PROBABILITY DISTANCES

SVANTE JANSON

Abstract. This is a survey of some important probability metrics, for probability distributions on a complete metric space. There are no new results.

1. Introduction

A probability metric or probability distance is a metric on a suitable set of probability distributions in some measurable space $S$. In this survey we give definitions and basic properties of (some of) the most important ones. (There are no new results.) See e.g. Zolotarev [19] or the books Rachev [14] and Rachev et al [15] (which is a second, enlarged edition of [14]) for a general theory and many other examples. A few proofs are given, and references are given to many other results, but many (usually simple) results are stated without proof or reference. Similarly, we give a few original references, but usually we ignore the history of the metrics studied here, and refer to e.g. the books just cited.

Although a probability metric $d(\mu, \nu)$ is formally defined for distributions $\mu$ and $\nu$, we follow common practice and write $d(X,Y) := d(\mathcal{L}(X), \mathcal{L}(Y))$ when $X$ and $Y$ are random variables with distributions $\mathcal{L}(X)$ and $\mathcal{L}(Y)$, and we often state the definitions below in this form. We switch between the versions for distributions and random variables without further comments, but we stress that $d(X,Y)$ thus depends only on the distributions of $X$ and $Y$. In particular,

$$d(X,Y) = 0 \iff X \overset{d}{=} Y.$$  \hfill (1.1)

Remark 1.1. We do not follow the elaborate terminology of e.g. [15]; our probability metrics are the simple probability metrics in [15]. Also, we do not distinguish between the terms “probability metric” and “probability distance”.

2. Notation and other preliminaries

Except in Section 5, $S$ is a complete separable metric space, equipped with its Borel $\sigma$-field $\mathcal{B}(S)$. The metric on $S$ is denoted by $d(x,y)$; we may use the more precise notation $(S,d)$ for the metric space when the metric is not obvious from the context. (There should not be any danger of confusing the metric $d$ on $S$ with probability metrics.)

$X, X_n, Y$ will generally denote random variables in $S$. 

Date: 17 January, 2020.
Partly supported by the Knut and Alice Wallenberg Foundation.
Let $o$ denote an arbitrary but fixed point in $\mathcal{S}$; the choice of $o$ does not matter. When $S = \mathbb{R}^q$, or more generally a Banach space, we take $o = 0$, and then $d(x, o) = \|x\|$ for $x \in \mathcal{S}$.

**Remark 2.1.** Some definitions and results extend to more general metric spaces, but there are also several technical problems, sometimes serious; e.g., with measurability if the space is not separable, and with existence of couplings if the space is not complete. See e.g. [2, Appendix III], [3, Section 8.3] and [15] for some results and limitations.

The assumption that $(\mathcal{S}, d)$ is complete can be relaxed to assuming that there exists an equivalent complete metric.

$\mathcal{M}(\mathcal{S})$ denotes the space of all signed Borel measures on $\mathcal{S}$, and $\mathcal{P}(\mathcal{S})$ denotes the subset of all probability measures. $\mathcal{M}(\mathcal{S})$ is a Banach space with the total variation norm $\|\cdot\|_{\mathcal{M}(\mathcal{S})}$. If $X$ is a random variable in $\mathcal{S}$, then $\mathcal{L}(X) \in \mathcal{P}(\mathcal{S})$ denotes its distribution.

$\delta_s$ denotes the Dirac measure at $s \in \mathcal{S}$, i.e., the distribution of the deterministic “random variable” $X := s$.

The weak topology in $\mathcal{P}(\mathcal{S})$ is defined in the standard way as the weak topology with respect to the space of bounded continuous functions on $\mathcal{S}$, see e.g. [2] or [3]. Recall that convergence in distribution $X_n \xrightarrow{d} X_n$ of random variables $X, X_n$ in $\mathcal{S}$, is defined as weak convergence (i.e., convergence in the weak topology) of their distributions $\mathcal{L}(X_n)$ to $\mathcal{L}(X)$.

Increasing and decreasing are used in weak sense: a function $f$ is increasing if $x \leq y \implies f(x) \leq f(y)$.

For $x, y \in \mathbb{R}^q$, we let $x \leq y$ denote the coordinate-wise partial order, i.e., $x_i \leq y_i$ for $i = 1, \ldots, q$, where $x = (x_i)_1^q$ and $y = (y_i)_1^q$.

$F_X(x) := \mathbb{P}(X \leq x)$ denotes the distribution function of a random variable $X$ with values in $\mathbb{R}$ or $\mathbb{R}^q$.

If $F : \mathbb{R} \to [0, 1]$ is a distribution function, then let

$$
F^{-1}(t) := \sup\{x : F(x) \leq t\} = \inf\{x : F(x) > t\}, \quad t \in (0,1). \tag{2.1}
$$

$F^{-1}$ is increasing and right-continuous. Furthermore, if $U$ is a uniformly distributed random variable on $(0,1)$, then $F^{-1}(U)$ is a random variable with the distribution function $F$. Fix a random variable $U \sim U(0,1)$, and for any real-valued random variable $X$, define $\bar{X} := F_X^{-1}(U)$. Then, thus,

$$
X \overset{d}{=} \bar{X} = F_X^{-1}(U). \tag{2.2}
$$

($U$ is fixed in the sequel.)

**Remark 2.2.** Equivalently, we may define $\bar{X}$ as the function $F_X^{-1}$ regarded as a random variable defined on the probability space $(0,1)$ (with Lebesgue measure); this is the same as the definition above for a specific choice of $U$.

We do not assume this.

A *coupling* of two random variables $X$ and $Y$ is a pair $(X', Y')$ of random variables on a common probability space such that $X' \overset{d}{=} X$ and $Y' \overset{d}{=} Y$.

If $X$ and $Y$ are real-valued random variables, then

$$
(\bar{X}, \bar{Y}) = (F_X^{-1}(U), F_Y^{-1}(U)) \tag{2.3}
$$
is a coupling of $X$ and $Y$ by (2.2); we call this the *monotone coupling* of $X$ and $Y$.

If $X$ is a random variable with values in $\mathbb{R}$, or more generally in a normed space with norm $|\cdot|$, then the $L_p$ norm of $X$ is defined by

$$
\|X\|_p := \begin{cases} 
(\mathbb{E}|X|^p)^{1/p}, & 0 < p < \infty, \\
\text{ess sup}|X|, & p = \infty.
\end{cases}
$$

(2.4)

For convenience, we extend the usual definition of metric, and allow a metric to take the value $+\infty$. If $d'$ is a metric in this sense on a set $E$, and $a \in E$, then $d'$ is finite, and thus a proper metric, on the set $\{x \in E : d'(x, a) < \infty\}$.

$\lfloor x \rfloor$ denotes the integer part of a real number $x$, and $\lceil x \rceil := -\lfloor -x \rfloor$ the smallest integer $\geq x$.

$\wedge y := \min(x, y)$ and $\vee y := \max(x, y)$.

Unspecified limits are as $n \to \infty$.

### 2.1. Lipschitz norms

Let $0 < \alpha \leq 1$. For a real-valued function $f : S \to \mathbb{R}$, define

$$
||f||_{\text{Lip}_\alpha} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)^\alpha},
$$

(2.5)

and let

$$
\text{Lip}_\alpha(S) := \{ f : S \to \mathbb{R} : ||f||_{\text{Lip}_\alpha} < \infty \}
$$

(2.6)

$$
= \{ f : S \to \mathbb{R} : |f(x) - f(y)| \leq Cd(x, y)^\alpha \text{ for some } C \text{ and all } x, y \in S \}.
$$

(2.7)

Note that $||f||_{\text{Lip}_\alpha} = 0$ if (and only if) $f$ is constant, so (2.5) is a seminorm only, and that $\text{Lip}_\alpha/\mathbb{R}$, i.e., $\text{Lip}_\alpha$ modulo constant functions, is a Banach space with the norm (2.5).

Furthermore, let $B(S)$ the space of bounded functions on $S$, with

$$
||f||_{B(S)} := \sup_{x \in S} |f(x)|,
$$

(2.8)

and let

$$
\text{BLip}_\alpha(S) := \text{Lip}_\alpha(S) \cap B(S).
$$

(2.9)

$\text{BLip}_\alpha(S)$ is a Banach space with either of the two norms

$$
||f||_{\text{BLip}_\alpha(S)} := \max\{ ||f||_{\text{Lip}(S)}, ||f||_{B(S)} \},
$$

(2.10)

$$
||f||'_{\text{BLip}_\alpha(S)} := ||f||_{\text{Lip}(S)} + ||f||_{B(S)}.
$$

(2.11)

These two norms are obviously equivalent, with

$$
||f||_{\text{BLip}_\alpha(S)} \leq ||f||'_{\text{BLip}_\alpha(S)} \leq 2||f||_{\text{BLip}_\alpha(S)}.
$$

(2.12)

We use the norm (2.10) unless anything else is said.

We may omit the subscript $\alpha$ when $\alpha = 1$, i.e., $\text{Lip} := \text{Lip}_1$ and $\text{BLip} := \text{BLip}_1$. 


Remark 2.3. If the diameter \( \text{diam}(\mathcal{S}) := \sup_{x,y \in \mathcal{S}} d(x, y) \) of \( \mathcal{S} \) is finite, then every function in \( \text{Lip}_\alpha \) is bounded, so \( \text{BLip}_\alpha = \text{Lip}_\alpha \) as sets. They are not quite the same as normed spaces, since the constant function 1 has norm \( \alpha \leq 1 \), and every function in \( \text{Lip}_\alpha \) has norm \( 1 \) in \( \text{Lip}_\alpha \) but not in \( \text{BLip}_\alpha \). However, the quotient spaces \( \text{BLip}_\alpha / \mathbb{R} \) and \( \text{Lip}_\alpha / \mathbb{R} \) are Banach spaces with equivalent norms, and if the diameter of \( \mathcal{S} \) is \( \leq 2^{1/\alpha} \), then \( \text{BLip}_\alpha / \mathbb{R} \) and \( \text{Lip}_\alpha / \mathbb{R} \) have the same norm so they are equal as Banach spaces.

Note also that we can always introduce a new, bounded, metric in \( \mathcal{S} \) by \( d_1(x, y) := d(x, y) \wedge 1 \); this is equivalent to \( d \), so it generates the same topology, but \( \mathcal{S} \) has diameter at most 1 for the new metric. Then \( \text{BLip}_\alpha(\mathcal{S}, d_1) = \text{BLip}_\alpha(\mathcal{S}, d) \) with equivalent norms, and

\[
\text{Lip}_\alpha(\mathcal{S}, d_1) = \text{BLip}_\alpha(\mathcal{S}, d_1) = \text{BLip}_\alpha(\mathcal{S}, d) \tag{2.13}
\]
as sets. Hence,

\[
\text{Lip}_\alpha(\mathcal{S}, d_1) / \mathbb{R} = \text{BLip}_\alpha(\mathcal{S}, d_1) / \mathbb{R} = \text{BLip}_\alpha(\mathcal{S}, d) / \mathbb{R} \tag{2.14}
\]
with equivalent norms. (If we instead use \( d_2(x, y) := d(x, y) \wedge 2^{1/\alpha} \), then the three spaces in (2.14) become isometric.)

2.2. Higher Lipschitz spaces. If \( \mathcal{S} = \mathbb{R}^d \), or more generally, if \( \mathcal{S} \) is a Banach space \( B \), we define also \( \text{Lip}_\alpha(\mathcal{S}) \) for \( \alpha > 1 \), as follows.

If \( B \) and \( B_1 \) are Banach spaces and \( f : V \to B_1 \) is a function defined on an open subset \( V \subseteq B \), then \( f \) is said to be (Fréchet) differentiable at a point \( x \in V \) if there exists a bounded linear operator \( Df(x) : B \to B_1 \) such that \( \| f(x + y) - f(x) - Df(x)y \|_{B_1} = o(\| y \|_B) \) as \( \| y \|_B \to 0 \). Further, \( f \) is differentiable in \( V \) if it is differentiable for every \( x \in V \); then \( Df \) is a function \( V \to L(B, B_1) \) (the space of bounded linear mappings \( B \to B_1 \)), and we may talk about its derivative \( D^2 f = DDF \), and so on; see e.g. [5]. Note that the \( m \)th derivative \( D^m f \) (if it exists) is a function from \( V \) into the Banach space of multilinear mappings \( B^m \to B_1 \); this space is equipped with the usual norm \( \sup \| T(x_1, \ldots, x_m) \|_{B_1} : \| x_1 \|_{B_1}, \ldots, \| x_m \|_B \leq 1 \). Let \( C^m(B, B_1) \) denote the space of \( m \) times continuously differentiable functions \( f : B \to B_1 \).

Given a Banach space \( B \) and a real number \( \alpha > 0 \), write \( \alpha = m + \gamma \) with \( m := \lfloor \alpha \rfloor - 1 \in \mathbb{N}_{\geq 0} \) and \( \gamma := \alpha - m \in (0, 1] \), and define, for \( f \in C^m(B, \mathbb{R}) \),

\[
\| f \|_{\text{Lip}_\alpha} := \sup_{x \neq y} \frac{\| D^m f(x) - D^m f(y) \|}{d(x, y)^\gamma}, \tag{2.15}
\]
and

\[
\text{Lip}_\alpha(B) := \{ f \in C^m(B, \mathbb{R}) : \| f \|_{\text{Lip}_\alpha} < \infty \}. \tag{2.16}
\]
Note that \( \| f \|_{\text{Lip}_\alpha} = 0 \) if and only if \( D^m f \) is constant. This holds, e.g. by Taylor’s formula [5, Théorème 5.6.1], if and only if \( f(x) \) is a polynomial function of degree \( \leq m \) in the sense that, for some bounded multilinear mappings \( T_k : B^k \to \mathbb{R} \),

\[
f(x) = \sum_{k=0}^m T_k(x, \ldots, x), \quad x \in B. \tag{2.17}
\]
(Here $T_0$ is a constant, obviously with $T_0 = f(0)$. It follows that $\text{Lip}_\alpha(B)$ regarded as a space of functions modulo polynomials (2.17) is a Banach space.

Note that for $0 < \alpha \leq 1$, we have $m = 0$ and $\text{Lip}_\alpha(S)$ is the same space as defined in (2.6).

3. Two constructions

Many probability metrics can be defined by the two constructions in the following subsections.

3.1. Minimal metrics. Let $\delta(X, Y)$ be a metric on random variables with values in $S$, defined for pairs of random variables $(X, Y)$ defined on a common probability space. (Thus, $\delta$ depends on the joint distribution of $X$ and $Y$, and is not a probability metric as defined in this paper. We assume that $\delta(X, Y) = 0 \iff X = Y$ a.s.) We allow $\delta$ to take the value $\infty$. (Equivalently, $\delta(X, Y)$ may be defined only for some $X$ and $Y$.) Then the corresponding minimal metric $\hat{\delta}$ is defined by

$$
\hat{\delta}(X, Y) := \inf\left\{ \delta(X', Y') : X' \overset{d}{=} X, \ Y' \overset{d}{=} Y \right\},
$$

thus taking the infimum over all couplings of $X$ and $Y$. The infimum is attained in all cases considered below, and then we may replace inf by min in (3.1).

$\hat{\delta}$ is symmetric by definition and satisfies the triangle inequality, e.g. as a consequence of [4, Lemma 1.1.6]; furthermore, at least in cases when the infimum in (3.1) is attained, $\hat{\delta}(X, Y) = 0$ implies $X \overset{d}{=} Y$, so $\hat{\delta}$ is really a probability metric, cf. (1.1).

3.2. Dual metrics. Let $F$ be a set of measurable functions $S \to \mathbb{R}$. Define a functional $\|\mu\|_F^*$ on the set of signed Borel measures $\mu \in \mathcal{M}(S)$ such that $\int_S |f| \, d\mu < \infty$ for every $f \in F$ by

$$
\|\mu\|_F^* := \sup\left\{ \left| \int_S f \, d\mu \right| : f \in F \right\} \in [0, \infty].
$$

This defines a seminorm on the space

$$
\mathcal{M}_F := \left\{ \mu \in \mathcal{M}(S) : \|\mu\|_F^* < \infty \right\}.
$$

If this seminorm is a norm, i.e., if $\int_S f \, d\mu = 0$ for all $f \in F$ implies $\mu = 0$, then

$$
d_F(\mu, \nu) := \|\mu - \nu\|_F^*
$$

defines a probability metric on $\mathcal{P}_F := \mathcal{P} \cap \mathcal{M}_F$. In terms of random variables, the definition is

$$
d_F(X, Y) := \sup\left\{ \left| \mathbb{E} f(X) - \mathbb{E} f(Y) \right| : f \in F \right\}.
$$

The most common case is that $F$ is the unit ball $\{ f \in \mathcal{F} : \| f \| \leq 1 \}$ where $\mathcal{F}$ either is a normed space of functions, or a seminormed space of functions with $\| f \| = 0 \iff f$ is a constant function (so that $\mathcal{F}$ can be regarded as a normed space of functions modulo constants). Then $d_F(\mu, \nu)$ is the norm of $\mu - \nu$ as an element of the dual space $\mathcal{F}^*$.

We may call $d_F$ the dual metric defined by the set $F$, or by the space $\mathcal{F}$.
Remark 3.1. The functions in $\mathcal{F}$, or perhaps in $\mathfrak{F}$, are regarded as (and often called) test functions.

Remark 3.2. The metric $d_\mathcal{F}$ in (3.4) and (3.5) is defined and finite for all probability measures on $\mathcal{S}$ if and only if for some (and then any) $o \in \mathcal{S}$, the set of functions $\{ x \mapsto f(x) - f(o) : f \in \mathcal{F} \}$ is uniformly bounded. When $\mathcal{F}$ is given by a normed space $\mathfrak{F}$ as above, this is equivalent to

$$|f(x) - f(o)| \leq C\|f\|, \quad f \in \mathfrak{F}, \quad x \in \mathcal{S}$$

for some constant $C$ not depending on $x$ or $f$. If $\mathfrak{F}$ is a Banach space, this is further equivalent to the property that every function $f \in \mathfrak{F}$ is bounded.

4. Ideal metrics

Let $\mathcal{S}$ be a Banach space. A probability metric $d$ is said to be ideal of order $\gamma$, where $\gamma \geq 0$, if it satisfies the following two properties [18; 19]:

(I1) For any random variables $X, Y, Z$ in $\mathcal{S}$ such that $Z$ is independent of $X$ and $Y$,

$$d(X + Z, Y + Z) \leq d(X, Z).$$

(4.1)

(I2) For any random variables $X, Y$ in $\mathcal{S}$ and $c \in \mathbb{R}$,

$$d(cX, cY) = |c|^\gamma d(X, Y).$$

(4.2)

(If $\gamma = 0$ and $c = 0$, we interpret $0^0 = 0$ in (4.2).)

Note that (I1) in particular implies translation invariance: for any constant $a \in \mathcal{S}$,

$$d(X + a, Y + a) = d(X, Y).$$

(4.3)

Furthermore, (I1) implies (and is equivalent to) that if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are two finite collections of independent random variables, then

$$d\left( \sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right) \leq \sum_{i=1}^n d(X_i, Y_i).$$

(4.4)

For the two constructions in Section 3, we have the following simple results.

Lemma 4.1. Let $\mathcal{S}$ be any complete separable metric space, and suppose that $\delta(X, Y)$ is a complete metric on random variables in $\mathcal{S}$. Then the minimal metric $\hat{\delta}$ is complete.

Proof. Suppose that $\mu_n$ is a Cauchy sequence of probability measures in $\mathcal{P}(\mathcal{S})$ for the metric $\delta$. By selecting a subsequence, we may assume that $\delta(\mu_n, \mu_{n+1}) < 2^{-n}$ for every $n \geq 1$. (If we show that the subsequence converges, then the original Cauchy sequence converges to the same limit.)

Then, for every $n \geq 1$ there exists random variables $X_n, Y_n \in \mathcal{S}$ such that $X_n \sim \mu_n$, $Y_n \sim \mu_{n+1}$, and $\delta(X_n, Y_n) < 2^{-n}$.

By [4, Lemma 1.1.6] and induction, there exists for every $n \geq 2$ a probability measure $\nu_n$ on $\mathcal{S}^n$ such that if $(Z_1^{(n)}, \ldots, Z_n^{(n)}) \sim \nu_n$, then $(Z_{n-1}^{(n)}, Z_n^{(n)}) \overset{d}{=} (X_{n-1}, Y_{n-1})$ and, when $n > 2$, $(Z_{n-1}^{(n)}, \ldots, Z_n^{(n)}) \overset{d}{=} \nu_{n-1}$. Consequently, by Kolmogorov’s extension theorem, there exists a probability measure $\nu$
on $S^\infty$ such that if $(Z_1, Z_2, \ldots) \sim \nu$, then $(Z_1^{(n)}, \ldots, Z_n^{(n)}) \sim \nu_n$ for every $n \geq 2$, and thus $(Z_k, Z_{k+1}) \overset{d}{=} (X_k, Y_k)$ for every $k \geq 1$. Thus
\[
\delta(Z_k, Z_{k+1}) = \delta(X_k, Y_k) < 2^{-k},
\]
and thus the sequence $Z_k$ is a Cauchy sequence for $\delta$. Hence, by assumption, there exists a random variable $Z$ such that $\delta(Z_n, Z) \to 0$. Since $Z_n \overset{d}{=} X_n \sim \mu_n$, this implies,
\[
\hat{\delta}(\mu_n, \mathcal{L}(Z)) \leq \delta(Z_n, Z) \to 0,
\]
and thus the sequence $\mu_n$ converges.

**Lemma 4.2.** Let $S$ be a Banach space and let $\delta(X, Y)$ be a metric on random variables in $S$ such that for any random variables $X$ and $Y$,
\[
\delta(X, Y) = \delta(X - Y, 0),
\]
\[
\delta(tX, 0) = |t|^\gamma \delta(X, 0) \quad t \in \mathbb{R}.
\]
Then the minimal metric $\hat{\delta}$ is ideal of order $\gamma$.

**Lemma 4.3.** Let $S$ be a Banach space and let $\mathfrak{F}$ be a (semi)normed space of functions $S \to \mathbb{R}$ such that if $f \in \mathfrak{F}$, $a \in S$ and $t \in \mathbb{R}$, then $f(\cdot + a), f(t \cdot) \in \mathfrak{F}$ and
\[
\|f(\cdot + a)\|_\mathfrak{F} = \|f\|_\mathfrak{F} \quad (5.1)
\]
\[
\|f(t \cdot)\|_\mathfrak{F} = |t|^\gamma \|f\|_\mathfrak{F}. \quad (5.2)
\]
Then the corresponding dual metric is ideal of order $\gamma$.

## 5. Total variation metric

In this section, $S = (S, \mathcal{B})$ may be any measure space.

The total variation distance of two probability measures $\mu, \nu \in \mathcal{P}(S)$ is defined as
\[
d_{TV}(\mu, \nu) := \sup\{ \|\mu(A) - \nu(A)\| : A \in \mathcal{B}\}. \quad (5.1)
\]
It is easy to see that
\[
d_{TV}(\mu, \nu) := \frac{1}{2}\|\mu - \nu\|, \quad (5.2)
\]
where $\|\mu - \nu\| := |\mu - \nu|(S)$ is the usual norm in $\mathcal{M}(S)$.

For random variables $X$ and $Y$, (5.1) takes the form
\[
d_{TV}(X, Y) := \sup\{ \|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\| : A \in \mathcal{B}\}. \quad (5.3)
\]

**Remark 5.1.** We may omit the absolute values in (5.1) and (5.3) without changing the supremum. (Since also $A^c \in \mathcal{B}$.)

**Remark 5.2.** Some authors prefer to define $d_{TV}(\mu, \nu) := \|\mu - \nu\|$, thus multiplying our $d_{TV}$ by 2. With our choice, we have $0 \leq d_{TV} \leq 1$. 

**Remark 5.3.** If $\mathcal{B}(S)$ is the space of bounded measurable functions $S \to \mathbb{R}$, with the norm (2.8), then $\|\mu\|$ is the dual norm in $\mathcal{M}(S)$, and thus the dual probability metric defined in Section 3.2 is $2d_{TV}$. Hence, $d_{TV}$ is the probability metric dual to the space $(\mathcal{B}(S), 2\|\cdot\|_E)$. Consequently, $d_{TV}$ equals the dual probability metric $d_F$ for any of the sets of test functions.
\{ f \in B_2(S) : |f| \leq \frac{1}{2} \}, \{ f \in B_1(S) : 0 \leq f \leq 1 \}, \text{ or, see the definition (5.1),} \{ 1_A : A \in B \}.

Given any probability measures \( \mu \) and \( \nu \) on \( S \), there exists a \( \sigma \)-finite measure \( \lambda \) such that both \( \mu \) and \( \nu \) are absolutely continuous with respect to \( \lambda \). (For example, \( \lambda := \mu + \nu \).) In this case, if \( \mu \) and \( \nu \) have densities (Radon–Nikodym derivatives) \( \frac{d\mu}{d\lambda} \) and \( \frac{d\nu}{d\lambda} \), then

\[
d_{TV}(\mu, \nu) = \frac{1}{2} \int_S \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right| d\lambda = \frac{1}{2} \left\| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right\|_{L_1(\lambda)} \tag{5.4}
\]

which implies

\[
d_{TV}(\mu, \nu) = 1 - \int_S \min \left\{ \frac{d\mu}{d\lambda}, \frac{d\nu}{d\lambda} \right\} d\lambda. \tag{5.5}
\]

In particular, \( d_{TV}(\mu, \nu) = 1 \) if and only if \( \mu \) and \( \nu \) are mutually singular.

It is easily seen that \( d_{TV} \) is a complete metric on \( P(S) \).

Consider now the special case that, as in the rest of the paper, \( S \) is a separable metric space. (So that the diagonal is measurable in \( S \times S \)). Then it is easy to see, e.g. using (5.5), that

\[
d_{TV}(X, Y) = \min \left\{ P(X' \neq Y') : X' \overset{d}{=} X, Y' \overset{d}{=} Y \right\}, \tag{5.6}
\]

where min as always indicates that the infimum is attained. In other words, \( d_{TV} \) is the minimal probability metric corresponding to \( \delta(X, Y) := P(X \neq Y) \). A coupling \( (X', Y') \) of \( X \) and \( Y \) attaining the minimum in (5.6) is called a maximal coupling. We repeat that such a coupling always exists.

**Remark 5.4.** The total variation metric is a rather strong metric. If \( S \) is countable and discrete, e.g. \( \mathbb{N} \) or \( \mathbb{Z} \), then weak convergence in \( P(S) \) is equivalent to convergence in \( d_{TV} \), but in any other separable metric space, convergence in \( d_{TV} \) is stronger, and often too strong to be useful. For example, a sequence of discrete real-valued random variables never converges in total variation to a continuous limit. \( \Box \)

**Remark 5.5.** If \( S \) is a Banach space (e.g., \( \mathbb{R} \)), then \( d_{TV} \) is an ideal metric of order 0. \( \Box \)

### 5.1. Hellinger metric

Again, let \( S = (S, \mathcal{B}) \) be an arbitrary measure space. As noted above, given any probability measures \( \mu \) and \( \nu \) on \( S \), there exists a \( \sigma \)-finite measure \( \lambda \) such that \( \mu \) and \( \nu \) are absolutely continuous with respect to \( \lambda \). The **Hellinger distance** then is defined by

\[
d_H(\mu, \nu) = \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}} \right\|_{L_2(\lambda)} \tag{5.7}
\]

\[
= \left( \frac{1}{2} \int_S \left| \sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}} \right|^2 d\lambda \right)^{1/2}. \tag{5.8}
\]

Equivalently,

\[
d_H(\mu, \nu)^2 = 1 - \int_S \sqrt{\frac{d\mu}{d\lambda}} \sqrt{\frac{d\nu}{d\lambda}} d\lambda, \tag{5.9}
\]
where the integral is called the Hellinger integral. Note that the Hellinger integral and the Hellinger distance do not depend on the choice of $\lambda$. The Hellinger integral may symbolically be written $\int_S \sqrt{d\mu} \, d\nu$.

**Remark 5.6.** Some authors define the Hellinger distance without the normalization factor $1/\sqrt{2}$ in (5.7), thus obtaining $\sqrt{2}d_H$ in our notation.

With our normalization, $0 \leq d_H(\mu, \nu) \leq 1$, and $h(\mu, \nu) = 1$ if and only if $\mu$ and $\nu$ are mutually singular. $\square$

It is easily seen that

$$d_H(\mu, \nu)^2 \leq d_{TV}(\mu, \nu) \leq d_H(\mu, \nu)\sqrt{2 - d_H(\mu, \nu)^2} \leq \sqrt{2}d_H(\mu, \nu). \quad (5.10)$$

Hence, convergence in the Hellinger metric is equivalent to convergence in total variation. (But rates may differ.) Furthermore, (5.10) implies that the Hellinger metric is complete on $\mathcal{P}(S)$, since $d_{TV}$ is.

The Hellinger integral and distance are convenient when considering product measures. If $S = \prod_{i \in I} S_i$ is a finite or countable product, and $X = (X_i)_{i \in I}$ and $Y = (Y_i)_{i \in I}$ are random variables in $S$ with independent components, then (5.9) leads to the formula

$$d_H(X, Y)^2 = 1 - \prod_{i \in I} (1 - d_H(X_i, Y_i)^2). \quad (5.11)$$

### 6. Prohorov Metric

The Prohorov distance, also called Lévy–Prohorov distance, of two distributions $\mu, \nu \in \mathcal{P}(S)$ is defined as

$$d_P(\mu, \nu) := \inf \{ \varepsilon > 0 : \nu(B) \leq \mu(B^\varepsilon) + \varepsilon, \mu(B) \leq \nu(B^\varepsilon) + \varepsilon, \forall B \in \mathcal{B}(S) \}, \quad (6.1)$$

where

$$B^\varepsilon := \{ x \in S : d(x, B) < \varepsilon \}. \quad (6.2)$$

The Prohorov metric is a metric on $\mathcal{P}(S)$ that generates the weak topology [2, Appendix III], [3, Theorem 8.3.2]. In other words, for random variables $X_n, X$ in $S$,

$$X_n \overset{d}{\to} X \iff d_P(X_n, X) \to 0. \quad (6.3)$$

It follows from the definition (6.1) that $0 \leq d_P(X, Y) \leq 1$.

**Remark 6.1.** The infimum in (6.1) equals the asymmetric version, i.e.,

$$d_P(\mu, \nu) = \inf \{ \varepsilon > 0 : \nu(B) \leq \mu(B^\varepsilon) + \varepsilon, \forall B \in \mathcal{B}(S) \}. \quad (6.4)$$

*Proof.* Let $d'$ be the infimum in (6.4). Obviously, $d' \leq d_P(\mu, \nu)$. On the other hand, suppose that $\varepsilon > d'$, and let $A \in \mathcal{B}(S)$. Let $B := S \setminus A^\varepsilon$. Then, $d(b, a) \geq \varepsilon$ for every $b \in B$ and $a \in A$, and thus $A \subseteq S \setminus B^\varepsilon$. Furthermore, $\nu(B) \leq \mu(B^\varepsilon) + \varepsilon$ by (6.4). Hence,

$$\mu(A) \leq \mu(S \setminus B^\varepsilon) = 1 - \mu(B^\varepsilon) \leq 1 + \varepsilon - \nu(B) = \nu(A^\varepsilon) + \varepsilon. \quad (6.5)$$

Since $A \in \mathcal{B}(S)$ is arbitrary, the definition (6.1) shows that $d_P(\mu, \nu) \leq \varepsilon$. Hence, $d_P(\mu, \nu) \leq d'$, and thus $d_P(\mu, \nu) = d'$. $\square$
Remark 6.2. It is easy to see that it suffices to take the infimum over closed $B$ (or open $B$) in (6.1) and (6.4). Similarly, we may replace the open neighbourhood $B^\varepsilon$ by the closed neighbourhood obtained by replacing $<$ by $\leq$ in (6.2).

Remark 6.3. The Prohorov metric $d_P$ is the minimal metric corresponding to the Ky Fan distance between random variables

$$
\delta_{KF}(X,Y) := \inf \{ \varepsilon > 0 : P(d(X,Y) > \varepsilon) < \varepsilon \},
$$

which itself metrizes convergence in probability. See [15, Corollary 7.5.2].

It is easy to see that (6.6) defines a complete metric on random variables, and thus the Prohorov metric $d_P$ is a complete metric on $\mathcal{P}(S)$, by Lemma 4.1. □

7. Lévy metric

For $S = \mathbb{R}$, the Lévy distance between two real-valued random variables is defined by

$$
d_L(X,Y) := \inf \{ \varepsilon > 0 : F_X(x-\varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x+\varepsilon) + \varepsilon \text{ for } x \in \mathbb{R} \},
$$

This is extended to random variables in $S = \mathbb{R}^q$ by

$$
d_L(X,Y) := \inf \{ \varepsilon > 0 : F_X(x-\varepsilon e) - \varepsilon \leq F_Y(x) \leq F_X(x+\varepsilon e) + \varepsilon \text{ for } x \in \mathbb{R}^q \},
$$

where $e = (1, \ldots, 1) \in \mathbb{R}^q$.

The Lévy metric is a metric that defines the weak topology on $\mathcal{P}(\mathbb{R}^q)$, i.e.,

$$
X_n \xrightarrow{d} X \iff d_L(X_n, X) \to 0.
$$

It is easy to see that the Lévy metric is complete.

Note that for $\mathbb{R}$, the definition (7.1) is the same as the definition (6.1) of the Prohorov metric, but considering only sets $B = (-\infty, x]$, $x \in \mathbb{R}$.

Thus, the condition in (6.1) is more restrictive, and

$$
d_L(X,Y) \leq d_P(X,Y).
$$

There is no corresponding converse inequality. The Lévy and Prohorov metrics on $\mathcal{P}(\mathbb{R})$ are equivalent in the sense that they define the same topology, but they are not uniformly equivalent.

Example 7.1. Let $X$ be uniformly distributed on the odd integers $1, \ldots, 2n-1$, and let $Y = X+1$. Then, $d_P(X,Y) = 1$ (take $B$ as the set of even integers in (6.1)), while $d_L(X,Y) = 1/n$. □

Remark 7.2. Similarly, for $\mathbb{R}^q$, (7.2) is the same as (6.1) with sets $B = \{ y : y \leq x \}$, $x \in \mathbb{R}^q$, provided we equip $\mathbb{R}^q$ with the $\ell_\infty$-metric $d((x_i)_i^q, (y_i)_i^q) := \max_i |x_i - y_i|$. (This metric is equivalent to the usual Euclidean distance, and therefore the corresponding Lévy distance is equivalent to the usual one.) □

---

1The historical relation is the reverse: Prohorov [12] introduced his general metric as an analogue of the Lévy metric on $\mathbb{R}$. 

---
8. Kolmogorov Metric

For $S = \mathbb{R}$, the Kolmogorov distance (or Kolmogorov–Smirnov distance) between two real-valued random variables is defined by

$$d_K(X, Y) := \sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| = \sup_{x \in \mathbb{R}} |P(X \leq x) - P(Y \leq x)|.$$  

(8.1)

(The Kolmogorov distance $d_K$ is often denoted $\rho(X, Y).$) By definition, $0 \leq d_K(X, Y) \leq 1.$ The Kolmogorov metric is complete and ideal of order 0. Comparing (8.1) and (7.1), we see that

$$d_L(X, Y) \leq d_K(X, Y).$$  

(8.2)

There is no corresponding converse inequality, and the Kolmogorov and Lévy distances are not equivalent. In fact, unlike the Lévy distance, the Kolmogorov distance does not define the weak topology on $\mathcal{P}(\mathbb{R}).$

Example 8.1. Let $X_n := 1/n$ (deterministically), so $\mathcal{L}(X_n) = \delta_{1/n}.$ Then, $X_n \xrightarrow{d} 0,$ and $d_L(X_n, 0) = 1/n \to 0,$ but $d_K(X_n, 0) = d_K(\delta_{1/n}, \delta_0) = 1$ for every $n.$

However, if $X$ has a continuous distribution, then

$$X_n \xrightarrow{d} X \iff d_K(X_n, X) \to 0.$$  

(8.3)

The definition (8.1) may also be written

$$d_{TV}(X, Y) := \sup\left\{ \left| P(X \in A) - P(Y \in A) \right| : A \in \mathcal{A} \right\},$$  

(8.4)

where $\mathcal{A}$ is the collection of intervals $\mathcal{A} := \{(-\infty, x] : -\infty < x < \infty\}.$ A comparison with (5.3) yields

$$d_K \leq d_{TV}.$$  

(8.5)

Again, there is no converse inequality. For example, if $X_n \xrightarrow{d} X$ where $X_n$ are discrete but $X$ continuous, then $d_K(X_n, X) \to 0$ by (8.3) while $d_{TV}(X_n, X) = 1$ for all $n.$

Remark 8.2. The Kolmogorov distance (8.1) is a dual metric, for the test functions $1_{(-\infty, x]}, x \in \mathbb{R}.$ Equivalently, $d_K$ equals the dual metric of the space $BV(\mathbb{R})$ of functions of bounded variation, with total variation as the (semi)norm.

8.1. Kolmogorov Metric in Higher Dimension. The Kolmogorov distance can be defined also for $S = \mathbb{R}^q$ by the same formula (8.1), now taking $x \in \mathbb{R}^d.$ Equivalently, it is given by (8.4) where $\mathcal{A}$ is the family of octants $\{y : y \leq x\}$ for $x \in \mathbb{R}^q.$

Obviously, (8.2) and (8.5) hold also for $\mathbb{R}^q.$

Different extensions of the Kolmogorov distance to $\mathbb{R}^q$ can be constructed by (8.4) for other families $\mathcal{A}$ of subsets of $\mathbb{R}^d,$ for example the family of all half-spaces, or of all convex sets. The latter yields the distance (used e.g. in [7])

$$d_{\text{conv}}(X, Y) := \sup\{ |P(X \in C) - P(Y \in C)| : C \subset \mathbb{R}^q \text{ is convex} \}.$$  

(8.6)
Note that in $\mathbb{R}$,
\[ d_K \leq d_{\text{conv}} \leq 2d_K. \] (8.7)

9. The Kantorovich–Rubinshtein metric

We return to a general complete separable metric space $\mathcal{S}$.

The Kantorovich–Rubinshtein distance is the dual distance defined by $\text{BLip}(\mathcal{S}) = \text{BLip}_1(\mathcal{S})$, i.e., see (3.5) and (2.10),
\[ d_{\text{KR}}(X,Y) = \sup \{ \| f(X) - f(Y) \| : \| f \|_{\text{BLip}(\mathcal{S})} \leq 1 \}. \] (9.1)

Another version uses instead the equivalent norm $\| f \|_{\prime \text{BLip}(\mathcal{S})}$ in (2.11) on $\text{BLip}(\mathcal{S})$; we denote this version by $d'_{\text{KR}}(X,Y)$ and note that (2.12) implies
\[ \frac{1}{2}d_{\text{KR}}(X,Y) \leq d'_{\text{KR}}(X,Y) \leq d_{\text{KR}}(X,Y). \] (9.2)

The Kantorovich–Rubinshtein metric $d_{\text{KR}}$ (or $d'_{\text{KR}}$) generates the weak topology in $\mathcal{P}(\mathcal{S})$ [3, Theorem 8.3.2]. In other words, for random variables $X_n, X$ in $\mathcal{S}$,
\[ X_n \overset{d}{\to} X \iff d_{\text{KR}}(X_n, X) \to 0. \] (9.3)

The Kantorovich–Rubinshtein metric is thus equivalent to the Prohorov metric, see (6.3). Moreover, they are uniformly equivalent, and, more precisely, [3, Theorem 8.10.43]
\[ \frac{2}{3}d_{\text{P}}(X,Y)^2 \leq \frac{2d_{\text{P}}(X,Y)^2}{2 + d_{\text{P}}(X,Y)} \leq d'_{\text{KR}}(X,Y) \leq d_{\text{KR}}(X,Y) \leq 3d_{\text{P}}(X,Y). \] (9.4)

Furthermore,
\[ d'_{\text{KR}}(X,Y) \leq 2d_{\text{P}}(X,Y). \] (9.5)

It follows from (9.4) that $d_{\text{KR}}$ and $d'_{\text{KR}}$ are complete metrics on $\mathcal{P}(\mathcal{S})$, since $d_{\text{P}}$ is.

10. The Kantorovich or Wasserstein metric

Let $\delta_1(X,Y) := \mathbb{E}d(X,Y)$ for random variables $X$ and $Y$ in $\mathcal{S}$ defined on a common probability space. This is a metric (for a fixed probability space), noting that $\delta_1(X,Y) = +\infty$ is possible in general; furthermore, $\delta_1$ is finite and (thus a proper metric) on the set of random variables $X$ such that
\[ \mathbb{E}d(X,o) < \infty, \] (10.1)
where we recall that $o$ is a fixed (but arbitrary) point in $\mathcal{S}$. The corresponding minimal metric (3.1) is
\[ \ell_1(X,Y) := \min \{ \mathbb{E}d(X',Y') : X' \overset{d}{=} X, Y' \overset{d}{=} Y \}, \] (10.2)
and it follows that $\ell_1$ is finite for random variables in $\mathcal{S}$ such that (10.1) holds; equivalently, $\ell_1$ is a proper metric on the set of probability measures
\[ \mathcal{P}_1(\mathcal{S}) := \{ \mu \in \mathcal{P}(\mathcal{S}) : \int_{\mathcal{S}} d(x,o) \, d\mu(x) < \infty \}. \] (10.3)

Furthermore, $\ell_1$ is a complete metric on $\mathcal{P}_1(\mathcal{S})$ by Lemma 4.1.
In the remainder of the present section, we consider only random variables and distributions satisfying (10.1) and (10.3).

The distance $\ell_1$ in (10.2) is known under many names, including the Kantorovich distance, the Wasserstein distance\textsuperscript{2}, the Fortet–Mourier distance, the Dudley distance,\textsuperscript{3} Gini’s measure of discrepancy, and (simply) the minimal $L_1$ distance; see [16] for a brief history. Moreover, $\ell_1$ is the special case $p = 1$ of the minimal $L_p$ distance defined in Section 11 and it is also the special case $\zeta_1$ of the Zolotarev distance defined in Section 12.

One reason for the many names of this metric (and for its importance) is that it has several quite different, but equivalent, definitions:

Theorem 10.1. The probability metric $\ell_1$ can be defined by any of the following methods:

(i) The definition (10.2) above, where the infimum always is attained and thus inf can be replaced by min.

(ii) $\ell_1$ is also the dual metric defined by $\operatorname{Lip}_1(S)$:

$$\ell_1(X, Y) = \sup \{ E|f(X) - Ef(Y)| : \|f\|_{\operatorname{Lip}(S)} \leq 1 \}. \quad (10.4)$$

(iii) A different duality:

$$\ell_1(X, Y) = \sup \{ E[f(X) + g(Y) : f, g \in C(S), f(x) + g(y) \leq d(x, y) \} \} \quad (10.5)$$

(iv) If $S = \mathbb{R}$, then the minimum in (10.2) is attained for the monotone coupling $(\bar{X}, \bar{Y}) = (F_X^{-1}(U), F_Y^{-1}(U))$ in (2.3):

$$\ell_1(X, Y) = \mathbb{E}|F_X^{-1}(U) - F_Y^{-1}(U)| = \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)| \, du. \quad (10.6)$$

(v) If $S = \mathbb{R}$, then,

$$\ell_1(X, Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)| \, dx, \quad (10.7)$$

the $L_1$ distance between the distribution functions. (Cf. the Kolmogorov distance (8.1), which is the $L_\infty$ distance.)

Proof. See [3, Theorem 8.10.45] and [15, Corollary 5.3.2] for (i); [3, Theorem 8.10.45] for (ii); [3, Lemma 8.10.44] and [15, Corollary 5.3.2] for (iii); [13, §2.3] and [15, Corollary 7.4.6], for (iv); [15, Theorems 5.5.1 and 7.4.4] for (v).

It follows from Theorem 10.1(ii) that

$$d_{KR} \leq \ell_1. \quad (10.8)$$

In particular, convergence in $\ell_1$ implies convergence in distribution. More precisely:

Theorem 10.2. The following are equivalent, for any $o \in S$:

(i) $\ell_1(X_n, X) \to 0$

(ii) $X_n \overset{d}{\to} X$ and $\mathbb{E}d(X_n, o) \to \mathbb{E}d(X, o)$.

\textsuperscript{2}After L.N. Vasershtein, but with the spelling Wasserstein.

\textsuperscript{3}But in [19] the Dudley distance means our $d'_{KR}$, see Section 9.
(iii) $X_n \overset{d}{\rightarrow} X$ and the random variables $\mathbb{E}d(X_n, o)$ are uniformly integrable.

**Remark 10.3.** If $S$ is a Banach space (e.g., $\mathbb{R}$), then by either Lemma 4.2 or Lemma 4.3, $\ell_1$ is an ideal metric of order 1. \hfill $\Box$

**Remark 10.4.** If the diameter $D := \text{diam}(S) < \infty$, then $\text{Lip}(S)/\mathbb{R} = \text{BLip}(S)/\mathbb{R}$ with equivalent norms, see Remark 2.3, and thus $d_{\text{KR}}$ and $\ell_1$ are equivalent; in fact, with $C = 1 \lor (D/2)$,

$$d_{\text{KR}} \leq \ell_1 \leq C d_{\text{KR}}. \tag{10.9}$$

In particular, if $\text{diam}(S) \leq 2$, then $\text{Lip}(S)/\mathbb{R}$ and $\text{BLip}(S)/\mathbb{R}$ are isometric, and $\ell_1 = d_{\text{KR}}$. \hfill $\Box$

11. **Minimal $L_p$ metric**

Let $0 < p \leq \infty$. The **minimal $L_p$ distance** $\ell_p$ is the minimal metric corresponding to $\delta_p(X, Y) := \|d(X, Y)\|_p = (\mathbb{E}d(X, Y)^p)^{1/p}$, when $p = \infty$ interpreted as $\delta_\infty(X, Y) := \|d(X, Y)\|_\infty = \text{ess sup} d(X, Y)$. I.e.,

$$\ell_p(X, Y) := \inf\{\|d(X, Y)\|_p : X' \overset{d}{=} X, Y' \overset{d}{=} Y\}, \tag{11.1}$$

where the infimum is taken over all couplings of $X$ and $Y$. The infimum is actually attained, see [15, Corollary 5.3.2]. Note that $\ell_p(X, Y)$ may be infinite. Note also that the special case $p = 1$ yields the Kantorovich metric $\ell_1$ in Section 10.

$\ell_p$ is also called the Mallows distance.

Recall that $\delta_p$ is a metric for $1 \leq p \leq \infty$; if $0 < p < 1$, instead $\delta_p(X, Y)^p$ is a metric. Consequently, $\ell_p$ is a probability metric if $1 \leq p \leq \infty$, and $\ell_p^p$ is a probability metric if $0 < p < 1$.

**Remark 11.1.** For any $p < 1$, $d^p$ is another metric on $S$ that defines the same topology as $d$. Obviously, $\ell_p^p$ equals the probability metric $\ell_1$ for the metric space $(S, d^p)$; hence the results for $\ell_1$ in Section 10 immediately extend to corresponding results for $\ell_p^p$, $0 < p \leq 1$.

The metric $\ell_p$ ($\ell_p^p$ if $p < 1$) is finite for random variables $X$ in $S$ such that the $p$th moment of $d(X, o)$ is finite, i.e.,

$$\mathbb{E}d(X, o)^p < \infty, \tag{11.2}$$

where again $o \in S$ is fixed but arbitrary. Equivalently, $\ell_p$ ($\ell_p^p$ if $p < 1$) is a proper metric on the set $\mathcal{P}_p(S)$ of all probability measures in $S$ with finite $p$th absolute moment in the sense

$$\mathcal{P}_p(S) := \{\mu \in \mathcal{P}(S) : \int_S d(x, o)^p \, d\mu(x) < \infty\}. \tag{11.3}$$

The following theorem generalizes Theorem 10.2, see e.g. [4, Theorem 1.1.9] (and [9, Theorem 5.5.9]).

**Theorem 11.2.** Let $0 < p < \infty$, and assume $\mathbb{E}d(X_n, o)^p < \infty$, $n \geq 1$, and $\mathbb{E}d(X, o)^p < \infty$. Then the following are equivalent:

(i) $\ell_p(X_n, X) \to 0$

(ii) $X_n \overset{d}{\rightarrow} X$ and $\mathbb{E}d(X_n, o)^p \to \mathbb{E}d(X, o)^p$. 

(iii) $X_n \overset{d}{\rightarrow} X$ and the random variables $\mathbb{E} d(X_n, o)^p$ are uniformly integrable.

By Lyapounov’s inequality, if $p < q$, then $\delta_p \leq \delta_q$, and thus

$$\ell_p \leq \ell_q, \quad 0 < p \leq q \leq \infty.$$  \hspace{1cm} (11.4)

Since $L^p$ is complete, for any probability space, completeness of $\ell_p$ follows immediately by Lemma 4.1.

**Theorem 11.3.** The metric $\ell_p$ (replaced by $\ell_p^q$ if $p < 1$) is complete, for any $0 < p \leq \infty$ and any complete separable metric space $\mathcal{S}$.

**Remark 11.4.** For $p = \infty$, the minimal $L_\infty$ distance is also given by the formula

$$\ell_\infty(\mu, \nu) := \inf \{ \varepsilon > 0 : \nu(B) \leq \mu(B^\varepsilon), \forall B \in B(\mathcal{S}) \},$$  \hspace{1cm} (11.5)

with $B^\varepsilon$ given by (6.2), see [15, (7.5.15)]. Cf. the definition of the Prohorov distance in (6.1), and note that thus

$$d_p \leq \ell_\infty.$$  \hspace{1cm} (11.6)

As in Remark 6.1, there is also an asymmetric version:

$$\ell_\infty(\mu, \nu) := \inf \{ \varepsilon > 0 : \nu(B) \leq \mu(B^\varepsilon) \forall B \in B(\mathcal{S}) \},$$  \hspace{1cm} (11.7)

11.1. **The Banach space case.** Suppose now that $\mathcal{S}$ is a Banach space, e.g. $\mathbb{R}$. Then $\delta_p(X, Y) = \|X - Y\|_p$ and thus

$$\ell_p(X, Y) := \inf \{ \|X' - Y'\|_p : X' = X, Y' = Y, \}.$$  \hspace{1cm} (11.8)

Furthermore, (11.2) becomes $\mathbb{E}\|X\|^p < \infty$, i.e., that the absolute $p$th moment is finite, and similarly for (11.3).

It follows by Lemma 4.2 that $\ell_p$ is an ideal metric of order 1 for $1 \leq p \leq \infty$, and that $\ell_p^q$ is an ideal metric of order $p$ for $p < 1$.

11.2. **The real case.** For real-valued random variables, the monotone coupling (2.3) is optimal in (11.1) for every $p \geq 1$, see [13, §2.3], [15, Corollary 7.4.6]. Thus:

**Theorem 11.5.** If $\mathcal{S} = \mathbb{R}$, and $p \geq 1$, then the minimum in (11.1) is attained for the monotone coupling $(\bar{X}, \bar{Y}) = (F_X^{-1}(U), F_Y^{-1}(U))$ in (2.3); thus,

$$\ell_p(X, Y) = \|F_X^{-1}(U) - F_Y^{-1}(U)\|_p = \left\{ \begin{array}{ll} \left( \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)|^p du \right)^{1/p}, & 1 \leq p < \infty, \\
\text{ess sup} |F_X^{-1}(u) - F_Y^{-1}(u)|, & p = \infty. \end{array} \right.$$  \hspace{1cm} (11.9)

Theorem 11.5 does not hold for $p < 1$; then the monotone coupling is not always optimal, as is seen by the following example.

**Example 11.6.** Let $X \sim \text{Be}(\frac{1}{2})$ and $Y := X - 1 \overset{d}{=} -X$. The monotone coupling is $(\bar{X}, \bar{Y}) \overset{d}{=} (X, X - 1)$, with $\|\bar{X} - \bar{Y}\|_p = \|1\|_p = 1$ for every $p > 0$, while the coupling $(X', Y') := (X, -X)$ has $\|X' - Y'\|_p = \|2X\|_p = 2^{1-1/p}$ which is smaller when $p < 1$. \hspace{1cm} $\square$
12. Zolotarev metrics

In this section we assume that either
(i) $0 < \alpha \leq 1$ and $S$ is any complete separable metric space, or
(ii) $0 < \alpha < \infty$ and $S$ is a Banach space $B$ (for example $S = \mathbb{R}^q$).

The Zolotarev distance $\zeta_\alpha$, introduced by Zolotarev [17, 18], then is defined as the dual metric given by the space $\text{Lip}_\alpha$ in Section 2.1 or 2.2, i.e., by

$$\zeta_\alpha(X,Y) := \sup \{ |\mathbb{E} f(X) - \mathbb{E} f(Y)| : \|f\|_{\text{Lip}_\alpha(S)} \leq 1 \},$$

(12.1)

with $\|\cdot\|_{\text{Lip}_\alpha(S)}$ given by (2.5) or (2.15). Note that this distance might be $\infty$, or undefined since $\mathbb{E} f(X)$ or $\mathbb{E} f(Y)$ might be undefined; we give simple conditions for it to be finite below.

Lemma 4.3 yields:

Theorem 12.1. $\zeta_\alpha$ is an ideal metric of order $\alpha$.

In the sequel, we treat the two (overlapping) cases (i) and (ii) above separately.

12.1. Zolotarev metric for $0 < \alpha \leq 1$.

Consider first the case $0 < \alpha \leq 1$, so $S$ is an arbitrary (complete, separable) metric space, and $\|\cdot\|_{\text{Lip}_\alpha(S)}$ is given by (2.5).

It is then easy to see that, for an arbitrary fixed $o \in S$, $\zeta_\alpha$ is (defined and) finite at least for random variables $X$ in $S$ such that the $\alpha$th moment of $d(X,o)$ is finite, i.e.,

$$\mathbb{E} d(X,o)^\alpha < \infty.$$

(12.2)

Equivalently, $\zeta_\alpha$ is a proper metric on the space $\mathcal{P}_\alpha(S)$ of probability measures with finite $\alpha$th moment defined in (11.3).

Theorem 12.2. If $0 < \alpha \leq 1$, then $\zeta_\alpha = \ell_\alpha^\alpha$.

Proof. First, if $\alpha = 1$, then $\zeta_1 = \ell_1$ by Theorem 10.1(ii) and (12.1).

For $\alpha < 1$, note as in Remark 11.1 that $d(x,y)^\alpha$ is a metric on $S$, equivalent to $d(x,y)$ and also complete. Furthermore, $\text{Lip}_\alpha(S) = \text{Lip}_1(S,d)$ equals $\text{Lip}_1(S,d^\alpha)$; hence, $\zeta_\alpha$ equals $\zeta_1 = \ell_1$ for the metric space $(S,d^\alpha)$, which equale $\ell_\alpha^\alpha$ by Remark 11.1. □

As a consequence, convergence in $\zeta_\alpha$ for $\alpha \leq 1$ implies weak convergence. More precisely, Theorems 12.2 and 11.2 yield:

Theorem 12.3. Let $0 < \alpha \leq 1$, and assume $\mathbb{E} d(X_n,o)^\alpha < \infty$, $n \geq 1$, and $\mathbb{E} d(X,o)^\alpha < \infty$. Then the following are equivalent:

(i) $\zeta_\alpha(X_n,X) \to 0$

(ii) $X_n \xrightarrow{d} X$ and $\mathbb{E} d(X_n,o)^\alpha \to \mathbb{E} d(X,o)^\alpha$.

(iii) $X_n \xrightarrow{d} X$ and the random variables $\mathbb{E} d(X_n,o)^\alpha$ are uniformly integrable.

Furthermore, Theorems 12.2 and 11.3 show completeness:

Theorem 12.4. Let $0 < \alpha \leq 1$. Then the probability metric $\zeta_\alpha$ is complete, for any complete separable metric space $S$. 

12.2. Zolotarev metric for a Banach space and $0 < \alpha < \infty$. Assume now that $S$ is a separable Banach space $B$. Then $\|x\|_{\text{Lip}_\alpha(S)}$ is given by (2.15), for any $\alpha > 0$. As in Section 2.2, we let $m := \lceil \alpha \rceil - 1$, so $m < \alpha \leq m + 1$. In particular, $m = 0 \iff 0 < \alpha \leq 1$, the case already treated (in greater generality) in Section 12.1 above.

Using a Taylor expansion [5, Théorème 5.6.1] of $f$ at 0, it is easily seen that $\zeta_\alpha(X, Y)$ is finite if

$$E\|X\|^\alpha < \infty \text{ and } E\|Y\|^\alpha < \infty,$$

and, furthermore, $X$ and $Y$ have the same moments up to order $m$, where the $k$th moment of $X$ is $EX^{\hat{k}}$, regarded as an element of the $k$th (completed) projective tensor power $B^{\hat{k}}$. (See [10] for tensor products and higher moments of Banach space valued random variables.)

**Remark 12.5.** For a Banach space $B$, the dual space of $B^{\hat{k}}$ is the space of bounded multilinear mappings $B^k \to \mathbb{R}$; hence $EX^{\hat{k}} = EY^{\hat{k}}$ if and only if $Eg(X, \ldots, X) = Eg(Y, \ldots, Y)$ for every bounded multilinear mapping $B^k \to \mathbb{R}$. Consequently, $X$ and $Y$ have the same moments up to order $m$ if and only if $Ef(X) = Ef(Y)$ for every function $f$ of the form (2.17), i.e., for every function $f$ with $\|f\|_{\text{Lip}_\alpha} = 0$. Conversely, the definition (12.1) implies that this condition is necessary for $\zeta_\alpha(X, Y)$ to be finite. Hence, if (12.3) holds, then $\zeta(X, Y) < \infty$ if and only if $X$ and $Y$ have the same moments up to order $m$. 

We define, for a given sequence $z = (z_1, \ldots, z_m)$ with $z_k \in B^{\hat{k}}$, $k = 1, \ldots, m$,

$$P_{\alpha, z}(B) := \{\mathcal{L}(X) : E\|X\|^\alpha < \infty, EX^{\hat{k}} = z_k, k = 1, \ldots, m\},$$

i.e., the set of probability measures on $B$ with finite absolute $\alpha$th moment and moments $z_1, \ldots, z_m$. Thus $\zeta_\alpha$ is finite on each $P_{\alpha, z}(B)$, and it is obviously a semi-metric there.

For $\alpha > 1$ (so $m \geq 1$) and a general (separable) Banach space $B$, we do not know whether $\zeta_\alpha$ always is a metric on $P_{\alpha, z}(B)$, and if so, whether it is complete. Moreover, according to Bentkus and Rachkauskas [1], it is not hard to show that in a general Banach space, convergence in $\zeta_\alpha$ does not imply weak convergence (convergence in distribution) when $\alpha > 1$; however, as far as we know they never published any details, and we do not know any explicit counter example.

For $\mathbb{R}^d$, and more generally for Hilbert spaces, there are no problems, as shown by the following theorem. (For a proof see [6]; the final assertion is proved already in [8].)

**Theorem 12.6.** If $H$ is a separable Hilbert space and $\alpha > 0$, then $\zeta_\alpha$ is a complete metric on the set $P_{\alpha, z}(H)$ of all probability measures on $H$ with a finite $\alpha$th absolute moment and given $k$th moments $z_k$, $1 \leq k < \alpha$. Moreover, if $X_n, X$ are $H$-valued random variables with distributions in $P_{\alpha, z}(H)$ and $\zeta_\alpha(X_n, X) \to 0$, then $X_n \overset{d}{\to} X$.

**Remark 12.7.** Suppose that we are given a sequence of random variables $X_n$ in $B$, and we want to show that some normalized variables $\tilde{X}_n$ converge in
\[ \zeta_{\alpha}, \text{i.e., } \zeta_{\alpha}(\tilde{X}_n, Y) \to 0 \text{ for some } Y. \]  
To begin with, we need \( \zeta_{\alpha}(\tilde{X}_n, Y) < \infty \) for all (large) \( n \), so by the criterion above, we want besides the moment condition \( \mathbb{E}\|\tilde{X}_n\|^\alpha < \infty \) and \( \mathbb{E}\|Y\|^\alpha < \infty \), also that the \( m \) first moments of \( X_n \) agree with those of \( Y \), and therefore do not depend on \( n \). We consider this condition for some ranges of \( \alpha \).

(i) For \( \alpha \leq 1 \) (\( m = 0 \)), this moment condition is vacuous.

(ii) For \( 1 < \alpha \leq 2 \) (\( m = 1 \)), we thus want \( \mathbb{E}\tilde{X}_n \) to be constant. This is harmless, and can always be achieved by centering to \( \tilde{X}_n := X_n - \mathbb{E}X_n \), which is very often done in any case.

(iii) For \( 2 < \alpha \leq 3 \) (\( m = 2 \)), the condition is more restrictive. Even if \( \tilde{X}_n \) is centered so that \( \mathbb{E}\tilde{X}_n = 0 \), we also need \( \text{Var} \tilde{X}_n \) to be independent of \( n \). In one dimension, \( S = \mathbb{R} \), this can be achieved by the usual standardization \( \bar{X}_n := (X_n - \mathbb{E}X_n)/\sqrt{\text{Var} X_n} \). In higher dimension, it is generally not enough to multiply by a suitable constant; one has to consider \( A_n(X_n - \mathbb{E}X_n) \) for suitable linear operators \( A_n : B \to B \). In an infinite-dimensional space, even this is typically impossible.

(iv) For \( \alpha > 3 \) (\( m \geq 3 \)), also the third moments have to agree. In general, this cannot be achieved by any linear normalization, and thus \( \zeta_{\alpha} \) with \( \alpha > 3 \) is in general not useful in this type of applications. (In principle, one might use it with \( 3 < \alpha \leq 4 \) if all \( X_n \) have symmetric distributions, so the third moments vanish by symmetry. We do not know any such applications.)

For applications, one is thus in practice restricted to \( 0 < \alpha \leq 3 \), and the range \( 2 < \alpha \leq 3 \) requires more work. Nevertheless, this range (in particular \( \alpha = 3 \)) is very useful in some applications, see e.g. [11].

□

References


Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden

Email address: svante.janson@math.uu.se

URL: http://www2.math.uu.se/~svante/