Taxonomy of Principal Distances & Divergences

A structured reference: families, equations, and example uses

Reorganized from Frank Nielsen's chart (Sony CSL) into clean families.

How to read this sheet — metric vs. divergence.
A **metric** (true distance) satisfies four rules: $d(p,q) \geq 0$; $d(p,q) = 0 \Leftrightarrow p = q$; *symmetry* $d(p,q) = d(q,p)$; and the *triangle inequality* $d(p,r) \leq d(p,q) + d(q,r)$. A **divergence** keeps only the first two (non-negativity and identity): it is generally *asymmetric*, $D(p\|q) \neq D(q\|p)$, and need not obey the triangle inequality. Divergences measure how one distribution differs from a reference; metrics measure mutual separation. The symbol $\|$ denotes the asymmetric "from / to" argument order.

1. Metric distances on vectors (geometry of points)

The classical distances between two points $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$. All are true metrics.

Euclidean distance, L_2 (Pythagoras, ~500 BC)

$$d_2(\mathbf{p},\mathbf{q}) = \sqrt{\textstyle\sum_i (p_i - q_i)^2}$$

Where it's used: Straight-line distance. k-means, nearest-neighbour search, least-squares regression.

Manhattan / city-block distance, L_1

$$d_1(\mathbf{p},\mathbf{q}) = \textstyle\sum_i |p_i - q_i|$$

Where it's used: Grid/route distance; robust to outliers; underlies LASSO sparsity.

Minkowski distance, L_k-norm (H. Minkowski, 1864–1909)

$$d_k(\mathbf{p},\mathbf{q}) = \Big(\textstyle\sum_i |p_i - q_i|^k\Big)^{1/k}$$

Where it's used: Tunable family: k=1 Manhattan, k=2 Euclidean, $k \to \infty$ Chebyshev ($\max_i |p_i - q_i|$).

Quadratic (generalized) distance

$$d_Q(\mathbf{p},\mathbf{q}) = \sqrt{(\mathbf{p}-\mathbf{q})^\top Q\,(\mathbf{p}-\mathbf{q})}\,, \quad Q \succeq 0$$

Where it's used: Feature-weighted / cross-bin distance, e.g. comparing colour histograms.

Mahalanobis metric (P. Mahalanobis, 1936)

$$d_\Sigma(\mathbf{p},\mathbf{q}) = \sqrt{(\mathbf{p}-\mathbf{q})^\top \Sigma^{-1} (\mathbf{p}-\mathbf{q})}$$

Where it's used: Scale- and correlation-aware distance ($Q = \Sigma^{-1}$). Outlier detection, classification, metric learning.

Hamming distance

$$d_H(\mathbf{p},\mathbf{q}) = \big|\{\, i : p_i \neq q_i \,\}\big|$$

Where it's used: Count of differing symbols. Error-correcting codes, DNA comparison, bit strings.

String / time-series distances. Edit-based or alignment-based dissimilarities such as **Levenshtein** (edit distance) and **Dynamic Time Warping**. Used for spell-checking, bioinformatics, speech and gesture recognition.

2. Riemannian & information geometry (curved manifolds)

Here "distance" becomes the length of the shortest path (*geodesic*) on a curved space. For statistical models, the natural curvature comes from the Fisher information.

Riemannian metric tensor & geodesic length (B. Riemann, 1826–1866)

$$ds^2 = g_{ij}\,dx^i\,dx^j, \qquad L(\gamma) = \int \sqrt{g_{ij}\,\dot{x}^i\dot{x}^j}\;dt$$

Where it's used: Shortest paths on curved surfaces/manifolds; foundation of general relativity and shape spaces.

Finsler metric tensor

$$g_{ij}(x,y) = \tfrac{1}{2}\,\frac{\partial^2 F^2(x,y)}{\partial y^i \partial y^j}$$

Where it's used: Generalizes Riemannian geometry to direction-dependent norms (anisotropic costs).

Fisher information matrix (R. A. Fisher, 1890–1962)

$$\mathbf{I}(\theta) = \mathbb{E}\Big[\Big(\tfrac{\partial}{\partial\theta}\ln p(X\mid\theta)\Big)\Big(\tfrac{\partial}{\partial\theta}\ln p(X\mid\theta)\Big)^{\top}\Big]$$

Where it's used: The "local entropy" / natural metric on a statistical model. Cramér–Rao bound, natural-gradient descent.

Fisher–Rao distance

$$\rho_{FR}(p,q) = \min_{\gamma}\int_0^1 \sqrt{\dot\gamma(t)^{\top}\mathbf{I}(\theta)\,\dot\gamma(t)}\;dt$$

Where it's used: Intrinsic geodesic distance between distributions; the Riemannian "gold standard" on statistical manifolds.

3. Statistical divergences between distributions

Asymmetric measures of difference between probability densities p, q (w.r.t. a base measure μ). Two great umbrella families (f-divergences and Bregman divergences) generate most of the others.

3a. f-divergences (Ali–Silvey 1966; Csiszár 1967)

A single template generates a whole zoo of divergences via a convex generator f with $f(1) = 0$:

$$D_f(p\|q) = \int p\, f\Big(\tfrac{q}{p}\Big)\,d\mu.$$

Its conjugate (reverse) is $D_{f^\star}(p\|q) = D_f(q\|p)$.

Kullback–Leibler divergence / relative entropy (1951) $\quad f(t) = -\log t$

$$\mathrm{KL}(p\|q) = \int p\,\log\frac{p}{q}\,d\mu = \mathbb{E}_p\Big[\log\tfrac{p}{q}\Big]$$

Where it's used: Maximum-likelihood, cross-entropy loss, variational inference, model selection. Note $H(p) = \mathrm{KL}(p\|u)$ up to a constant.

Pearson χ^2 divergence (K. Pearson, 1857–1936) $\quad f(t) = (t-1)^2$

$$\chi^2(p\|q) = \int \frac{(q-p)^2}{p}\,d\mu$$

Where it's used: Goodness-of-fit testing; local (second-order) approximation to KL.

Hellinger distance $\quad (f(t) = (\sqrt{t}-1)^2)$

$$H(p,q) = \sqrt{\int \big(\sqrt{p}-\sqrt{q}\big)^2 d\mu} = \sqrt{2\big(1-\textstyle\int\sqrt{pq}\,d\mu\big)}$$

Where it's used: A symmetric, bounded *true metric* between densities. Robust statistics, density estimation.

Total variation distance $(f(t) = \frac{1}{2}|t-1|)$

$$\mathrm{TV}(p,q) = \tfrac{1}{2}\int |p-q|\,d\mu$$

Where it's used: The strongest "statistical distinguishability"; coupling arguments, mixing times.

Amari α-divergence (S. Amari, 1985)

$$f_\alpha(t) = \frac{4}{1-\alpha^2}\Big(1 - t^{\frac{1+\alpha}{2}}\Big), \quad -1 < \alpha < 1; \quad \begin{cases} t\log t & \alpha = 1 \\ -\log t & \alpha = -1 \end{cases}$$

Where it's used: One dial spanning KL (α=1), reverse-KL (α=−1) and Hellinger (α=0). Core of information geometry.

3b. Bregman divergences (L. Bregman, 1967)

Generated by a strictly convex potential F: the gap between F and its tangent plane at θ_2.

$$B_F(\theta_1\|\theta_2) = F(\theta_1) - F(\theta_2) - \langle\,\theta_1 - \theta_2,\ \nabla F(\theta_2)\rangle.$$

Legendre duality: $D_{F^*}\big(\nabla F(\theta_1)\|\nabla F(\theta_2)\big) = B_F(\theta_2\|\theta_1)$.

Squared Euclidean. $F(\mathbf{x}) = \|\mathbf{x}\|^2 \Rightarrow B_F = \|\theta_1 - \theta_2\|^2$. Centroid clustering / k-means.

Kullback–Leibler (discrete). F = negative Shannon entropy $\Rightarrow B_F = \mathrm{KL}$. *KL is the unique divergence that is **both** an f-divergence and a Bregman divergence.*

Itakura–Saito divergence (Burg entropy, $F = -\sum_i \log x_i$)

$$\mathrm{IS}(p\|q) = \sum_i \Big(\frac{p_i}{q_i} - \log\frac{p_i}{q_i} - 1\Big)$$

Where it's used: Scale-invariant spectral distortion. Speech/audio coding, non-negative matrix factorization (NMF).

3c. Overlap / α-power family (Bhattacharyya, Chernoff, Rényi)

All built from the affinity integral $\int p^\alpha q^{\,1-\alpha}\,d\mu$.

Bhattacharyya distance (A. Bhattacharyya, 1967)

$$d_B(p,q) = -\log\int \sqrt{p\,q}\;d\mu \quad \big(\text{coefficient } BC = \textstyle\int\sqrt{pq}\big)$$

Where it's used: Class-separability measure; object tracking; feature selection.

Chernoff divergence / information (H. Chernoff, 1952)

$$C_\alpha(p\|q) = -\log\int p^\alpha q^{1-\alpha}d\mu, \quad C(p,q) = \max_{\alpha\in(0,1)} C_\alpha(p\|q)$$

Where it's used: Optimal error exponent in binary hypothesis testing.

Rényi divergence (A. Rényi, 1961)

$$R_\alpha(p\|q) = \frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}d\mu \xrightarrow[\alpha\to 1]{} \mathrm{KL}(p\|q)$$

Where it's used: Differential privacy accounting, information theory, generalized entropies.

3d. Symmetrized & Jensen-type divergences

Jeffreys divergence (symmetric KL)

$$J(p,q) = \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)$$

Where it's used: A symmetric KL when direction is arbitrary.

Jensen–Shannon divergence

$$\mathrm{JS}(p,q) = \tfrac{1}{2}\mathrm{KL}\big(p\,\big\|\,m\big) + \tfrac{1}{2}\mathrm{KL}\big(q\,\big\|\,m\big), \quad m = \tfrac{p+q}{2}$$

Where it's used: Symmetric, always finite, and $\sqrt{\text{JS}}$ is a metric. Original GAN objective; comparing text/topic distributions.

Burbea–Rao / Jensen divergence

$$J_F(p,q) = \frac{F(p)+F(q)}{2} - F\Big(\frac{p+q}{2}\Big)$$

Where it's used: The "Jensen gap" of a convex F; JS is the case $F = -H$ (negative Shannon entropy).

4. Entropies (the functionals divergences are built from)

Entropy measures the uncertainty/spread of a single distribution; most divergences above are differences or gaps of these functionals.

Shannon / Boltzmann–Gibbs entropy (Boltzmann 1878; Shannon 1948)

$$H(p) = -\int p \log p \; d\mu \qquad \big(\text{physics: } S = -k\int p \log p \, d\mu\big)$$

Where it's used: Information content, source coding, thermodynamics.

Rényi entropy (1961)

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \int p^\alpha \, d\mu$$

Where it's used: Additive generalization of Shannon entropy; collision/min-entropy in cryptography.

Tsallis entropy — non-additive (1988)

$$T_\alpha(p) = \frac{1}{1-\alpha}\Big(\int p^\alpha \, d\mu - 1\Big)$$

Where it's used: Non-extensive (long-range correlated) systems in statistical physics.

Sharma–Mittal entropy (two-parameter unifier)

$$h_{\alpha,\beta}(p) = \frac{1}{1-\beta}\left(\Big(\int p^\alpha d\mu\Big)^{\frac{1-\beta}{1-\alpha}} - 1\right)$$

Where it's used: Unifies Shannon, Rényi and Tsallis entropies as limiting cases.

5. Distances between sets & whole metric spaces

Hausdorff distance

$$d_{\text{Haus}}(X,Y) = \max\Big\{ \sup_{x\in X} \rho(x,Y),\ \sup_{y\in Y} \rho(X,y)\Big\}$$

Where it's used: How far two *sets* are; shape/image matching, template comparison.

Gromov–Hausdorff distance

$$d_{GH}(X,Y) = \inf_{\phi_X,\phi_Y} \rho^Z_{\text{Haus}}\big(\phi_X(X), \phi_Y(Y)\big), \quad \phi \text{ isometric embeddings}$$

Where it's used: Compares whole *metric spaces* (shapes, graphs) up to isometry. Manifold learning, 3-D shape matching.

6. Optimal transport & integral probability metrics (IPMs)

An IPM measures distance by the largest gap a test function from a class $\mathcal{F}$ can produce: $\gamma_{\mathcal{F}}(p,q) = \sup_{f\in\mathcal{F}} \big|\int f\,dp - \int f\,dq\big|$.

Wasserstein distance / Earth Mover's Distance (EMD, 1998)

$$W_\alpha(p,q) = \Big(\inf_{\gamma\in\Gamma(p,q)} \int \rho(x,y)^\alpha \, d\gamma(x,y)\Big)^{1/\alpha}$$

Where it's used: "Minimum cost to morph p into q." WGANs, image retrieval, domain adaptation. EMD $= W_1$ (with $\rho = L_1$).

Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD}(p,q) = \sup_{\|f\|_{\mathcal{H}} \le 1} \left| \mathbb{E}_p f - \mathbb{E}_q f \right|$$

Where it's used: IPM over an RKHS ball; kernel two-sample tests, training generative models.

Stein discrepancy. IPM built from a Stein operator; needs the score $\nabla \log p$ but *not* the normalizing constant. Sampler/MCMC diagnostics, Stein variational gradient descent.

Kolmogorov(–Smirnov) distance

$$K(p,q) = \sup_x \left| F_p(x) - F_q(x) \right|$$

Where it's used: Sup-distance between CDFs; classic non-parametric goodness-of-fit test.

Lévy–Prokhorov distance

$$\mathrm{LP}_\rho(p,q) = \inf \left\{ \varepsilon > 0 : \; p(A) \le q(A^\varepsilon) + \varepsilon \;\; \forall A \in \mathcal{B}(\mathcal{X}) \right\}$$

Where it's used: Metrizes weak convergence of probability measures (convergence in distribution).

7. Quantum geometry (density matrices)

Replace probability densities by a density matrix ρ (or $\mathbf{P}, \mathbf{Q}$). Integrals become traces.

Von Neumann entropy (J. von Neumann, 1927)

$$S(\rho) = -k \operatorname{Tr}(\rho \log \rho)$$

Where it's used: Quantum analogue of Shannon entropy; entanglement and quantum information.

Von Neumann (quantum relative) divergence

$$D(\mathbf{P} \| \mathbf{Q}) = \operatorname{Tr}\big(\mathbf{P}(\log \mathbf{P} - \log \mathbf{Q}) - \mathbf{P} + \mathbf{Q}\big)$$

Where it's used: Quantum analogue of KL divergence; distinguishability of quantum states.

Log-Det divergence

$$D(\mathbf{P} \| \mathbf{Q}) = \langle \mathbf{P}, \mathbf{Q}^{-1} \rangle - \log\det(\mathbf{P}\mathbf{Q}^{-1}) - \dim \mathbf{P}$$

Where it's used: Bregman divergence on positive-definite matrices. Covariance comparison, metric learning (ITML).

8. The big picture — how the families nest

- $L_1 \subset L_2 \subset L_\infty$ are all special cases of the **Minkowski** L_k family; **Mahalanobis** is the **Quadratic** distance with $Q = \Sigma^{-1}$.
- **f-divergences** contain KL, reverse-KL, χ^2, Hellinger, total variation and the α-divergence.
- **Bregman divergences** contain squared-Euclidean, KL (discrete), Itakura–Saito, Mahalanobis and Log-Det.
- **KL is the unique divergence lying in *both*** the f-divergence and Bregman families.
- **Bhattacharyya, Chernoff, Rényi** are all read off the affinity $\int p^\alpha q^{1-\alpha} d\mu$.
- **Jensen–Shannon** is a **Burbea–Rao** divergence; **Jeffreys** is symmetrized KL.
- **Wasserstein, MMD, Stein, Kolmogorov** are **IPMs**; **EMD** = Wasserstein-1.
- **Fisher–Rao** is the Riemannian (geodesic) distance; locally it agrees with $\sqrt{2\,\mathrm{KL}}$.
- **Von Neumann** & **Log-Det** are the quantum / matrix counterparts of KL and a Bregman divergence.

Compiled as a clean reorganization of the "Taxonomy of principal distances and divergences" chart by Frank Nielsen. Generators (f, F) and constants follow standard conventions; some sources differ on the argument order of $D(p\|q)$.