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*. $P(a)$ 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 $\mathbb{R}$ with the *Borel* σ*-algebra*, so $X : \Omega \to \mathbb{R}$. The range $\mathcal{X}$ of the 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* (or simply, *distribution*) *function* CDF, 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* (or *density*, *probability*) *function* PDF 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 $\int_a^b p(t)dt = P(a \leq X \leq b)$. In the discrete case,

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-44342-2_14

the PDF is $\sum_{x_i \le x} p(x_i)$, where $p(x) = P(X = x)$ is the *probability mass function*. But $p(x) = 0$ for each fixed x 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.

We usually present the discrete version of probability metrics, but many of them are defined on any measurable space; see [Bass89, Bass13, 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. Often, 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. A quasi-distance between distributions is also called **divergence** or *distance statistic*.

Below we denote $p_X = p(x) = P(X = x)$, $F_X = F(x) = P(X \le x)$, $p(x, y) = P(X = x, Y = y)$. We denote by $\mathbb{E}[X]$ 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 *covariance* between the random variables X and Y is $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. The *variance* and *standard deviation* of X are $Var(X) = Cov(X, X)$ and $\sigma(X) = \sqrt{Var(X)}$, respectively. The *correlation* between X and Y is $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$; cf. Chap. 17.

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.
- **Lukaszyk–Karmovski metric**
  The **Lukaszyk–Karmovski metric** (2001) on $\mathbb{Z}$ with $\mathcal{X} \subset \mathbb{R}$ is defined by

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} |x - y| p(x) p(y).$$

For continuous random variables, it is defined by $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|F(x)F(y)dxdy$. This function can be positive for $X = Y$. Such possibility is excluded, and so, it will be a distance metric, if and only if it holds

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|\delta(x-\mathbb{E}[X])\delta(y-\mathbb{E}[Y])dxdy = |\mathbb{E}[X]-\mathbb{E}[Y]|.$$

- **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 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 **variational distance** (or *total variation distance*, *Kolmogorov 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 two 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-hemimetric in Chap. 3) is defined (Van der Lubbe, 1979) for $\alpha > 0, \beta > 1$ by

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

  The case $\alpha = 1, \beta = 2$ corresponds to the *Bayesian distance* in Devijver, 1974; the case $\beta = \frac{1}{\alpha}$ was considered in Trouborst et al., 1974.
- **Mahalanobis semimetric**
  The **Mahalanobis semimetric** is a semimetric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}^n$) defined by

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

  for a given positive-semidefinite matrix A; its square is a **Bregman quasi-distance** (cf. Chap. 13). Cf. also the **Mahalanobis distance** in Chap. 17.

- **Engineer semimetric**
  The **engineer semimetric** is a semimetric 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 (1948) by

$$\sup_{x \in \mathbb{R}} |P_1(X \leq x) - P_2(X \leq x)|.$$

  This metric is used, for example, in Biology as a measure of sexual dimorphism.
  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 **Crnkovic–Drachma distance** is defined by

$$\sup_{x \in \mathbb{R}}(P_1(X \leq x) - P_2(X \leq x)) \ln \frac{1}{\sqrt{(P_1(X \leq x)(1 - P_1(X \leq x))}} +$$

$$+ \sup_{x \in \mathbb{R}}(P_2(X \leq x) - P_1(X \leq x)) \ln \frac{1}{\sqrt{(P_1(X \leq x)(1 - P_1(X \leq x))}}.$$

- **Cramér–von Mises distance**
  The **Cramér–von Mises distance** (1928) is defined on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$) by

$$\omega^2 = \int_{-\infty}^{+\infty} (P_1(X \leq x) - P_2(X \leq x))^2 dP_2(x).$$

  The **Anderson–Darling distance** (1954) on $\mathcal{P}$ is defined by

$$\int_{-\infty}^{+\infty} \frac{(P_1(X \leq x) - P_2)(X \leq x))^2}{(P_2(X \leq x)(1 - P_2(X \leq x))} dP_2(x).$$

In Statistics, above distances of Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and, below, χ^2**-distance** are the main measures of *goodness of fit* between estimated, P_2, and theoretical, P_1, distributions.
They and other distances were generalized (for example by Kiefer, 1955, and Glick, 1969) on *K-sample setting*, i.e., some convenient generalized distances $d(P_1, \ldots, P_K)$ were defined. Cf. **m-hemimetric** in Chap. 3.

- **Energy distance**
The **energy distance** (Székely, 2005) between cumulative density functions $F(X)$, $F(Y)$ of two independent random vectors $X, Y \in \mathbb{R}^n$ is defined by

$$d(F(X), F(Y)) = 2\mathbb{E}[||(X - Y||] - \mathbb{E}[||X - X'||] - \mathbb{E}[||(Y - Y'||],$$

where X, X' are *iid* (independent and identically distributed), Y, Y' are iid and $||.||$ is the length of a vector. Cf. **distance covariance** in Chap. 17.
It holds $d(F(X), F(Y)) = 0$ if and only if X, Y are iid.
- **Prokhorov metric**
Given a metric space $(\mathcal{X}, d)$, the **Prokhorov metric** on $\mathcal{P}$ is defined (1956) by

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

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.
- **Levy–Sibley metric**
The **Levy–Sibley metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$ only) defined by

$$\inf\{\epsilon > 0 : P_1(X \le x - \epsilon) - \epsilon \le P_2(X \le x) \le P_1(X \le x + \epsilon) + \epsilon \text{ for any } x \in \mathbb{R}\}.$$

It is a special case of the **Prokhorov metric** for $(\mathcal{X}, d) = (\mathbb{R}, |x - y|)$.
- **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 $F = \{f : \mathcal{X} \to \mathbb{R}, ||f||_\infty + Lip_d(f) \le 1\}$, and $Lip_d(f) = \sup_{x,y \in \mathcal{X}, x \ne y} \frac{|f(x) - f(y)|}{d(x,y)}$.
- **Szulga metric**
Given a metric space $(\mathcal{X}, d)$, the **Szulga metric** (1982) 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 semimetric**
  The **Zolotarev semimetric** is a semimetric on $\mathcal{P}$, defined (1976) 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 $\{\beta \in G^* : \hat{g}(\beta) = 0\}$ has empty interior; here G^* is the dual group of G, and $\hat{g}$ 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

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

  It 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 $\int_{y\in G} f(xy^{-1})dP(y)$.
  In particular, this metric can be defined on the space of probability measures on $\mathbb{R}^n$, where g is a PDF satisfying above conditions.
- **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 semimetric**
  The **bi-discrepancy semimetric** (evaluating the proximity of distributions P_1, P_2 over different collections $\mathcal{A}_1, \mathcal{A}_2$ of measurable sets) is defined by

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

  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** (1974) is a semimetric, evaluating the proximity of probability distributions P_1, P_2 (on different spaces $\mathcal{X}_1, \mathcal{X}_2$) and defined as follows:

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

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 (1956) 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** (1931) 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 nondecreasing continuous function with $f(0) = 0$, and $f(2t) \le C f(t)$ 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 nondecreasing 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** (1973) 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.

- **Bregman divergence**
  Given a differentiable strictly convex function $\phi(p) : \mathbb{R}^n \to \mathbb{R}$ and $\beta \in (0, 1)$, the **skew Jensen** (or *skew Burbea–Rao*) **divergence** on $\mathcal{P}$ is (Basseville–Cardoso, 1995)

$$J_\phi^{(\beta)}(P_1, P_2) = \beta\phi(p_1) + (1 - \beta)\phi(p_2) - \phi(\beta p_1 + (1 - \beta)p_2).$$

  The **Burbea–Rao distance** (1982) is the case $\beta = \frac{1}{2}$ of it, i.e., it is

$$\sum_x \left(\frac{\phi(p_1(x)) + \phi(p_2(x))}{2} - \phi(\frac{p_1(x) + (p_2(x)}{2}) \right).$$

  The **Bregman divergence** (1967) is a quasi-distance on $\mathcal{P}$ defined by

$$\sum_x (\phi(p_1(x)) - \phi(p_2(x)) - (p_1(x) - p_2(x))\phi'(p_2(x))) = \lim_{\beta \to 1} \frac{1}{\beta} J_\phi^{(\beta)}(P_1, P_2).$$

  The **generalised Kullback–Leibler distance** $\sum_x p_1(x) \ln \frac{p_1(x)}{p_2(x)} - \sum_x (p_1(x) - p_2(x))$ and **Itakura–Saito distance** (cf. Chap. 21) $\sum_x \frac{p_1(x)}{p_2(x)} - \ln \frac{p_1(x)}{p_2(x)} - 1$ are the cases $\phi(p) = \sum_x p(x) \ln p(x) - \sum_x p(x)$ and $\phi(p) = -\sum_x \ln p(x)$ of the Bregman divergence. Cf. **Bregman quasi-distance** in Chap. 13.
  Csizár, 1991, proved that the **Kullback–Leibler distance** is the only **Bregman divergence** which is an f**-divergence**.

- f**-divergence**
  Given a convex function $f(t) : \mathbb{R}_{\ge 0} \to \mathbb{R}$ with $f(1) = 0$, $f'(1) = 0$, $f''(1) = 1$, the f**-divergence** (independently, Csizár, 1963, Morimoto, 1963, Ali–Silvey, 1966, Ziv–Zakai, 1973, and Akaike, 1974) on $\mathcal{P}$ is defined by

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

  The cases $f(t) = t \ln t$ and $f(t) = (t-1)^2$ correspond to the **Kullback–Leibler distance** and to the χ^2**-distance** below, respectively. The case $f(t) = |t - 1|$ corresponds to the **variational distance**, 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**.

Semimetrics can also be obtained, as the square root of the f-divergence, in the cases $f(t) = (t-1)^2/(t+1)$ (the **Vajda–Kus semimetric**), $f(t) = |t^a - 1|^{1/a}$ 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 semimetric**).

- **α-divergence**
  Given $\alpha \in \mathbb{R}$, the **α-divergence** (independently, Csizár, 1967, Havrda–Charvát, 1967, Cressie–Read, 1984, and Amari, 1985) is defined as $KL(P_1, P_2)$, $KL(P_2, P_1)$ for $\alpha = 1, 0$ and for $\alpha \neq 0, 1$, it is

$$\frac{1}{\alpha(1-\alpha)} \left(1 - \sum_x p_2(x) \left(\frac{p_1(x)}{p_2(x)} \right)^\alpha \right).$$

  The **Amari divergence** come from the above by the transformation $\alpha = \frac{1+t}{2}$.
- **Harmonic mean similarity**
  The **harmonic mean similarity** is a similarity on $\mathcal{P}$ defined by

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

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

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

  Cf. more general **quantum fidelity similarity** in Chap. 24.
- **Hellinger metric**
  In terms of the **fidelity similarity** ρ, the **Hellinger metric** (or **Matusita distance**, *Hellinger–Kakutani metric*) on $\mathcal{P}$ is defined by

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

- **Bhattacharya distance** 1
  In terms of the **fidelity similarity** ρ, the **Bhattacharya distance** 1 (1946) is

$$(\arccos \rho(P_1, P_2))^2$$

  for $P_1, P_2 \in \mathcal{P}$. Twice this distance is the **Rao distance** from Chap. 7. It is used also in Statistics and Machine Learning, where it is called the *Fisher distance*.
  The **Bhattacharya distance** 2 (1943) on $\mathcal{P}$ is defined by

$$-\ln \rho(P_1, P_2).$$

- **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 half of χ^2-distance is also called *Kagan's divergence*.
  The probabilistic **symmetric χ^2-measure** is a distance on $\mathcal{P}$, defined by

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

- **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 (1951) 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 $\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} p_1(x, y) \ln \frac{p_1(x,y)}{p_1(x)p_1(y)}$.
  The **exponential divergence** is defined by $\sum_x p_1(x)(\ln \frac{p_1(x)}{p_2(x)})^2$.

- **Distance to normality**
  For a continuous distribution P on $\mathbb{R}$, the *differential entropy* is defined by

$$h(P) = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx.$$

  It is $\ln(\delta\sqrt{2\pi e})$ for a *normal* (or *Gaussian*) *distribution* $g_{\delta,\mu}(x) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(x-\mu)^2}{2\delta^2}\right)$ with variance δ^2 and mean μ.
  The **distance to normality** (or *negentropy*) of P is the **Kullback–Leibler distance** $KL(P, g) = \int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{g(x)}\right) dx = h(g) - h(P)$, where q is a normal distribution with the same variance as P. So, it is nonnegative and equal to 0 if and only if $P = g$ almost everywhere. Cf. **Shannon distance**.
  Also, $h(u_{a,b}) = \ln(b-a)$ for an *uniform distribution* with minimum a and maximum $b > a$, i.e., $u_{a,b}(x) = \frac{1}{b-a}$, if $x \in [a, b]$, and it is 0, otherwise. It holds $h(u_{a,b}) \geq h(P)$ for any distribution P with support contained in $[a, b]$; so, $h(u_{a,b}) - h(P)$ can be called the *distance to uniformity*. Tononi, 2008, used it in his model of consciousness.
- **Jeffrey distance**
  The **Jeffrey distance** (or *J-divergence*, *KL2-distance*) is a symmetric version of the **Kullback–Leibler distance** defined (1946) on $\mathcal{P}$ by

$$KL(P_1, P_2) + KL(P_2, P_1) = \sum_x ((p_1(x) - p_2(x)) \ln \frac{p_1(x)}{p_2(x)}.$$

  The **Aitchison distance** (1986) is defined by $\sqrt{\sum_x (\ln \frac{p_1(x)g(p_1)}{p_2(x)g(p_2)})^2}$, where $g(p) = (\prod_x p(x))^{\frac{1}{n}}$ is the geometric mean of components $p(x)$ of p.
- **Resistor-average distance**
  The **resistor-average distance** is (Johnson–Simanović, 2000) a symmetric version of the **Kullback–Leibler distance** on $\mathcal{P}$ which is defined by the harmonic sum

$$\left(\frac{1}{KL(P_1, P_2)} + \frac{1}{KL(P_2, P_1)}\right)^{-1}.$$

- **Jensen–Shannon divergence**
  Given a number $\beta \in [0, 1]$ and $P_1, P_2 \in \mathcal{P}$, let P_3 denote $\beta P_1 + (1-\beta)P_2$. The **skew divergence** and the **Jensen–Shannon divergence** between P_1 and P_2 are defined on $\mathcal{P}$ as $KL(P_1, P_3)$ and $\beta KL(P_1, P_3) + (1-\beta)KL(P_2, P_3)$, respectively. Here KL is the **Kullback–Leibler distance**; cf. **clarity similarity**.
  In terms of *entropy* $H(P) = -\sum_x p(x) \ln p(x)$, the Jensen–Shannon divergence is $H(\beta P_1 + (1-\beta)P_2) - \beta H(P_1) - (1-\beta)H(P_2)$, i.e., the **Jensen divergence** (cf. **Bregman divergence**).

Let $P_3 = \frac{1}{2}(P_1 + P_2)$, i.e., $\beta = \frac{1}{2}$. Then the skew divergence and twice the Jensen–Shannon divergence are called **K-divergence** and **Topsøe distance** (or *information statistics*), respectively. The Topsøe distance is a symmetric version of $KL(P_1, P_2)$. It is not a metric, but its square root is a metric.

- **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.
- **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 nondecreasing function on $\mathbb{R}$, and g is a continuous convex function on $\mathbb{R}_{\geq 0}$ (cf. **f-divergence**).
  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*) on $\mathcal{P}$ is defined (1954) by

$$\max_{t \in (0,1)} D_t(P_1, P_2),$$

  where $0 < t < 1$ and $D_t(P_1, P_2) = -\ln \sum_x (p_1(x))^t (p_2(x))^{1-t}$ (called the *Chernoff coefficient*) which is proportional to the **Rényi distance**.
- **Rényi distance**
  Given $t \in \mathbb{R}$, the **Rényi distance** (or *order t Rényi entropy*, 1961) is a quasi-distance on $\mathcal{P}$ defined as the **Kullback–Leibler distance** $KL(P_1, P_2)$ if $t = 1$, and, otherwise, by

$$\frac{1}{1-t} \ln \sum_x p_2(x) \left(\frac{p_1(x)}{p_2(x)} \right)^t.$$

  For $t = \frac{1}{2}$, one half of the Rényi distance is the **Bhattacharya distance** 2. Cf. **f-divergence** and **Chernoff distance**.

- **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).
- **Transportation distance**
  Given a metric space $(\mathcal{X}, d)$, the **transportation distance** (and/or, according to Villani, 2009, **Monge–Kantorovich–Wasserstein–Rubinstein–Ornstein–Gini–Dall'Aglio–Mallows–Tanaka distance**) is the metric defined by

$$W_1(P_1, P_2) = \inf \mathbb{E}_S[d(X, Y)] = \inf_S \int_{(X,Y) \in \mathcal{X} \times \mathcal{X}} d(X, Y) dS(X, Y),$$

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 $x, y \in \mathcal{X}$. Cf. **Dudley metric**.

In general, for a Borel function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, the **$c$-transportation distance** $T_c(P_1, P_2)$ is $\inf \mathbb{E}_S[c(X, Y)]$. It is the minimal total transportation cost if $c(X, Y)$ is the cost of transporting a unit of mass from the location X to the location Y. Cf. the **Earth Mover's distance** (Chap. 21), which is a discrete form of it.

The **L_p-Wasserstein distance** is $W_p = (T_{d^p})^{1/p} = (\inf \mathbb{E}_S[d^p(X, Y)])^{1/p}$. For $(\mathcal{X}, d) = (\mathbb{R}, |x - y|)$, it is also called the *L_p-metric between distribution functions* (CDF) F_i with $F_i^{-1}(x) = \sup_u(P_i(X \leq x) < u)$, and can be written as

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

For $p = 1$, this metric is called **Monge–Kantorovich metric** (or **Wasserstein metric**, **Fortet–Mourier metric**, **Hutchinson metric**, **Kantorovich–Rubinstein metric**). For $p = 2$, it is the **Levy–Fréchet metric** (Fréchet, 1957).

- **Ornstein $\overline{d}$-metric**

  The **Ornstein $\overline{d}$-metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} = \mathbb{R}^n$) defined (1974) 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.

- **Distances between belief assignments**

  In *Bayesian* (or *subjective*, *evidential*) interpretation, a probability can be assigned to any statement, even if no random process is involved, as a way to represent its subjective plausibility, or the degree to which it is supported by the available evidence, or, mainly, degree of belief. Within this approach, *imprecise probability* generalizes Probability Theory to deal with scarce, vague, or conflicting information. The main model is *Dempster–Shafer theory*, which allows evidence to be combined.

Given a set X, a (basic) **belief assignment** is a function $m : P(X) \to [0,1]$ (where $P(X)$ is the set of all subsets of X) with $m(\emptyset) = 0$ and $\sum_{A \subset P(X)} m(A) = 1$. Probability measures are a special case in which $m(A) > 0$ only for singletons.
For the classic probability $P(A)$, it holds then $\text{Bel}(A) \leq P(A) \leq \text{Pl}(A)$, where the *belief function* and *plausibility function* are defined, respectively, by

$$\text{Bel}(A) = \sum_{B:B \subset A} m(B) \text{ and } \text{Pl}(A) = \sum_{B:B \cap A \neq \emptyset} m(B) = 1 - \text{Bel}(\overline{A}).$$

The original (Dempster, 1967) *conflict factor* between two belief assignments m_1 and m_2 was defined as $c(m_1, m_2) = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$. This is not a distance since $c(m, m) > 0$. The combination of m_1 and m_2, seen as independent sources of evidence, is defined by $m_1 \oplus m_2(A) = \frac{1}{1-c(m_1,m_2)} \sum_{B \cap C = A} m_1(B) m_2(C)$.
Usually, a distance between m_1 and m_2 estimates the difference between these sources in the form $d_U = |U(m_1) - U(m_2)|$, where U is an uncertainty measure; see Sarabi-Jamab et al., 2013, for a comparison of their performance. In particular, this distance is called:

confusion (Hoehle, 1981) if $U(m) - \sum_A m(A) \log_2 \text{Bel}(A)$;
dissonance (Yager, 1983) if $U(m) = E(m) = -\sum_A m(A) \log_2 \text{Pl}(A)$;
Yager's factor (Eager, 1983) if $U(m) = 1 - \sum_{A \neq \emptyset} \frac{m(A)}{|A|}$;
possibility-based (Smets, 1983) if $U(m) = -\sum_A \log_2 \sum_{B:A \subset B} m(B)$;
U-uncertainty (Dubois–Prade, 1985) if $U(m) = I(m) = -\sum_A m(A) \log_2 |A|$;
Lamata–Moral's (1988) if $U(m) = \log_2(\sum_A m(A)|A|)$ and $U(m) = E(m) + I(m)$;
discord (Klir–Ramer, 1990) if $U(m) = D(m) = -\sum_A m(A) \log_2(1 - \sum_B m(B) \frac{|B \setminus A|}{|B|})$ and a variant: $U(m) = D(m) + I(m)$;
strife (Klir–Parviz, 1992) if $U(m) = -\sum_A m(A) \log_2(\sum_B m(B) \frac{|A \cap B|}{|A|})$;
Pal et al.'s (1993) if $U(m) = G(m) = -\sum_A \log_2 m(A)$ and $U(m) = G(m) + I(m)$;
total conflict (George–Pal, 1996) if $U(m) = \sum_A m(A) \sum_B (m(B)(1 - \frac{|A \cap B|}{|A \cup B|}))$.

Among other distances used are the **cosine distance** $1 - \frac{m_1^T m_2}{||m_1|| ||m_2||}$, the **Mahalanobis distance** $\sqrt{(m_1 - m_2)^T A (m_1 - m_2)}$ for some matrices A, and *pignistic-based* one (Tessem, 1993) $\max_A \{| \sum_{B \neg \emptyset} (m_1(B) - m_2(B) \frac{|A \cap B|}{|B|} |\}$.