots \times U}_{k \text{ times}}$ the k-fold Cartesian product of U, and $\mathcal{P}_k = \mathcal{P}_k(U)$ the space of all probability measures defined on the σ-algebra $\mathcal{B}_k = \mathcal{B}_k(U)$ of Borel subsets of U^k. We will use the terms *probability measure* and *law* interchangeably. For any set $\{\alpha, \beta, \dots, \gamma\} \subseteq \{1, 2, \dots, k\}$ and for any $P \in \mathcal{P}_k$ let us define the marginal of P on the coordinates $\alpha, \beta, \dots, \gamma$ by $T_{\alpha,\beta,\dots,\gamma}P$. For example, for any Borel subsets A and B of U, $T_1 P(A) = P(A \times U \times \cdots \times U)$, $T_{1,3}P(A \times B) = P(A \times U \times B \times \cdots \times U)$. Let $\mathbb{B}$ be the operator in U^2 defined by $\mathbb{B}(x, y) := (y, x)$ $(x, y \in U)$. All metrics $\mu(X, Y)$ cited in Sect. 2.2 [see (2.2.1)–(2.2.9)] are completely determined by the joint distributions $\Pr_{X,Y}$ $(\Pr_{X,Y} \in \mathcal{P}_2(\mathbb{R}))$ of the RVs $X, Y \in \mathfrak{X}(\mathbb{R})$.

In the next definition we will introduce the notion of probability distance, and thus we will describe the primary, simple, and compound metrics in a uniform way. Moreover, the space where the RVs X and Y take values will be extended to U, an arbitrary s.m.s.

Definition 2.5.1. A mapping μ defined on $\mathcal{P}_2$ and taking values in the extended interval $[0, \infty]$ is said to be a *probability semidistance with parameter* $\mathbb{K} := \mathbb{K}_\mu \geq 1$ (or *p. semidistance* for short) in $\mathcal{P}_2$ if it possesses the following three properties:

(1) (*Identity property* (**ID**)). If $P \in \mathcal{P}_2$ and $P(\cup_{x \in U}\{(x, x)\}) = 1$, then $\mu(P) = 0$;
(2) (*Symmetry* (**SYM**)). If $P \in \mathcal{P}_2$, then $\mu(P \circ \mathbb{B}^{-1}) = \mu(P)$;
(3) (*Triangle inequality* (**TI**)). If $P_{13}, P_{12}, P_{23} \in \mathcal{P}_2$ and there exists a law $Q \in \mathcal{P}_3$ such that the following "consistency" condition holds:

$$T_{13}Q = P_{13}, \qquad T_{12}Q = P_{12}, \qquad T_{23}Q = P_{23}, \tag{2.5.1}$$

then

$$\mu(P_{13}) \leq \mathbb{K}[\mu(P_{12}) + \mu(P_{23})].$$

If $\mathbb{K} = 1$, then μ is said to be a *p. semimetric*. If we strengthen the condition **ID** to

$\widetilde{\mathbf{ID}}$: if $P \in P_2$, then

$$P(\cup\{(x, x) : x \in U\}) = 1 \iff \mu(P) = 0,$$

then we say that μ is a *probability distance with parameter* $\mathbb{K} = \mathbb{K}_\mu \geq 1$ (or *p. distance* for short).

Definition 2.5.1 acquires a visual form in terms of RVs, namely: let $\mathfrak{X} := \mathfrak{X}(U)$ be the set of all RVs on a given probability space $(\Omega, \mathcal{A}, \Pr)$ taking values in $(U, \mathcal{B}_1)$. By $\mathcal{L}\mathfrak{X}_2 := \mathcal{L}\mathfrak{X}_2(U) := \mathcal{L}\mathfrak{X}_2(U; \Omega, \mathcal{A}, \Pr)$ we denote the space of all joint distributions $\Pr_{X,Y}$ generated by the pairs $X, Y \in \mathfrak{X}$. Since $\mathcal{L}\mathfrak{X}_2 \subseteq \mathcal{P}_2$, the notion of a p. (semi-)distance is naturally defined on $\mathcal{L}\mathfrak{X}_2$. Considering μ on the subset $\mathcal{L}\mathfrak{X}_2$, we will put

$$\mu(X,Y) := \mu(\Pr_{X,Y})$$

and call μ a *p. semidistance on* $\mathfrak{X}$. If μ is a p. distance, then we use the phrase *p. distance* on $\mathfrak{X}$. Each p. semidistance μ on $\mathfrak{X}$ is a semidistance on $\mathfrak{X}$ in the sense of Definition 2.4.2.[11] Then the relationships **ID**, $\widetilde{\mathbf{ID}}$, **SYM**, and **TI** have simple "metrical" interpretations:

$$\mathbf{ID}^{(*)} \qquad \Pr(X=Y)=1 \Rightarrow \mu(X,Y)=0,$$

$$\widetilde{\mathbf{ID}}^{(*)} \qquad \Pr(X=Y)=1 \iff \mu(X,Y)=0,$$

$$\mathbf{SYM}^{(*)} \quad \mu(X,Y)=\mu(Y,X),$$

$$\mathbf{TI}^{(*)} \qquad \mu(X,Z) < \mathbb{K}[\mu(X,Z)+\mu(Z,Y)].$$

Definition 2.5.2. A mapping $\mu : \mathcal{L}\mathfrak{X}_2 \to [0,\infty]$ is said to be a *p. semidistance* on $\mathfrak{X}$ (resp. *distance*) *with parameter* $\mathbb{K} := \mathbb{K}_\mu \geq 1$ if $\mu(X,Y) = \mu(\Pr_{X,Y})$ satisfies the properties $\mathbf{ID}^{(*)}$ [resp. $\widetilde{\mathbf{ID}}^{(*)}$], $\mathbf{SYM}^{(*)}$, and $\mathbf{TI}^{(*)}$ for all RVs $X, Y, Z \in \mathfrak{X}(U)$.

Example 2.5.1. Let $H \in \mathcal{H}$ (Example 2.4.1) and (U,d) be an s.m.s. Then $\mathcal{L}_H(X,Y) = EH(d(Z,V))$ is a p. distance in $\mathfrak{X}(U)$. Clearly, $\mathcal{L}_H$ is finite in the subspace of all X with finite moment $EH(d(X,a))$ for some $a \in U$. Kruglov's distance $\mathbf{Kr}(X,Y) := \mathbf{Kr}(F_X, F_Y)$ is a p. semidistance in $\mathfrak{X}(\mathbb{R})$.

Examples of p. metrics in $\mathfrak{X}(U)$ are the Ky Fan metric

$$\mathbf{K}(X,Y) := \inf\{\varepsilon > 0 : \Pr(d(X,Y) > \varepsilon) < \varepsilon\}, \quad X, Y \in \mathfrak{X}(U), \tag{2.5.2}$$

and the $\mathcal{L}_p$*-metrics* $(0 \leq p \leq \infty)$

$$\mathcal{L}_p(X,Y) := \{Ed^p(X,Y)\}^{\min(1,1/p)}, \quad 0 < p < \infty, \tag{2.5.3}$$

$$\mathcal{L}_\infty(X,Y) := \operatorname{ess\,sup} d(X,Y) := \inf\{\varepsilon > 0 : \Pr(d(X,Y) > \varepsilon) = 0\}, \tag{2.5.4}$$

$$\mathcal{L}_0(X,Y) := EI\{X,Y\} := \Pr(X,Y). \tag{2.5.5}$$

The engineer's metric **EN**, Kolmogorov metric $\boldsymbol{\rho}$, Kantorovitch metric $\boldsymbol{\kappa}$, and the Lévy metric **L** (Sect. 2.2) are p. semimetrics in $\mathfrak{X}(\mathbb{R})$.

Remark 2.5.1. Unlike Definition 2.5.2, Definition 2.5.1 is free of the choice of the initial probability space and depends solely on the structure of the metric space U. The main reason for considering not arbitrary but separable metric spaces (U,d) is that we need the measurability of the metric d to connect the metric structure of U with that of $\mathfrak{X}(U)$. In particular, the measurability of d enables us to handle, in a well-defined way, probability metrics such as the Ky Fan metric **K** and $\mathcal{L}_p$-metrics.

[11]If we replace "semidistance" with "distance," then the statement continues to hold.

Note that $\mathcal{L}_0$ does not depend on the metric d, so one can define $\mathcal{L}_0$ on $\mathfrak{X}(U)$, where U is an arbitrary measurable space, while in (2.5.2)–(2.5.4) we need $d(X, Y)$ to be an RV. Thus the natural class of spaces appropriate for our investigation is the class of s.m.s.

2.6 Universally Measurable Separable Metric Spaces

What follows is an exposition of some basic results regarding universally measurable separable metric spaces (u.m.s.m.s.). As we will see, the notion of u.m.s.m.s. plays an important role in TPM.

Definition 2.6.1. Let P be a Borel probability measure on a metric space (U, d). We say that P is *tight* if for each $\varepsilon > 0$ there is a compact $K \subseteq U$ with $P(K) \geq 1 - \varepsilon$.[12]

Definition 2.6.2. An s.m.s. (U, d) is *universally measurable* (u.m.) if every Borel probability measure on U is tight.

Definition 2.6.3. An s.m.s. (U, d) is *Polish* if it is topologically complete [i.e., there is a topologically equivalent metric e such that (U, e) is complete]. Here the topological equivalence of d and e simply means that for any $x, x_1, x_2, \ldots$ in U

$$d(x_n, x) \to 0 \iff e(x_n, x) \to 0.$$

Theorem 2.6.1. *Every Borel subset of a Polish space is u.m.*

Proof. See Billingsley (1968, Theorem 1.4), Cohn (1980, Proposition 8.1.10), and Dudley (2002, p. 391). □

Remark 2.6.1. Theorem 2.6.1 provides us with many examples of u.m. spaces but does not exhaust this class. The topological characterization of u.m.s.m.s. is a well-known open problem.[13]

In his famous paper on measure theory, Lebesgue (1905) claimed that the projection of any Borel subset of $\mathbb{R}^2$ onto $\mathbb{R}$ is a Borel set. As noted by Souslin and his teacher Lusin (1930), this is in fact not true. As a result of the investigations surrounding this discovery, a theory of such projections (the so-called analytic or Souslin sets) was developed. Although not a Borel set, such a projection was shown to be Lebesgue-measurable; in fact it is u.m. This train of thought leads to the following definition.

---

[12] See (Dudley, 2002, Sect. 11.5).

[13] See Billingsley (1968, Appendix III, p. 234)

Definition 2.6.4. Let S be a Polish space, and suppose that f is a measurable function mapping S onto an s.m.s. U. In this case, we say that U is *analytic*.

Theorem 2.6.2. *Every analytic s.m.s. is u.m.*

Proof. See Cohn (1980, Theorem 8.6.13, p. 294) and Dudley (2002, Theorem 13.2.6). □

Example 2.6.1. Let $\mathbb{Q}$ be the set of rational numbers with the usual topology. Since $\mathbb{Q}$ is a Borel subset of the Polish space R, then $\mathbb{Q}$ is u.m.; however, $\mathbb{Q}$ is not itself a Polish space.

Example 2.6.2. In any uncountable Polish space, there are analytic (hence u.m.) non-Borel sets.[14]

Example 2.6.3. Let $C[0,1]$ be the space of continuous functions $f : [0,1] \to \mathbb{R}$ under the uniform norm. Let $E \subseteq C[0,1]$ be the set of f that fail to be differentiable at some $t \in [0,1]$. Then a theorem of Mazukiewicz (1936) says that E is an analytic, non-Borel subset of $C[0,1]$. In particular, E is u.m.

Recall again the notion of *Hausdorff metric* $r := r_\rho$ in the space of all subsets of a given metric space (S,ρ)

$$\begin{aligned} r(A,B) &= \max\left\{\sup_{x\in A}\inf_{y\in B}\rho(x,y), \sup_{y\in B}\inf_{x\in A}\rho(x,y)\right\} \\ &= \inf\{\varepsilon > 0 : A^\varepsilon \supseteq B, B^\varepsilon \supseteq A\}, \end{aligned} \tag{2.6.1}$$

where A^ε is the open ε-neighborhood of A, $A^\varepsilon = \{x : d(x.A) < \varepsilon\}$.

As we noticed in the space 2^S of all subsets $A \neq \emptyset$ of S, the Hausdorff distance r is actually only a semidistance. However, in the space $\mathcal{C} = \mathcal{C}(S)$ of all closed nonempty subsets, r is a metric (Definition 2.4.1) and takes on both finite and infinite values, and if S is a bounded set, then r is a finite metric on $\mathcal{C}$.

Theorem 2.6.3. *Let (S,ρ) be a metric space, and let $(\mathcal{C}(S), r)$ be the space described previously. If (S,ρ) is separable (or complete, or totally bounded), then $(\mathcal{C}(S), r)$ is separable (or complete, or totally bounded).*

Proof. See Hausdorff (1949, Sect. 29) and Kuratowski (1969, Sects. 21 and 23). □

Example 2.6.4. Let $S = [0,1]$, and let ρ be the usual metric on S. Let $\mathcal{R}$ be the set of all finite complex-valued Borel measures m on S such that the Fourier transform

$$\widehat{m}(t) = \int_0^1 \exp(iut)m(\mathrm{d}u)$$

[14]See Cohn (1980, Corollary 8.2.17) and Dudley (2002, Proposition 13.2.5).

vanishes at $t = \pm\infty$. Let $\mathcal{M}$ be the class of sets $E \in \mathcal{C}(S)$ such that there is some $m \in \mathcal{R}$ concentrated on E. Then $\mathcal{M}$ is an analytic, non-Borel subset of $(\mathcal{C}(S), r_\rho)$.[15]

We seek a characterization of u.m.s.m.s. in terms of their Borel structure.

Definition 2.6.5. A measurable space M with σ-algebra $\mathcal{M}$ is *standard* if there is a topology τ on M such that (M, τ) is a compact metric space and the Borel σ-algebra generated by τ coincides with $\mathcal{M}$.

An s.m.s. is standard if it is a Borel subset of its completion.[16] Obviously, every Borel subset of a Polish space is standard.

Definition 2.6.6. Say that two s.m.s. U and V are called Borel-isomorphic if there is a one-to-one correspondence f of U onto V such that $B \in \mathcal{B}(U)$ if and only if $f(B) \in \mathcal{B}(V)$.

Theorem 2.6.4. *Two standard s.m.s. are Borel-isomorphic if and only if they have the same cardinality.*

Proof. See Cohn (1980, Theorem 8.3.6) and Dudley (2002, Theorem 13.1.1). □

Theorem 2.6.5. *Let U be an s.m.s. The following are equivalent:*

(1) U is u.m.
(2) For each Borel probability m on U there is a standard set $S \in \mathcal{B}(U)$ such that $m(S) = 1$.

Proof. $1 \Rightarrow 2$: Let m be a law on U. Choose compact $K_n \subseteq U$ with $m(K_n) \geq 1 - 1/n$. Put $S = \cup_{n\geq 1} K_n$. Then S is σ-compact and, hence, standard. Thus, $m(S) = 1$, as desired.

$2 \Leftarrow 1$: Let m be a law on U. Choose a standard set $S \in \mathcal{B}(U)$ with $m(S) = 1$. Let $\overline{U}$ be the completion of U. Then S is Borel in its completion $\overline{S}$, which is closed in $\overline{U}$. Thus, S is Borel in $\overline{U}$. It follows from Theorem 2.6.1 that

$$1 = m(S) = \sup\{m(K) : K \text{ compact}\}.$$

Thus, every law m on U is tight, so that U is u.m. □

Corollary 2.6.1. *Let (U, d) and (V, e) be Borel-isomorphic separable metric spaces. If (U, d) is u.m., then so is (V, e).*

Proof. Suppose that m is a law on V. Define a law n on U by $n(A) = m(f(A))$, where $f : U \to V$ is a Borel isomorphism. Since U is u.m., there is a standard set $\subseteq U$ with $n(S) = 1$. Then $f(S)$ is a standard subset of V with $m(f(S)) = 1$. Thus, by Theorem 2.6.5, V is u.m. □

[15] See Kaufman (1984).
[16] See Dudley (2002, p. 347).

The following result, which is in essence due to Blackwell (1956), will be used in an important way later on.[17]

Theorem 2.6.6. *Let U be a u.m. separable metric space, and suppose that* Pr *is a probability measure on U. If $\mathcal{A}$ is a countably generated sub-σ-algebra of $\mathcal{B}(U)$, then there is a real-valued function $P(B|x)$, $B \in \mathcal{B}(U)$, $x \in U$, such that*

(1) For each fixed $B \in \mathcal{B}(U)$ the mapping $x \to P(B|x)$ is an $\mathcal{A}$-measurable function on U;
(2) For each fixed $x \in U$ the set function $B \to P(B|x)$ is a law on U;
(3) For each $A \in \mathcal{A}$ and $B \in \mathcal{B}(U)$ we have $\int_A P(B|x)\,\mathrm{Pr}(\mathrm{d}x) = \mathrm{Pr}(A \cap B)$;
(4) There is a set $N \in \mathcal{A}$ with $\mathrm{Pr}(N) = 0$ such that $P(B|x) = 1$ whenever $x \in U - N$.

Proof. Choose a sequence $F_1, F_2, \ldots$ of sets in $\mathcal{B}(U)$ that generates $\mathcal{B}(U)$ and is such that a subsequence generates $\mathcal{A}$. We will prove that there exists a metric e on U such that (U, d) and (U, e) are Borel-isomorphic and for which the sets $F_1, F_2, \ldots$ are *clopen*, i.e., open and closed. □

Claim. If (U, d) is an s.m.s. and $A_1, A_2, \ldots$ is a sequence of Borel subsets of U, then there is some metric e on U such that

(i) (U, e) is an s.m.s. isometric with a closed subset of $\mathbb{R}$;
(ii) $A_1, A_2, \ldots$ are clopen subsets of (U, e);
(iii) (U, d) and (U, e) are Borel-isomorphic (Definition 2.6.6).

Proof of claim. Let $B_1, B_2, \ldots$ be a countable base for the topology of (U, d). Define sets $C_1, C_2, \ldots$ by $C_{2n-1} = A_n$ and $C_{2n} = B_n$ $(n = 1, 2, \ldots)$ and $f : U \to \mathbb{R}$ by $f(x) = \sum_{n=1}^{\infty} 2I_{C_n}(x)/3^n$. Then f is a Borel isomorphism of (U, d) onto $f(U) \subseteq K$, where K is the Cantor set

$$K := \left\{ \sum_{n=1}^{\infty} \alpha_n/3^n : \alpha_n \text{ take value 0 or 2} \right\}.$$

Define the metric e by $e(x, y) = |f(x) - f(y)|$, so that (U, e) is isometric with $f(U) \subseteq K$. Then $A_n = f^{-1}\{x \in K; x(n) = 2\}$, where $x(n)$ is the nth digit in the ternary expansion of $x \in K$. Thus, A_n is clopen in (U, e), as required.

Now (U, e) is (Corollary 2.6.1) u.m., so there are compact sets $K_1 \subseteq K_2 \subseteq \cdots$ with $\mathrm{Pr}(K_n) \to 1$. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be the (countable) algebras generated by the sequences $F_1, F_2, \ldots$ and $F_1, F_2, \ldots, K_1, K_2, \ldots$, respectively. Then define $P_1(B|x)$ so that (1) and (3) are satisfied for $B \in \mathcal{G}_2$. Since $\mathcal{G}_2$ is countable, there is some set $N \in \mathcal{A}$ with $\mathrm{Pr}(N) = 0$ and such that for $x \in N$,

(a) $P_1(\cdot|x)$ is a finitely additive probability on $\mathcal{G}_2$,
(b) $P_1(A|x) = 1$ for $A \in \mathcal{A} \cap \mathcal{G}_2$ and $x \in A$,
(c) $P_1(K_n|x) \to 1$ as $n \to \infty$.

[17] See Theorem 3.3.1 in Sect. 3.3.

Claim. For $x \in N$ the set function $B \to P_1(B|x)$ is countably additive on $\mathcal{G}_1$.

Proof of claim. Suppose that $H_1, H_2, \ldots$ are disjoint sets in $\mathcal{G}_1$ whose union is U. Since the H_n are clopen and the K_n are compact in (U, e), there is, for each n, some $M = M(n)$ such that $K_n \subseteq H_1 \cup H_2 \cup \cdots \cup H_M$. Finite additivity of $P_1(x, \cdot)$ on $\mathcal{G}_2$ yields, for $x \notin N$, $P_1(K_n|x) \leq \sum_{i=1}^{M} P_1(H_i|x) \leq \sum_{i=1}^{\infty} P_1(H_i|x)$. Let $n \to \infty$, and apply (c) to obtain $\sum_{i=1}^{\infty}(P_1(H_i|x) = 1$, as required.

In view of the claim, for each $x \in N$ we define $B \to P(B|x)$ as the unique countably additive extension of P_1 from $\mathcal{G}_1$ to $\mathcal{B}(U)$. For $x \in N$ put $P(B|x) = \Pr(B)$. Clearly, (2) holds. Now the class of sets in $\mathcal{B}(U)$ for which (1) and (3) hold is a monotone class containing $\mathcal{G}_1$, and so coincides with $\mathcal{B}(U)$.

Claim. Condition (4) holds.

Proof of claim. Suppose that $A \in \mathcal{A}$ and $x \in A - N$. Let A_0 be the $\mathcal{A}$-atom containing x. Then $A_0 \subseteq A$, and there is a sequence $A_1, A_2, \ldots$ in $\mathcal{G}_1$ such that $A_0 = A_1 \cap A_2 \cap \cdots$. From (b), $P(A_n|x) = 1$ for $n \geq 1$, so that $P(A_0|x) = 1$, as desired. □

Corollary 2.6.2. *Let U and V be u.m.s.m.s., and let* Pr *be a law on $U \times V$. Then there is a function $P : \mathcal{B}(V) \times U \to \mathbb{R}$ such that*

(1) For each fixed $B \in \mathcal{B}(V)$ the mapping $x \to P(B|x)$ is measurable on U;
(2) For each fixed $x \in U$ the set function $B \to P(B|x)$ is a law on V;
(3) For each $A \in \mathcal{B}(U)$ and $B \in \mathcal{B}(V)$ we have

$$\int_{\mathcal{A}} P(B|x) P_1(\mathrm{d}x) = \Pr(A \cap B),$$

where P_1 is the marginal of Pr *on U.*

Proof. Apply the preceding theorem with $\mathcal{A}$ the σ-algebra of rectangles $A \times U$ for $A \in \mathcal{B}(U)$. □

2.7 Equivalence of the Notions of Probability (Semi-) distance on $\mathcal{P}_2$ and on $\mathfrak{X}$

As we saw in Sect. 2.5, every p. (semi-)distance on $\mathcal{P}_2$ induces (by restriction) a p. (semi-)distance on $\mathfrak{X}$. It remains to be seen whether every p. (semi-)distance on $\mathfrak{X}$ arises in this way. This will certainly be the case whenever

$$\mathcal{L}\mathfrak{X}_2(U, (\Omega, \mathcal{A}, \Pr)) = \mathcal{P}_2(U). \tag{2.7.1}$$

Note that the left member depends not only on the structure of (U, d) but also on the underlying probability space.

In this section we will prove the following facts.

1. There is some probability space $(\Omega, \mathcal{A}, \Pr)$ such that (2.7.1) holds for every separable metric space U.
2. If U is a separable metric space, then (2.7.1) holds for every nonatomic probability space $(\Omega, \mathcal{A}, \Pr)$ if and only if U is u.m.

We need a few preliminaries.

Definition 2.7.1. [18] If $(\Omega, \mathcal{A}, \Pr)$ is a probability space, then we say that $A \in \mathcal{A}$ is an *atom* if $\Pr(A) > 0$ and $\Pr(B) = 0$ or $\Pr(A)$ for each measurable $B \subseteq A$. A probability space is *nonatomic* if it has no atoms.

Lemma 2.7.1. [19] *Let ν be a law on a complete s.m.s. (U, d) and suppose that $(\Omega, \mathcal{A}, \Pr)$ is a nonatomic probability space. Then there is a U-valued RV X with distribution $\mathcal{L}(X) = \nu$.*

Proof. Denote by d^* the following metric on U^2: $d^*(x, y) := d(x_1, x_2) + d(y_1, y_2)$ for $x = (x_1, y_1)$ and $y = (x_2, y_2)$. For each k there is a partition of U^2 comprising nonempty Borel sets $\{A_{ik} : i = 1, 2, \dots\}$ with $\operatorname{diam}(A_{ik}) < 1/k$ and such that A_{ik} is a subset of some $A_{j,k-1}$.

Since $(\Omega, \mathcal{A}, \Pr)$ is nonatomic, we see that for each $\mathcal{C} \in \mathcal{A}$ and for each sequence p_i of nonnegative numbers such that $p_1 + p_2 + \cdots = \Pr(\mathcal{C})$ there exists a partitioning $\mathcal{C}_1, \mathcal{C}_2, \dots$ of $\mathcal{C}$ such that $\Pr(\mathcal{C}_i) = p_i$, $i = 1, 2, \dots$ [20]

Therefore, there exist partitions $\{B_{ik} : i = 1, 2, \dots\} \subseteq \mathcal{A}$, $k = 1, 2, \dots$, such that $B_{ik} \subseteq B_{jk-1}$ for some $j = j(i)$ and $\Pr(B_{ik}) = \nu(A_{ik})$ for all i, k. For each pair (i, j) let us pick a point $x_{ik} \in A_{ik}$ and define U^2-valued $X_k(\omega) = x_{ik}$ for $\omega \in B_{ik}$. Then $d^*(X_{k+m}(\omega), X_k(\omega)) < 1/k$, $m = 1, 2, \dots$, and since (U^2, d^*) is a complete space, there exists the limit $X(\omega) = \lim_{k\to\infty} X_k(\omega)$. Thus

$$d^*(X(\omega), X_k(\omega)) \le \lim_{m\to\infty} [d^*(X_{k+m}(\omega), X(\omega)) + d^*(X_{k+m}(\omega), X_k(\omega))] \le \frac{1}{k}.$$

Let $P_k := \Pr_{X_k}$ and $P^* := \Pr_X$. Further, our aim is to show that $P^* = \nu$. For each closed subset $A \subseteq U$

$$P_k(A) = \Pr(X_k \in A) \le \Pr(X \in A^{1/k}) = P^*(A^{1/k}) \le P_k(A^{2/k}), \qquad (2.7.2)$$

where $A^{1/k}$ is the open $1/k$-neighborhood of A. On the other hand,

$$P_k(A) = \sum\{P_k(x_{ik}) : x_{ik} \in A\} = \sum\{\Pr(B_{ik}) : x_{ik} \in A\}$$

[18] See Loeve (1963, p. 99) and Dudley (2002, p. 82).

[19] See Berkes and Phillip (1979).

[20] See, for example, Loeve (1963, p. 99).

$$= \sum\{\nu(A_{ik}) : x_{ik} \in A\} \le \sum\{\nu(A_{ik} \cap A^{1/k}) : x_{ik} \in A\}$$
$$\le \nu(A^{1/k}) \le \sum\{\nu(A_{ik}) : x_{ik} \in A^{2/k}\} \le P_k(A^{2/k}). \qquad (2.7.3)$$

Further, we can estimate the value $P_k(A^{2/k})$ in the same way as in (2.7.2) and (2.7.3), and thus we get the inequalities

$$P^*(A^{1/k}) \le P_k(A^{2/k}) \le P^*(A^{2/k}), \qquad (2.7.4)$$
$$\nu(A^{1/k}) \le P_k(A^{2/k}) \le \nu(A^{3/k}). \qquad (2.7.5)$$

Since $\nu(A^{1/k})$ tends to $\nu(A)$ with $k \to \infty$ for each closed set A and, analogously, $P^*(A^{1/k}) \to P^*(A)$ as $k \to \infty$, then by (2.7.4) and (2.7.5) we obtain the equalities

$$P^*(A) = \lim_{k\to\infty} P_k(A^{2/k}) = \nu(A)$$

for each closed A, and hence $P^* = \nu$. □

Theorem 2.7.1. *There is a probability space* $(\Omega, \mathcal{A}, \Pr)$ *such that for every s.m.s.* U *and every Borel probability* μ *on* U *there is an RV* $X : \Omega \to U$ *with* $\mathcal{L}(X) = \mu$.

Proof. Define $(\Omega, \mathcal{A}, \Pr)$ as the measure-theoretic (von Neumann) product[21] of the probability spaces $(C, \mathcal{B}(C), \nu)$, where C is some nonempty subset of $\mathbb{R}$ with Borel σ-algebra $\mathcal{B}(C)$ and ν is some Borel probability on $(C, \mathcal{B}(C))$.

Now, given an s.m.s. U, there is some set $C \subseteq \mathbb{R}$ Borel-isomorphic with U (Claim 2.6 in Theorem 2.6.6). Let $f : C \to U$ supply the isomorphism. If μ is a Borel probability on U, then let ν be a probability on C such that $f(\nu) := \nu f^{-1} = \mu$. Define $X : \Omega \to U$ as $X = f \circ \pi$, where $\pi : \Omega \to C$ is a projection onto the factor $(C, \mathcal{B}(C), \nu)$. Then $\mathcal{L}(X) = \mu$, as desired. □

Remark 2.7.1. The preceding result establishes claim (i) made at the beginning of the section. It provides one way of ensuring (2.7.1): simply insist that all RVs be defined on a "superprobability space" as in Theorem 2.7.1. We make this assumption throughout the sequel.

The next theorem extends the Berkes and Phillips's lemma 2.7.1 to the case of u.m.s.m.s. U.

Theorem 2.7.2. *Let* U *be an s.m.s. The following statements are equivalent.*

(1) U *is u.m.*
(2) If $(\Omega, \mathcal{A}, \Pr)$ *is a nonatomic probability space, then for every Borel probability* P *on* U *there is an RV* $X : \Omega \to U$ *with law* $\mathcal{L}(X) = P$.

[21] See Hewitt and Stromberg (1965, Theorems 22.7 and 22.8, p. 432–133).

Proof. 1 ⇒ 2: Since U is u.m., there is some standard set $S \in \mathcal{B}(U)$ with $P(S) = 1$ (Theorem 2.6.5). Now there is a Borel isomorphism f mapping S onto a Borel subset B of $\mathbb{R}$ (Theorem 2.6.4). Then $f(P) := P \circ f^{-1}$ is a Borel probability on $\mathbb{R}$. Thus, there is an RV $g : \Omega \to \mathbb{R}$ with $\mathcal{L}(g) = f(P)$ and $g(\Omega) \subseteq B$ (Lemma 2.7.1 with $(U, d) = (\mathbb{R}, |\cdot|)$). We may assume that $g(\Omega) \subseteq B$ since $\Pr(g^{-1}(B)) = 1$. Define $x : \Omega \to U$ by $x(\omega) = f^{-1}(g(\omega))$. Then $\mathcal{L}(X) = \nu$, as claimed.

2 ⇒ 1: Now suppose that ν is a Borel probability on U. Consider an RV $X : \Omega \to U$ on the (nonatomic) probability space $((0,1), \mathcal{B}(0,1), \lambda)$ with $\mathcal{L}(X) = \nu$. Then range(X) is an analytic subset of U with $\nu^*(\text{range}(X)) = 1$. Since range$(X)$ is u.m. (Theorem 2.6.2), there is some standard set $S \subseteq \text{range}(X)$ with $P(S) = 1$. This follows from Theorem 2.6.5. The same theorem shows that U is u.m. □

Remark 2.7.2. If U is a u.m.s.m.s., we operate under the assumption that all U-valued RVs are defined on a nonatomic probability space. Then (2.7.1) will be valid.

References

Berkes I, Phillip W (1979) Approximation theorems for independent and weakly independent random vectors. Ann Prob 7:29–54

Billingsley P 1968 Convergence of probability measures. Wiley, New York

Birnbaum ZW, Orliz W (1931) Über die verallgemeinerung des begriffes der zueinander Konjugierten Potenzen. Stud Math 3:1–67

Blackwell D (1956) On a class of probability spaces. In: Proceedings of the 3rd Berkeley symposium on mathematical statistics and probability, vol 2, pp 1–6

Cohn DL (1980) Measure theory. Birkhauser, Boston

Dudley RM (1976) Probabilities and metrics: convergence of laws on metric spaces, with a view to statistical testing. In: Aarhus University Mathematics Institute lecture notes series no. 45, Aarhus

Dudley RM (2002) Real analysis and probability, 2nd edn. Cambridge University Press, New York

Dunford N, Schwartz J (1988) Linear operators, vol 1. Wiley, New York

Gibbs A, Su F (2002) On choosing and bounding probability metrics. Int Stat Rev 70:419–435

Hausdorff F (1949) Set theory. Dover, New York

Hennequin PL, Tortrat A (1965) Théorie des probabilités et quelques applications. Masson, Paris

Hewitt E, Stromberg K (1965) Real and abstract analysis. Springer, New York

Kaufman R (1984) Fourier transforms and descriptive set theory. Mathematika 31:336–339

Kruglov VM (1973) Convergence of numerical characteristics of independent random variables with values in a Hilbert space. Theor Prob Appl 18:694–712

Kuratowski K (1969) Topology, vol II. Academic, New York

Lebesgue H (1905) Sur les fonctions representables analytiquement. J Math Pures Appl V:139–216

Loeve M (1963) Probability theory, 3rd edn. Van Nostrand, Princeton

Lukacs E (1968) Stochastic convergence. D.C. Heath, Lexington, MA

Lusin N (1930) Lecons Sur les ensembles analytiaues. Gauthier-Villars, Paris

Massey FJ Jr (1950) A note on the power of a non-parametric test. Annal Math Statist 21:440–443

Mazukiewicz S (1936) Uberdie Menge der differenzierbaren Funktionen. Fund Math 27:247–248

Mostafaei H, S Kordnourie (2011) Probability metrics and their applications. Appl Math Sci 5(4):181–192

Thompson R (1966) Bias of the one-sample Cramér-Von Mises test. J Am Stat Assoc 61:246–247

Thompson R (1979) Bias and monotonicity of goodness-of-fit tests. J Am Stat Assoc 74:875–876

http://www.springer.com/978-1-4614-4868-6

The Methods of Distances in the Theory of Probability and Statistics
Rachev, S.T.; Klebanov, L.; Stoyanov, S.V.; Fabozzi, F.
2013, XVI, 619 p., Hardcover
ISBN: 978-1-4614-4868-6