Chapter 14
Distances in Probability Theory

A *probability space* is a *measurable space* $(\Omega, \mathcal{A}, P)$, where $\mathcal{A}$ is the set of all measurable subsets of Ω, and P is a measure on $\mathcal{A}$ with $P(\Omega) = 1$. The set Ω is called a *sample space*. An element $a \in \mathcal{A}$ is called an *event*. In particular, an *elementary event* is a subset of Ω that contains only one element. $P(a)$ is called the *probability* of the event a. The measure P on $\mathcal{A}$ is called a *probability measure*, or *(probability) distribution law*, or simply *(probability) distribution*.

A *random variable* X is a measurable function from a probability space $(\Omega, \mathcal{A}, P)$ into a measurable space, called a *state space* of possible values of the variable; it is usually taken to be the real numbers with the *Borel σ-algebra*, so $X : \Omega \to \mathbb{R}$. The range $\mathcal{X}$ of the random variable X is called the *support* of the distribution P; an element $x \in \mathcal{X}$ is called a *state*.

A distribution law can be uniquely described via a *cumulative distribution function* (CDF, *distribution function*, *cumulative density function*) $F(x)$ which describes the probability that a random value X takes on a value at most x: $F(x) = P(X \leq x) = P(\omega \in \Omega : X(\omega) \leq x)$.

So, any random variable X gives rise to a *probability distribution* which assigns to the interval $[a, b]$ the probability $P(a \leq X \leq b) = P(\omega \in \Omega : a \leq X(\omega) \leq b)$, i.e., the probability that the variable X will take a value in the interval $[a, b]$.

A distribution is called *discrete* if $F(x)$ consists of a sequence of finite jumps at x_i; a distribution is called *continuous* if $F(x)$ is continuous. We consider (as in the majority of applications) only discrete or *absolutely continuous* distributions, i.e., the CDF function $F : \mathbb{R} \to \mathbb{R}$ is *absolutely continuous*. It means that, for every number $\epsilon > 0$, there is a number $\delta > 0$ such that, for any sequence of pairwise disjoint intervals $[x_k, y_k]$, $1 \leq k \leq n$, the inequality $\sum_{1 \leq k \leq n} (y_k - x_k) < \delta$ implies the inequality $\sum_{1 \leq k \leq n} |F(y_k) - F(x_k)| < \epsilon$.

A distribution law also can be uniquely defined via a *probability density function* (PDF, *density function*, *probability function*) $p(x)$ of the underlying random variable. For an absolutely continuous distribution, the CDF is almost everywhere differentiable, and the PDF is defined as the derivative $p(x) = F^{'}(x)$ of the CDF; so, $F(x) = P(X \leq x) = \int_{-\infty}^{x} p(t)dt$, and

M.M. Deza and E. Deza, *Encyclopedia of Distances*,

DOI 10.1007/978-3-642-00234-2_14,

$\int_a^b p(t)dt = P(a \le X \le b)$. In the discrete case, the PDF (the density of the random variable X) is defined by its values $p(x_i) = P(X = x)$; so $F(x) = \sum_{x_i \le x} p(x_i)$. In contrast, each elementary event has probability zero in any continuous case.

The random variable X is used to "push-forward" the measure P on Ω to a measure dF on $\mathbb{R}$. The underlying probability space is a technical device used to guarantee the existence of random variables and sometimes to construct them.

For simplicity, we usually present the discrete version of probability metrics, but many of them are defined on any measurable space; see [Bass89], [Cha08]. For a probability distance d on random quantities, the conditions $P(X = Y) = 1$ or equality of distributions imply (and characterize) $d(X, Y) = 0$; such distances are called [Rach91] *compound* or *simple* distances, respectively. In many cases, some *ground* distance d is given on the state space $\mathcal{X}$ and the presented distance is a lifting of it to a distance on distributions.

In Statistics, many of the distances below, between distributions P_1 and P_2, are used as measures of *goodness of fit* between estimated, P_2, and theoretical, P_1, distributions. Also, in Statistics, a distance that not satisfy the triangle inequality, is often called a **distance statistic**; a *statistic* is a function of a sample which is independent of its distribution.

Below we use the notation $\mathbb{E}[X]$ for the *expected value* (or *mean*) of the random variable X: in the discrete case $\mathbb{E}[X] = \sum_x xp(x)$, in the continuous case $\mathbb{E}[X] = \int xp(x)dx$. The *variance* of X is $\mathbb{E}[(X - \mathbb{E}[X])^2]$. Also we denote $p_X = p(x) = P(X = x)$, $F_X = F(x) = P(X \le x)$, $p(x, y) = P(X = x, Y = y)$.

14.1 Distances on random variables

All distances in this section are defined on the set $\mathbf{Z}$ of all random variables with the same support $\mathcal{X}$; here $X, Y \in \mathbf{Z}$.

- ***p*-average compound metric**
  Given $p \ge 1$, the ***p*-average compound metric** (or *L_p-metric between variables*) is a metric on $\mathbf{Z}$ with $\mathcal{X} \subset \mathbb{R}$ and $\mathbb{E}[|Z|^p] < \infty$ for all $Z \in \mathbf{Z}$, defined by

$$(\mathbb{E}[|X - Y|^p])^{1/p} = (\sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} |x - y|^p p(x, y))^{1/p}.$$

  For $p = 2$ and ∞, it is called, respectively, the *mean-square distance* and *essential supremum distance* between variables.

- **Absolute moment metric**
  Given $p \geq 1$, the **absolute moment metric** is a metric on $\mathbf{Z}$ with $\mathcal{X} \subset \mathbb{R}$ and $\mathbb{E}[|Z|^p] < \infty$ for all $Z \in \mathbf{Z}$, defined by
  $$(|(\mathbb{E}[|X|^p])^{1/p} - (\mathbb{E}[|Y|^p])^{1/p}|.$$
  For $p = 1$ it is called the *engineer metric*.
- **Indicator metric**
  The **indicator metric** is a metric on $\mathbf{Z}$, defined by
  $$\mathbb{E}[1_{X \neq Y}] = \sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} 1_{x \neq y} p(x,y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{X}, x \neq y} p(x,y).$$
  (Cf. **Hamming metric** in Chap. 1.)
- **Ky Fan metric** K
  The **Ky Fan metric** K is a metric K on $\mathbf{Z}$, defined by
  $$\inf\{\epsilon > 0 : P(|X - Y| > \epsilon) < \epsilon\}.$$
  It is the case $d(x,y) = |X - Y|$ of the **probability distance**.
- **Ky Fan metric** K^*
  The **Ky Fan metric** K^* is a metric K^* on $\mathbf{Z}$, defined by
  $$\mathbb{E}\left[\frac{|X - Y|}{1 + |X - Y|}\right] = \sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} \frac{|x - y|}{1 + |x - y|} p(x,y).$$
- **Probability distance**
  Given a metric space $(\mathcal{X}, d)$, the **probability distance** on $\mathbf{Z}$ is defined by
  $$\inf\{\epsilon > 0 : P(d(X,Y) > \epsilon) < \epsilon\}.$$

14.2 Distances on distribution laws

All distances in this section are defined on the set $\mathcal{P}$ of all distribution laws such that corresponding random variables have the same range $\mathcal{X}$; here $P_1, P_2 \in \mathcal{P}$.

- **L_p-metric between densities**
  The **L_p-metric between densities** is a metric on $\mathcal{P}$ (for a countable $\mathcal{X}$), defined, for any $p \geq 1$, by
  $$(\sum_x |p_1(x) - p_2(x)|^p)^{\frac{1}{p}}.$$

For $p = 1$, one half of it is called the **total variation metric** (or *variational distance*, *trace-distance*). For $p = 2$, it is the **Patrick-Fisher distance**. The *point metric* $\sup_x |p_1(x) - p_2(x)|$ corresponds to $p = \infty$.

The **Lissak-Fu distance** with parameter $\alpha > 0$ is defined as $\sum_x |p_1(x) - p_2(x)|^\alpha$.

- **Bayesian distance**
  The *error probability in classification* is the following error probability of the optimal Bayes rule for the classification into 2 classes with a priori probabilities $\phi, 1-\phi$ and corresponding densities p_1, p_2 of the observations:

$$P_e = \sum_x \min(\phi p_1(x), (1-\phi)p_2(x)).$$

The **Bayesian distance** on $\mathcal{P}$ is defined by $1 - P_e$.

For the classification into m classes with *a priori* probabilities ϕ_i, $1 \leq i \leq m$, and corresponding densities p_i of the observations, the error probability becomes

$$P_e = 1 - \sum_x p(x) \max_i P(C_i|x),$$

where $P(C_i|x)$ is the *a posteriori* probability of the class C_i given the observation x and $p(x) = \sum_{i=1}^m \phi_i P(x|C_i)$. The *general mean distance between m classes C_i* (cf. m-hemi-metric in Chap. 3) is defined (Van der Lubbe 1979), for $\alpha > 0$ and $\beta > 1$, by

$$\sum_x p(x)(\sum_i P(C_i|x)^\beta)^\alpha.$$

The case $\alpha = 1, \beta = 2$ corresponds to the *Bayesian distance* in Devijver (1974); the case $\beta = \frac{1}{\alpha}$ was considered in Trouborst, Baker, Boekee and Boxma (1974).

- **Mahalanobis semi-metric**
  The **Mahalanobis semi-metric** (or *quadratic distance*) is a semi-metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}^n$), defined by

$$\sqrt{(\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X])^T A^{-1} (\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X])}$$

for a given positive-definite matrix A.

- **Engineer semi-metric**
  The **engineer semi-metric** is a semi-metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$), defined by

$$|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]| = |\sum_x x(p_1(x) - p_2(x))|.$$

- **Stop-loss metric of order** m
  The **stop-loss metric of order** m is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$), defined by
$$\sup_{t\in\mathbb{R}} \sum_{x\geq t} \frac{(x-t)^m}{m!}(p_1(x)-p_2(x)).$$
- **Kolmogorov–Smirnov metric**
  The **Kolmogorov–Smirnov metric** (or *Kolmogorov metric*, *uniform metric*) is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$), defined by
$$\sup_{x\in\mathbb{R}} |P_1(X\leq x)-P_2(X\leq x)|.$$
  The **Kuiper distance** on $\mathcal{P}$ is defined by
$$\sup_{x\in\mathbb{R}}(P_1(X\leq x)-P_2(X\leq x))+\sup_{x\in\mathbb{R}}(P_2(X\leq x)-P_1(X\leq x)).$$
  (Cf. **Pompeiu–Eggleston metric** in Chap. 9.)
  The **Anderson–Darling distance** on $\mathcal{P}$ is defined by
$$\sup_{x\in\mathbb{R}} \frac{|(P_1(X\leq x)-P_2(X\leq x)|}{\ln\sqrt{(P_1(X\leq x)(1-P_1(X\leq x))}}.$$
  The **Crnkovic–Drachma distance** is defined by
$$\sup_{x\in\mathbb{R}}(P_1(X\leq x)-P_2(X\leq x))\ln\frac{1}{\sqrt{(P_1(X\leq x)(1-P_1(X\leq x))}}+$$
$$+\sup_{x\in\mathbb{R}}(P_2(X\leq x)-P_1(X\leq x))\ln\frac{1}{\sqrt{(P_1(X\leq x)(1-P_1(X\leq x))}}.$$
  The above three distances are used in Statistics as measures of *goodness of fit*, especially, for VaR (Value at Risk) measurements in Finance.
- **Cramer–von Mises distance**
  The **Cramer–von Mises distance** is a distance on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$), defined by
$$\int_{-\infty}^{+\infty}(P_1(X\leq x)-P_2(X\leq x))^2dx.$$
  This is the squared L_2**-metric** between cumulative density functions.
- **Levy–Sibley metric**
  The **Levy metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$ only), defined by
$$\inf\{\epsilon>0: P_1(X\leq x-\epsilon)-\epsilon\leq P_2(X\leq x)\leq P_1(X\leq x+\epsilon)+\epsilon \text{ for any } x\in\mathbb{R}\}.$$
  It is a special case of the **Prokhorov metric** for $(\mathcal{X}, d)=(\mathbb{R}, |x-y|)$.

- **Prokhorov metric**
Given a metric space $(\mathcal{X}, d)$, the **Prokhorov metric** on $\mathcal{P}$ is defined by

$$\inf\{\epsilon > 0 : P_1(X \in B) \leq P_2(X \in B^\epsilon) + \epsilon \text{ and } P_2(X \in B) \leq P_1(X \in B^\epsilon) + \epsilon\},$$

where B is any Borel subset of $\mathcal{X}$, and $B^\epsilon = \{x : d(x, y) < \epsilon, y \in B\}$.
It is the smallest (over all joint distributions of pairs (X, Y) of random variables X, Y such that the marginal distributions of X and Y are P_1 and P_2, respectively) **probability distance** between random variables X and Y.
- **Dudley metric**
Given a metric space $(\mathcal{X}, d)$, the **Dudley metric** on $\mathcal{P}$ is defined by

$$\sup_{f \in F} |\mathbb{E}_{P_1}[f(X)] - \mathbb{E}_{P_2}[f(X)]| = \sup_{f \in F} |\sum_{x \in \mathcal{X}} f(x)(p_1(x) - p_2(x))|,$$

where $F = \{f : \mathcal{X} \to \mathbb{R}, ||f||_\infty + Lip_d(f) \leq 1\}$, and $Lip_d(f) = \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.
- **Szulga metric**
Given a metric space $(\mathcal{X}, d)$, the **Szulga metric** on $\mathcal{P}$ is defined by

$$\sup_{f \in F} |(\sum_{x \in \mathcal{X}} |f(x)|^p p_1(x))^{1/p} - (\sum_{x \in \mathcal{X}} |f(x)|^p p_2(x))^{1/p}|,$$

where $F = \{f : X \to \mathbb{R}, Lip_d(f) \leq 1\}$, and $Lip_d(f) = \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.
- **Zolotarev semi-metric**
The **Zolotarev semi-metric** is a semi-metric on $\mathcal{P}$, defined by

$$\sup_{f \in F} |\sum_{x \in \mathcal{X}} f(x)(p_1(x) - p_2(x))|,$$

where F is any set of functions $f : \mathcal{X} \to \mathbb{R}$ (in the continuous case, F is any set of such bounded continuous functions); cf. **Szulga metric**, **Dudley metric**.
- **Convolution metric**
Let G be a separable locally compact abelian group, and let $C(G)$ be the set of all real bounded continuous functions on G vanishing at infinity. Fix a function $g \in C(G)$ such that $|g|$ is integrable with respect to the Haar measure on G, and $\{\beta \in G^* : \widehat{g}(\beta) = 0\}$ has empty interior; here G^* is the dual group of G, and $\widehat{g}$ is the Fourier transform of g.

The **convolution metric** (or *smoothing metric*) is defined (Yukich 1985), for any two finite signed Baire measures P_1 and P_2 on G, by

$$\sup_{x \in G} | \int_{y \in G} g(xy^{-1})(dP_1 - dP_2)(y)|.$$

This metric can also be seen as the difference $T_{P_1}(g) - T_{P_2}(g)$ of *convolution operators* on $C(G)$ where, for any $f \in C(G)$, the operator $T_P f(x)$ is $\int_{y \in G} f(xy^{-1})dP(y)$.

- **Discrepancy metric**
  Given a metric space $(\mathcal{X}, d)$, the **discrepancy metric** on $\mathcal{P}$ is defined by

  $$\sup\{|P_1(X \in B) - P_2(X \in B)| : B \text{ is any closed ball}\}.$$

- **Bi-discrepancy semi-metric**
  The **bi-discrepancy semi-metric** is a semi-metric evaluating the proximity of distributions P_1, P_2 (over different collections $\mathcal{A}_1, \mathcal{A}_2$ of measurable sets), defined in the following way:

  $$D(P_1, P_2) + D(P_2, P_1),$$

  where $D(P_1, P_2) = \sup\{\inf\{P_2(C) : B \subset C \in \mathcal{A}_2\} - P_1(B) : B \in \mathcal{A}_1\}$ (*discrepancy*).
- **Le Cam distance**
  The **Le Cam distance** is a semi-metric, evaluating the proximity of probability distributions P_1, P_2 (on different spaces $\mathcal{X}_1, \mathcal{X}_2$), defined in the following way:

  $$\max\{\delta(P_1, P_2), \delta(P_2, P_1)\},$$

  where $\delta(P_1, P_2) = \inf_B \sum_{x_2 \in \mathcal{X}_2} |BP_1(X_2 = x_2) - BP_2(X_2 = x_2)|$ is the *Le Cam deficiency*. Here $BP_1(X_2 = x_2) = \sum_{x_1 \in \mathcal{X}_1} p_1(x_1)b(x_2|x_1)$, where B is a probability distribution over $\mathcal{X}_1 \times \mathcal{X}_2$, and

  $$b(x_2|x_1) = \frac{B(X_1 = x_1, X_2 = x_2)}{B(X_1 = x_1)} = \frac{B(X_1 = x_1, X_2 = x_2)}{\sum_{x \in \mathcal{X}_2} B(X_1 = x_1, X_2 = x)}.$$

  So, $BP_2(X_2 = x_2)$ is a probability distribution over $\mathcal{X}_2$, since $\sum_{x_2 \in \mathcal{X}_2} b(x_2|x_1) = 1$.

  Le Cam distance is not a probabilistic distance, since P_1 and P_2 are defined over different spaces; it is a distance between statistical experiments (models).
- **Skorokhod–Billingsley metric**
  The **Skorokhod–Billingsley metric** is a metric on $\mathcal{P}$, defined by

  $$\inf_f \max \left\{ \sup_x |P_1(X \le x) - P_2(X \le f(x))|, \sup_x |f(x) - x|, \sup_{x \ne y} \left| \ln \frac{f(y) - f(x)}{y - x} \right| \right\},$$

  where $f : \mathbb{R} \to \mathbb{R}$ is any strictly increasing continuous function.

- **Skorokhod metric**
The **Skorokhod metric** is a metric on $\mathcal{P}$, defined by

$$\inf\{\epsilon > 0 : \max\{\sup_x |P_1(X < x) - P_2(X \le f(x))|, \sup_x |f(x) - x|\} < \epsilon\},$$

where $f : \mathbb{R} \to \mathbb{R}$ is a strictly increasing continuous function.

- **Birnbaum–Orlicz distance**
The **Birnbaum–Orlicz distance** is a distance on $\mathcal{P}$, defined by

$$\sup_{x \in \mathbb{R}} f(|P_1(X \le x) - P_2(X \le x)|),$$

where $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is any non-decreasing continuous function with $f(0) = 0$, and $f(2t) \le Cf(t)$ for any $t > 0$ and some fixed $C \ge 1$. It is a **near-metric**, since the C**-triangle inequality** $d(P_1, P_2) \le C(d(P_1, P_3) + d(P_3, P_2))$ holds.

Birnbaum–Orlicz distance is also used, in Functional Analysis, on the set of all integrable functions on the segment $[0, 1]$, where it is defined by $\int_0^1 H(|f(x) - g(x)|)dx$, where H is a non-decreasing continuous function from $[0, \infty)$ onto $[0, \infty)$ which vanishes at the origin and satisfies the *Orlicz condition*: $\sup_{t>0} \frac{H(2t)}{H(t)} < \infty$.

- **Kruglov distance**
The **Kruglov distance** is a distance on $\mathcal{P}$, defined by

$$\int f(P_1(X \le x) - P_2(X \le x))dx,$$

where $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is any even strictly increasing function with $f(0) = 0$, and $f(s + t) \le C(f(s) + f(t))$ for any $s, t \ge 0$ and some fixed $C \ge 1$. It is a **near-metric**, since the C**-triangle inequality** $d(P_1, P_2) \le C(d(P_1, P_3) + d(P_3, P_2))$ holds.

- **Burbea–Rao distance**
Consider a continuous convex function $\phi(t) : (0, \infty) \to \mathbb{R}$ and put $\phi(0) = \lim_{t \to 0} \phi(t) \in (-\infty, \infty]$. The convexity of ϕ implies non-negativity of the function $\delta_\phi : [0, 1]^2 \to (-\infty, \infty]$, defined by $\delta_\phi(x, y) = \frac{\phi(x)+\phi(y)}{2} - \phi(\frac{x+y}{2})$ if $(x, y) \ne (0, 0)$, and $\delta_\phi(0, 0) = 0$.

The corresponding **Burbea–Rao distance** on $\mathcal{P}$ is defined by

$$\sum_x \delta_\phi(p_1(x), p_2(x)).$$

- **Bregman distance**
Consider a differentiable convex function $\phi(t) : (0, \infty) \to \mathbb{R}$, and put $\phi(0) = \lim_{t \to 0} \phi(t) \in (-\infty, \infty]$. The convexity of ϕ implies that the

function $\delta_\phi : [0,1]^2 \to (-\infty, \infty]$ defined by continuous extension of $\delta_\phi(u,v) = \phi(u) - \phi(v) - \phi^{'}(v)(u-v)$, $0 < u, v \leq 1$, on $[0,1]^2$ is non-negative.

The corresponding **Bregman distance** on $\mathcal{P}$ is defined by

$$\sum_1^m \delta_\phi(p_i, q_i).$$

(Cf. **Bregman quasi-distance**.)

- **f-divergence of Csizar**
The **f-divergence of Csizar** is a function on $\mathcal{P} \times \mathcal{P}$, defined by

$$\sum_x p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right),$$

where f is a continuous convex function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$.

The cases $f(t) = t \ln t$ and $f(t) = (t-1)^2/2$ correspond to the **Kullback–Leibler distance** and to the **χ^2-distance** below, respectively. The case $f(t) = |t-1|$ corresponds to the L_1-*metric between densities*, and the case $f(t) = 4(1-\sqrt{t})$ (as well as $f(t) = 2(t+1) - 4\sqrt{t}$) corresponds to the squared **Hellinger metric**.

Semi-metrics can also be obtained, as the square root of the f-divergence of Csizar, in the cases $f(t) = (t-1)^2/(t+1)$ (the **Vajda–Kus semi-metric**), $f(t) = |t^a - 1|^{1/a}$ with $0 < a \leq 1$ (the **generalized Matusita distance**), and $f(t) = \frac{(t^a+1)^{1/a} - 2^{(1-a)/a}(t+1)}{1-1/\alpha}$ (the **Osterreicher semi-metric**).

- **Fidelity similarity**
The **fidelity similarity** (or *Bhattacharya coefficient*, *Hellinger affinity*) on $\mathcal{P}$ is

$$\rho(P_1, P_2) = \sum_x \sqrt{p_1(x) p_2(x)}.$$

- **Hellinger metric**
In terms of the **fidelity similarity** ρ, the **Hellinger metric** (or *Hellinger-Kakutani metric*) on $\mathcal{P}$ is defined by

$$(2 \sum_x (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2)^{\frac{1}{2}} = 2(1 - \rho(P_1, P_2))^{\frac{1}{2}}.$$

Sometimes, $(\sum_x (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2)^{\frac{1}{2}}$ is called the **Matusita distance**, while $(\sum_x (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2$ is called the *squared-chord distance.*

- **Harmonic mean similarity**
The **harmonic mean similarity** is a similarity on $\mathcal{P}$, defined by

$$2 \sum_x \frac{p_1(x) p_2(x)}{p_1(x) + p_2(x)}.$$

- **Bhattacharya distance** 1
  In terms of the **fidelity similarity** ρ, the **Bhattacharya distance** 1 on $\mathcal{P}$ is
$$(\arccos \rho(P_1, P_2))^2.$$
  Twice this distance is used also in Statistics and Machine Learning, where it is called the **Fisher distance**.
- **Bhattacharya distance** 2
  In terms of the **fidelity similarity** ρ, the **Bhattacharya distance** 2 on $\mathcal{P}$ is
$$-\ln \rho(P_1, P_2).$$
- χ^2**-distance**
  The χ^2**-distance** (or **Pearson** χ^2**-distance**) is a quasi-distance on $\mathcal{P}$, defined by
$$\sum_x \frac{(p_1(x) - p_2(x))^2}{p_2(x)}.$$
  The **Neyman** χ^2**-distance** is a quasi-distance on $\mathcal{P}$, defined by
$$\sum_x \frac{(p_1(x) - p_2(x))^2}{p_1(x)}.$$
  The probabilistic **symmetric** χ^2**-measure** is a distance on $\mathcal{P}$, defined by
$$2\sum_x \frac{(p_1(x) - p_2(x))^2}{p_1(x) + p_2(x)}.$$
  The half of the probabilistic **symmetric** χ^2**-measure** is called *squared* χ^2.
- **Separation quasi-distance**
  The **separation distance** is a quasi-distance on $\mathcal{P}$ (for a countable $\mathcal{X}$) defined by
$$\max_x \left(1 - \frac{p_1(x)}{p_2(x)}\right).$$
  (Not to be confused with **separation distance** in Chap. 9.)
- **Kullback–Leibler distance**
  The **Kullback–Leibler distance** (or *relative entropy*, *information deviation*, *information gain*, *KL-distance*) is a quasi-distance on $\mathcal{P}$, defined by
$$KL(P_1, P_2) = \mathbb{E}_{P_1}[\ln L] = \sum_x p_1(x) \ln \frac{p_1(x)}{p_2(x)},$$
  where $L = \frac{p_1(x)}{p_2(x)}$ is the *likelihood ratio*. Therefore,
$$KL(P_1, P_2) = -\sum_x (p_1(x) \ln p_2(x)) + \sum_x (p_1(x) \ln p_1(x)) = H(P_1, P_2) - H(P_1),$$

where $H(P_1)$ is the *entropy* of P_1, and $H(P_1, P_2)$ is the *cross-entropy* of P_1 and P_2.

If P_2 is the product of marginals of P_1 (say, $p_2(x,y) = p_1(x)p_1(y)$), the KL-distance $KL(P_1, P_2)$ is called the *Shannon information quantity* and (cf. **Shannon distance**) is equal to $\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} p_1(x,y) \ln \frac{p_1(x,y)}{p_1(x)p_1(y)}$.

- **Skew divergence**
The **skew divergence** is a quasi-distance on $\mathcal{P}$, defined by
$$KL(P_1, aP_2 + (1-a)P_1),$$
where $a \in [0,1]$ is a constant, and KL is the **Kullback–Leibler distance**.
The cases $a = 1$ and $a = \frac{1}{2}$ correspond to $KL(P_1, P_2)$ and K*-divergence*.

- **Jeffrey divergence**
The **Jeffrey divergence** (or J*-divergence*, *divergence distance*, $KL2$*-distance*) is a symmetric version of the **Kullback–Leibler distance**, defined by
$$KL(P_1, P_2) + KL(P_2, P_1) = \sum_x (p_1(x) - p_2(x)) \ln \frac{p_1(x)}{p_2(x)}.$$
For $P_1 \to P_2$, the Jeffrey divergence behaves like the χ^2**-distance**.

- **Jensen–Shannon divergence**
The **Jensen–Shannon divergence** is defined by
$$aKL(P_1, P_3) + (1-a)KL(P_2, P_3),$$
where $P_3 = aP_1 + (1-a)P_2$, and $a \in [0,1]$ is a constant (cf. **clarity similarity**).
In terms of *entropy* $H(P) = -\sum_x p(x) \ln p(x)$, the Jensen–Shannon divergence is equal to $H(aP_1 + (1-a)P_2) - aH(P_1) - (1-a)H(P_2)$.

- **Topsøe distance**
Let P_3 denote $\frac{1}{2}(P_1+P_2)$. The **Topsøe distance** (or *information statistics*) is a symmetric version of the **Kullback–Leibler distance** (or rather of the K*-divergence* $KL(P_1, P_3)$):
$$KL(P_1, P_3) + KL(P_2, P_3) = \sum_x \left(p_1(x) \ln \frac{p_1(x)}{p_3(x)} + p_2(x) \ln \frac{p_2(x)}{p_3(x)} \right).$$
The Topsøe distance is twice the **Jensen–Shannon divergence** with $a = \frac{1}{2}$. Some authors use the term *Jensen–Shannon divergence* only for this value of a. It is not a metric, but its square root is a metric.
The **Taneja distance** is defined by
$$\sum_x p_3(x) \ln \frac{p_3(x)}{\sqrt{p_1(x)p_2(x)}}.$$

- **Resistor-average distance**
The Johnson–Simanović's **resistor-average distance** is a symmetric version of the **Kullback–Leibler distance** on $\mathcal{P}$ which is defined by the harmonic sum
$$\left(\frac{1}{KL(P_1,P_2)}+\frac{1}{KL(P_2,P_1)}\right)^{-1}.$$
Cf. **resistance metric** for graphs in Chap. 15.
- **Ali–Silvey distance**
The **Ali–Silvey distance** is a quasi-distance on $\mathcal{P}$, defined by the functional
$$f(\mathbb{E}_{P_1}[g(L)]),$$
where $L=\frac{p_1(x)}{p_2(x)}$ is the *likelihood ratio*, f is a non-decreasing function on $\mathbb{R}$, and g is a continuous convex function on $\mathbb{R}_{\geq 0}$ (cf. **f-divergence of Csizar**).
The case $f(x)=x$, $g(x)=x\ln x$ corresponds to the **Kullback–Leibler distance**; the case $f(x)=-\ln x$, $g(x)=x^t$ corresponds to the **Chernoff distance**.
- **Chernoff distance**
The **Chernoff distance** (or *Rényi cross-entropy*) is a distance on $\mathcal{P}$, defined by
$$\max_{t\in[0,1]} D_t(P_1,P_2),$$
where $0\leq t\leq 1$ and $D_t(P_1,P_2)=-\ln\sum_x(p_1(x))^t(p_2(x))^{1-t}$ (called the *Chernoff coefficient* or *Hellinger path*), which is proportional to the **Rényi distance**.
The case $t=\frac{1}{2}$ corresponds to the **Bhattacharya distance** 2.
- **Rényi distance**
The **Rényi distance** (or *order t Rényi entropy*) is a quasi-distance on $\mathcal{P}$, defined, for any constant $0\leq t<1$, by
$$\frac{1}{1-t}\ln\sum_x p_2(x)\left(\frac{p_1(x)}{p_2(x)}\right)^t.$$
The limit of the Rényi distance, for $t\to 1$, is the **Kullback–Leibler distance**. For $t=\frac{1}{2}$, one half of the Rényi distance is the **Bhattacharya distance** 2 (cf. **f-divergence of Csizar** and **Chernoff distance**).
- **Clarity similarity**
The **clarity similarity** is a similarity on $\mathcal{P}$, defined by
$$(KL(P_1,P_3)+KL(P_2,P_3))-(KL(P_1,P_2)+KL(P_2,P_1))=$$
$$=\sum_x\left(p_1(x)\ln\frac{p_2(x)}{p_3(x)}+p_2(x)\ln\frac{p_1(x)}{p_3(x)}\right),$$

where KL is the **Kullback–Leibler distance**, and P_3 is a fixed probability law. It was introduced in [CCL01] with P_3 being the probability distribution of English.

- **Shannon distance**
Given a *measure space* $(\Omega, \mathcal{A}, P)$, where the set Ω is finite and P is a probability measure, the *entropy* (or *Shannon information entropy*) of a function $f : \Omega \to X$, where X is a finite set, is defined by
$$H(f) = -\sum_{x \in X} P(f = x) \log_a (P(f = x));$$
here $a = 2$, e, or 10 and the unit of entropy is called a *bit*, *nat*, or *dit* (digit), respectively. The function f can be seen as a partition of the measure space. For any two such partitions $f : \Omega \to X$ and $g : \Omega \to Y$, denote by $H(f, g)$ the entropy of the partition $(f, g) : \Omega \to X \times Y$ (*joint entropy*), and by $H(f|g)$ the *conditional entropy* (or *equivocation*); then the **Shannon distance** between f and g is a metric defined by
$$H(f|g) + H(g|f) = 2H(f, g) - H(f) - H(g) = H(f, g) - I(f; g),$$
where $I(f; g) = H(f) + H(g) - H(f, g)$ is the *Shannon mutual information.*
If P is the uniform probability law, then Goppa showed that the Shannon distance can be obtained as a limiting case of the **finite subgroup metric**.
In general, the **information metric** (or **entropy metric**) between two random variables (information sources) X and Y is defined by
$$H(X|Y) + H(Y|X) = H(X, Y) - I(X; Y),$$
where the *conditional entropy* $H(X|Y)$ is defined by $\sum_{x \in X} \sum_{y \in Y} p(x, y)$ ln $p(x|y)$, and $p(x|y) = P(X = x|Y = y)$ is the conditional probability.
The **Rajski distance** (or *normalized information metric*) is defined (Rajski 1961, for discrete probability distributions X, Y) by
$$\frac{H(X|Y) + H(Y|X)}{H(X, Y)} = 1 - \frac{I(X; Y)}{H(X, Y)}.$$
It is equal to 1 if X and Y are independent. (Cf., a different one, **normalized information distance** in Chap. 11).

- **Kantorovich–Mallows–Monge–Wasserstein metric**
Given a metric space $(\mathcal{X}, d)$, the **Kantorovich–Mallows–Monge–Wasserstein metric** is defined by
$$\inf \mathbb{E}_S[d(X, Y)],$$

where the infimum is taken over all joint distributions S of pairs (X, Y) of random variables X, Y such that marginal distributions of X and Y are P_1 and P_2.

For any **separable** metric space $(\mathcal{X}, d)$, this is equivalent to the **Lipschitz distance between measures** $\sup_f \int f d(P_1 - P_2)$, where the supremum is taken over all functions f with $|f(x) - f(y)| \leq d(x, y)$ for any $x, y \in \mathcal{X}$.

More generally, the L_p**-Wasserstein distance** for $\mathcal{X} = \mathbb{R}^n$ is defined by

$$(\inf \mathbb{E}_S[d^p(X, Y)])^{1/p},$$

and, for $p = 1$, it is also called the $\overline{\rho}$*-distance*. For $(\mathcal{X}, d) = (\mathbb{R}, |x - y|)$, it is also called the L_p*-metric between distribution functions* (CDF), and can be written as

$$\begin{aligned}(\inf \mathbb{E}[|X - Y|^p])^{1/p} &= \left(\int_{\mathbb{R}} |F_1(x) - F_2(x)|^p dx \right)^{1/p} \\ &= \left(\int_0^1 |F_1^{-1}(x) - F_2^{-1}(x)|^p dx \right)^{1/p}\end{aligned}$$

with $F_i^{-1}(x) = \sup_u (P_i(X \leq x) < u)$.

The case $p = 1$ of this metric is called the **Monge–Kantorovich metric** or **Hutchinson metric** (in Fractal Theory), **Wasserstein metric**, **Fortet–Mourier metric**.

- **Ornstein $\overline{d}$-metric**
  The **Ornstein $\overline{d}$-metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} = \mathbb{R}^n$), defined by

$$\frac{1}{n} \inf \int_{x,y} \left(\sum_{i=1}^{n} 1_{x_i \neq y_i} \right) dS,$$

where the infimum is taken over all joint distributions S of pairs (X, Y) of random variables X, Y such that marginal distributions of X and Y are P_1 and P_2.