# The Many Faces of Information Geometry

*Frank Nielsen*

Information geometry [Ama16, AJLS17, Ama21] aims at unravelling the geometric structures of families of probability distributions and at studying their uses in information sciences. Information sciences is an umbrella term regrouping statistics, information theory, signal processing, machine learning and AI, etc. Information geometry was born independently from econometrician H. Hotelling (1930) and statistician C. R. Rao (1945) from the mathematical curiosity of considering a parametric family of probability distributions, called the statistical model, as a Riemannian manifold equipped with the Fisher metric tensor [Nie20]. Information geometry tackles problems by using the concepts of differential geometry (like curvature) with tensor calculus. In his pioneer work, Rao considered the Riemannian geodesic distance and geodesic balls on the manifold to study classification and hypothesis testing problems in statistics.

Let $(\mathcal{X}, \mathcal{F}, \mu)$ denote a probability space [Kee10] (with sample space $\mathcal{X}$, $\sigma$-algebra $\mathcal{F}$, and finite positive measure $\mu$, usually chosen as the Lebesgue mesure $\mu_L$ or the counting measure $\mu_c$), and consider a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability distributions, all dominated by $\mu$. Let $p_\theta(x) := \frac{dP_\theta(x)}{d\mu}$ denote the Radon-Nikodym derivative, the probability density function of random variable $X \sim p_\theta$. By definition, the Fisher Riemannian metric $g_F$ expressed in the $\theta$-coordinate system is the Fisher information matrix (FIM) of the random variable $X$: $[g_F]_\theta := I_X(\theta)$ with

$$I_X(\theta) := E_{p_\theta}\left[s_\theta(x) s_\theta(x)^\top\right],$$

where $s_\theta(x) := \nabla_\theta \log p_\theta(x)$ is called the score function [Kee10]. The Fisher metric is also referred to as the Shahshahani metric in mathematical biology. Because the FIM is the covariance matrix of the score (since $E_{p_\theta}[s_\theta(x)] = 0$), $I_X(\theta)$ is necessarily positive semidefinite, and positive-definite for regular statistical models [Ama16]. The FIM is covariant under reparameterization: for any smooth invertible mapping $\eta(\theta)$ with invertible Jacobian matrix $\left[\frac{\partial \theta_j}{\partial \eta_j}\right]_{ij}$, we have

$$I_\eta(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]_{ij}^\top \times I_\theta(\theta(\eta)) \times \left[\frac{\partial \theta_i}{\partial \eta_j}\right]_{ij}.$$

*Frank Nielsen is a senior researcher (Fellow) of Sony Computer Science Laboratories, Inc., Tokyo, Japan. His email address is* Frank.Nielsen@acm.org.

*Communicated by* Notices *Associate Editor Richard Levine.*



DOI: https://doi.org/10.1090/noti2403

The Riemannian Fisher length element induced by the Fisher metric

$$\mathrm{d}s_\theta = \sqrt{\mathrm{d}\theta^\top \times [g_F]_\theta \times \mathrm{d}\theta}$$

is invariant under any smooth invertible reparameterization: $\mathrm{d}s_\theta = \mathrm{d}s_\eta$ with $\mathrm{d}s_\eta^2 = \mathrm{d}\eta^\top \times [g_F]_\eta \times \mathrm{d}\eta$. Nowadays, the Riemannian manifold $(\mathcal{P}, g_F)$ is commonly called the Fisher-Rao manifold [Nie20], and its induced Riemannian geodesic length distance $\rho_g(\theta_1, \theta_2)$ is called the Fisher-Rao distance $\rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2}) \coloneqq \rho_{g_F}(\theta_1, \theta_2)$ with

$$\rho_{g_F}(\theta_1, \theta_2) \coloneqq \int_0^1 \mathrm{d}s_{\gamma(t)}\,\mathrm{d}t,$$

where $\gamma(t)$ denotes the Riemannian geodesic [GN14] with the boundary conditions $\gamma(0) = \theta_1$ and $\gamma(1) = \theta_2$. Thus the Fisher-Rao distance used to evaluate the dissimilarities between probability distributions is invariant under reparameterization. For example, the Fisher-Rao distance remains the same whether the family of normal distributions are parameterized by $\lambda = (\mu, \sigma)$, $\lambda' = (\mu, \sigma^2)$, or $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$:

$$\rho_{\mathrm{Rao}}(p_{\lambda_1}, p_{\lambda_2}) = \rho_{\mathrm{Rao}}(p_{\lambda_1'}, p_{\lambda_2'}) = \rho_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2}).$$

This parameterization invariance property of statistical distances highlights the power of modeling geometrically statistical models. The Fisher-Rao manifolds may have different constant sectional curvatures $\kappa$: for example, the curvatures of the Fisher-Rao manifolds of univariate normal distributions, univariate zero-centered multivariate normal distributions, and categorical distributions are $\kappa = -\frac{1}{2} < 0$, $\kappa = 0$, and $\kappa = \frac{1}{4} > 0$, respectively. Information geometry elucidates the role played by curvature in statistics. Since any Riemannian manifold of dimension $D$ can be isometrically embedded in a Euclidean space of dimension at most $D_E = \frac{1}{2}D(D+1)(3D+11)$ using Nash's embedding theorem, we may visualize the Fisher-Rao manifold $(\mathcal{P}, g_F)$ as a $D$-dimensional surface of $\mathbb{R}^{D_E}$. This extrinsic view of geometry is helpful to intuitively grasp the notion of tangent planes $T_{p_\theta}$ and tangent vectors $v \in T_{p_\theta}$ at any $p_\theta \in \mathcal{P}$ and allows one to visualize geodesics on surfaces. However, let us point out that differential geometry defines intrinsically these notions [GN14].

Using the Fisher metric can be justified from several theoretical viewpoints [Ama16]:

- First, the FIM occurs when locally approximating the Kullback-Leibler (KL) divergence [Kee10]. In statistics, estimating densities using the Maximum Likelihood Estimator (MLE) or the maximum entropy principle under moment constraints (MaxEnt) can be interpreted as KL divergence minimization problems [Kee10] (to be detailed below). The KL divergence between densities $p_{\theta_1}$ and $p_{\theta_2}$ is defined by:

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] \coloneqq \int_{\mathcal{X}} p_{\theta_1}(x) \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \mathrm{d}\mu(x).$$

  The delimiter ":" indicates that the divergence is oriented: $D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] \neq D_{\mathrm{KL}}[p_{\theta_2} : p_{\theta_1}]$. The KL divergence can be expressed as

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = h^\times[p_{\theta_1} : p_{\theta_2}] - h[p_{\theta_1}],$$

  where $h^\times[p_{\theta_1} : p_{\theta_2}]$ denotes the cross-entropy $h^\times[p_{\theta_1} : p_{\theta_2}] \coloneqq -\int p_{\theta_1} \log p_{\theta_2} \mathrm{d}\mu(x)$ and $h[p_{\theta_1}] \coloneqq -\int p_{\theta_1} \log p_{\theta_1} \mathrm{d}\mu(x)$ is Shannon entropy. Hence, the KL divergence is also called relative entropy in information theory. The second-order Taylor approximation of the KL divergence yields

$$D_{\mathrm{KL}}[p_\theta : p_{\theta+d\theta}] = \frac{1}{2} d\theta^\top \times I_\theta(\theta) \times d\theta \approx \frac{1}{2}\mathrm{d}s_\theta^2.$$

  More generally, the FIM is used in the local approximations of $f$-divergences [Ama16]:

$$I_f[p_\theta : p_{\theta+d\theta}] = \frac{1}{2} f''(1)\, d\theta^\top \times I_\theta(\theta) \times d\theta,$$

  where

$$I_f[p_{\theta_1} : p_{\theta_2}] \coloneqq \int_{\mathcal{X}} p_{\theta_1}(x) f\left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}\right) \mathrm{d}\mu(x)$$

  for a convex function $f(u)$ satisfying $f(1) = 0$, and strictly convex at 1. The KL divergence is an $f$-divergence obtained for $f(u) = -\log(u)$ with $f''(u) = 1$. The $f$-divergences are said to be separable because they can be written as integrals of scalar divergences:

$$I_f[p : q] = \int_{\mathcal{X}} i_f[p(x) : q(x)] \mathrm{d}\mu(x)$$

  with $i_f[a : b] \coloneqq af(b/a)$. The $f$-divergences enjoy the following monotonicity property:

$$I_f[p_Y : q_Y] \leq I_f[p_X : q_X],$$

  where $p_Y$ and $q_Y$ are the densities induced by a Markov kernel from measurable space $(\mathcal{X}, \mathcal{F})$ to measurable space $(\mathcal{Y}, \mathcal{G})$ [Ama21]. To give a concrete example, consider the $f$-divergences between two (normalized) histograms $p = (p_1, \ldots, p_{2n})$ and $q = (q_1, \ldots, q_{2n})$ with $2n$ bins representing multinomial probability laws. Then $I_f[p' : q'] \leq I_f[p : q]$, where $p' = (p_1', \ldots, p_n')$ and $q' = (q_1, \ldots, q_n')$ with $p_i' = p_{2(i-1)+1} + p_{2i}$ and $q_i' = q_{2(i-1)+1} + q_{2i}$ reduced histograms obtained by merging consecutive bins (a very special deterministic Markov kernel from measurable space $([2n], 2^{[2n]})$ to measurable space $([n], 2^{[n]})$

where $[s] := \{1, \ldots, s\}$). The only separable statistical divergences satisfying this monotonicity property are $f$-divergences when $n > 2$ [Ama16].

- A sufficient statistic [Kee10] $Y = T(X)$ for the parameter $\theta$ of a random variable $X \sim p_\theta$ is such that the conditional probability of $X$ given $Y = T(X)$ does not depend on $\theta$. That is, all statistical information concerning parameter $\theta$ is contained in $Y = T(X)$. For example, $T(X) = (T_1(X), T_2(X))$ with $T_1(X) = \sum_{i=1}^n X_i$ and $T_2(X) = \sum_{i=1}^n X_i^2$ is a sufficient statistic for the parameter $\theta = (\mu, \sigma)$ of a random vector $(X_1, \ldots, X_n)$ of $n$ independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim p_\theta$ following a normal distribution $p_{\mu,\sigma}$. In general, we have $I_Y(\theta) \leq I_X(\theta)$ with equality if and only if (iff) $Y$ is a sufficient statistic [Kee10]. Let us observe that for $n$ i.i.d. random variables $X_i \sim p_\theta$, we have $I_{X_1,\ldots,X_n}(\theta) = nI_X(\theta)$ with $X \sim p_\theta$. Sufficiency also characterizes the equality in the monotonicity inequality of $f$-divergences: $I_f[p_Y : q_Y] \leq I_f[p_X : q_X]$ with equality iff $Y = T(X)$ is a sufficient statistic.
- Let $\hat{\theta}_n$ be an unbiased estimator of $\theta$ of an i.i.d. random vector $(X_1, \ldots, X_n) \sim p_\theta$. Then the following Cramér-Rao lower bound (CRLB) on the variance of $\hat{\theta}_n$ holds:

$$\mathrm{Var}[\hat{\theta}_n] = E\left[\left(\hat{\theta}_n - E[\hat{\theta}_n]\right)^2\right] \succeq \frac{1}{n} I_\theta^{-1}(\theta),$$

where $A \succeq B$ iff matrix $A - B$ is positive semidefinite. The notation $\succeq$ indicates the comparison with respect to (w.r.t.) the Loewner partial ordering of positive semidefinite matrices. Thus the inverse of the FIM provides a lower bound on the accuracy of any unbiased estimator. An estimator is said to be Fisher efficient when its variance asymptotically matches the CRLB when $n \to \infty$.

The Fisher metric is the only invariant metric under Markovian morphisms of statistical models [Ama16]. However, let us point out that there are infinitely many counterparts of the FIM in quantum information geometry, and that other alternative Riemannian information metrics can be explored (e.g., the Wasserstein information metric [Li21]).

In statistics, two special types of statistical models, called the exponential families and the mixture families, are often handled:

- An exponential family [Kee10] is a set of parametric densities $\mathcal{E} := \{p_\theta(x)\mathrm{d}\mu\}$ such that

$$p_\theta(x) := \exp\left(\sum_{i=1}^D t_i(x)\theta_i - F(\theta)\right),$$

where the $(t_1(x), \ldots, t_D(x))$ form the minimal sufficient statistic. Function $F(\theta)$ is used to normalize the densities:

$$F(\theta) = \log\left(\int_{\mathcal{X}} \exp\left(\sum_{i=1}^D t_i(x)\theta_i\right)\mathrm{d}\mu(x)\right),$$

and called the cumulant function (or log-partition in statistical physics). $F(\theta)$ is strictly convex for (full regular) exponential families [Kee10]. For example, the family of $d$-variate normal distributions is an exponential family of order $D = \frac{D(D+3)}{2}$ w.r.t. the Lebesgue measure $\mu_L$, and the family of Poisson distributions is a discrete exponential family of order $D = 1$ w.r.t. the counting measure $\mu_c$. Exponential families have all finite-dimensional minimal sufficient statistics.
- A mixture family [Nie20] is a set of parametric densities $\mathcal{M} := \{m_\theta(x)\mathrm{d}\mu\}$ such that $m_\theta(x) = \sum_{i=1}^D \theta_i p_i(x) + (1 - \sum_{i=1}^D \theta_i)p_0(x)$, where functions $1, p_0(x), p_1(x), \ldots, p_D(x)$ are linearly independent functions. Statistical mixtures such as Gaussian mixture models with prescribed component densities are examples of mixture families. Mixture families are closed under convex combinations. It can be shown that the negentropy of mixture densities is a strictly convex function [Nie20]: $F(\theta) = -h[m_\theta]$.

For these two types of statistical models, the Fisher metric is a Hessian metric [Shi07] since the FIM is the Hessian of some strictly convex potential function $F(\theta)$: $I(\theta) = \nabla^2 F(\theta)$. This is easily checked for exponential families as the FIM can be written under mild regularity conditions [Ama16] as $I_X(\theta) = -E_{p_\theta}[\nabla^2 \log p_\theta(x)]$.

In general, calculating in closed-form the Fisher-Rao distances may be difficult since it requires to solve the Riemannian geodesic equation with boundary conditions, and to integrate the length elements along geodesics. For example, although the Fisher-Rao distance between univariate normal distributions is available in closed-form, we do not have a closed-form formula for the Fisher-Rao distance between multivariate normals [Nie20]. Thus in practice, the Fisher-Rao between multivariate normals is numerically approximated. We shall now explain that the Fisher-Rao manifolds with Hessian metrics carry another beautiful geometric dual structure which is well suited for computation in applications: namely, these exponential/mixture families can be modeled as Hessian manifolds [Shi07], and are commonly called dually flat spaces in information geometry [Ama21].

In the second half of the 20th century, information geometry gained a momentum with the pathbreaking work of N. Chentsov. Chentsov shared statistician A. Wald's viewpoint that all problems in statistics can be

viewed as decision problems, and therefore investigated a theoretical framework for characterizing optimal decision rules in statistics using category theory and Markovian morphisms [Čen82]. A family of probability distributions should be invariant both under smooth one-to-one transformations of its parameter $\theta$ and under transformations of the corresponding random variables by sufficient statistics. This precisely defines the statistical invariance. Chentsov's breakthrough consisted in separating the metric tensor $g$ (used to measure angles between vectors and lengths of vectors in a tangent plane) from its induced Levi-Civita connection ${}^{g}\nabla$ used by default on a Riemannian manifold for obtaining (locally) length minimizing geodesics. This novel insight allowed Chentsov to model statistical models as differentiable manifolds equipped with affine connections $\nabla$ more general than the Levi-Civita connection of Fisher-Rao manifolds. More precisely, Chentsov discovered the existence of a unique totally symmetric third-order tensor fulfilling the statistical invariance which is nowadays called the Amari-Chentsov tensor or skewness tensor, and used that tensor to build invariant affine connections. A. Kolmogorov called Chentsov's field of research "geometrostatistics" in Russian (translated as "geometrical statistics" in the english monograph [Čen82]).

In short, an affine connection $\nabla$ [GN14] defines the following:

- A way to differentiate a vector field $X$ on a manifold (or more generally a tensor field $T$) by another vector field $Y$: namely, the covariant derivatives denoted by $\nabla_Y X$ (or $\nabla_Y T$).
- The parallel transport $\prod_{c(t)}^{\nabla} v$ for a tangent vector $v \in T_{c(0)}$ along any smooth curve $c(t)$. An affine connection allows to parallel transport vectors of different tangent planes onto a common tangent plane in order to measure their subtended angles using the metric tensor in that common tangent plane.
- $\nabla$-geodesics $\gamma$ defined as autoparallel curves: $\nabla_{\dot\gamma}\dot\gamma = 0$. Geodesics $\gamma^{\nabla}(t)$ are calculated by solving the second-order non-linear ordinary differential equation (ODE):

$$\ddot\theta^i + \sum_{j,k=1}^{n} \Gamma^i_{jk}\dot\theta^j\dot\theta^k = 0, \quad i \in \{1,\dots,D\},$$

  where $\theta = (\theta^1,\dots,\theta^D)$, $\dot\theta^l := \frac{d}{dt}\theta^l$, and $\Gamma^i_{jk}$ are the Christoffel symbols [GN14] ($D^3$ smooth functions) defining the affine connection. In physics, geodesics represent free particle trajectories.

The curvature tensor $R^{\nabla}$ and the torsion tensor $T^{\nabla}$ of a manifold $(M, g, \nabla)$ are induced by the chosen connection $\nabla$ [GN14]. The fundamental theorem of Riemannian

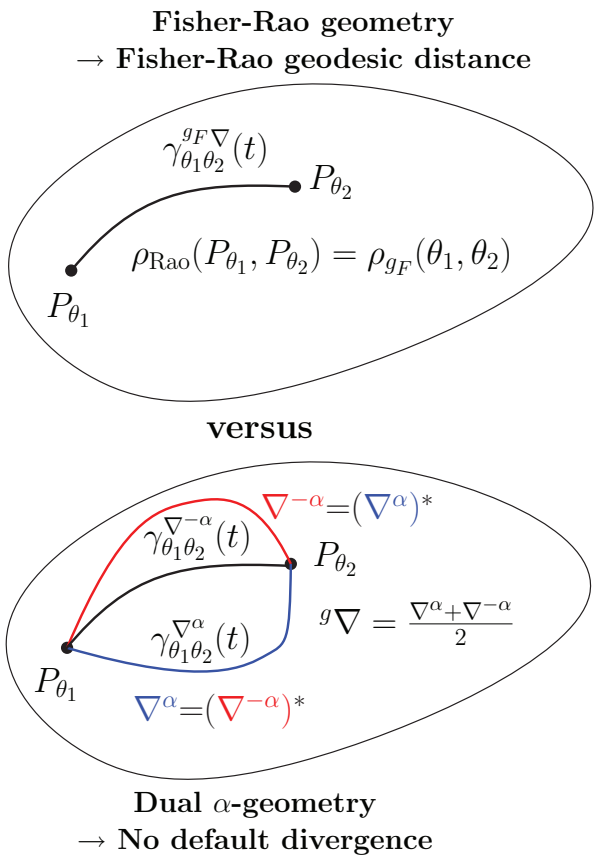



**Figure 1.** Fisher-Rao geometry vs. dual $\alpha$-geometry.

geometry [GN14] states that there exists a unique torsion-free affine connection which preserves the metric, meaning that for any two vectors $v_1$ and $v_2$ of the tangent plane $T_p$, and a smooth curve $c(t)$ with $c(0) = p$, we have for any $t$: $g(v_1, v_2) = g\left(\prod_{c(t)}^{\nabla} v_1, \prod_{c(t)}^{\nabla} v_2\right)$. This unique torsion-free affine connection is called the Levi-Civita metric connection. Historically, affine connections were studied by É. Cartan in the 1920s, and used in the Einstein-Cartan theory of gravity. Chentsov considered regular exponential families in his monograph, and by considering their invariance, discovered the so-called exponential connection $\nabla^e$.

The field of information geometry was shaped by S.-i. Amari who dreamt of a mathematical theory of neuroscience. Amari pioneered the dualistic statistical structures of information geometry: that is, Amari showed that given any torsion-free affine connection $\nabla$, there exists a dual torsion-free affine connection $\nabla^*$ such that the mid-connection $\frac{\nabla+\nabla^*}{2}$ corresponds to the Levi-Civita connection. This duality ensures that the primal and dual parallel transports are metric compatible: $g(v_1, v_2) = g\left(\prod_{c(t)}^{\nabla} v_1, \prod_{c(t)}^{\nabla^*} v_2\right)$. Notice that the lengths of dually parallel-transported vectors by $\nabla$ and $\nabla^*$ may

vary along $c(t)$ but their inner product is kept constant. For parametric statistical models $\mathcal{P} = \{p_\theta\}$, Amari reported the $\alpha$-geometry (for $\alpha \in \mathbb{R}$) which is a manifold equipped with the Fisher metric and a pair of dual connections $(\nabla^{-\alpha}, \nabla^\alpha)$ coupled to the Fisher metric $g_F$: $\frac{\nabla^{-\alpha}+\nabla^\alpha}{2} = \nabla^0 = {}^{g_F}\nabla$. Amari's $\alpha$-connections $\nabla^\alpha$ are defined by their Christoffel symbols $\Gamma^\alpha_{ki,j}$:

$$\Gamma^\alpha_{ki,j}(\theta) = E_{p_\theta}\left[\left(\partial_k\partial_i l_\theta(x) + \frac{1-\alpha}{2}\partial_k l_\theta(x)\partial_i l_\theta(x)\right)\partial_j l_\theta(x)\right],$$

where $l_\theta(x) := \log p_\theta(x)$ is the log-likelihood function and $\partial_s := \frac{\partial}{\partial\theta_s}$. Chentsov's exponential connection $\nabla^e$ corresponds to Amari's $\nabla^1$ connection. The $e$-connection was also studied by Efron who defined geometrically the notion of statistical curvature to study the higher-order asymptotic theory of statistical estimators in a landmark paper [Efr75] which has been recognized as one of the first successful applications of differential geometry to statistical inference. The dual connection of $\nabla^e$ is $\nabla^m := (\nabla^e)^* = \nabla^{-1}$, and called the mixture connection. This connection was proposed by P. Dawid, a discussant of Efron's paper [Efr75]. Thus the Fisher-Rao geometry can be interpreted as the 0-geometry enhanced with the Fisher-Rao geodesic distance (Figure 1). The Riemannian geodesics are ${}^g\nabla$-autoparallel and have the property to locally minimize the geodesic lengths [GN14]. In general, the $\alpha$-geometry is not associated with any statistical divergence when $\alpha \neq 0$. But the $\alpha$-geometry may be recovered from the divergence geometry of invariant $f$-divergences on the probability simplex [Egu83] (to be detailed below).

A connection $\nabla$ is said to be flat [GN14, Shi07] when it has zero torsion and when there exists a local coordinate system $\theta$ such that the Christoffel symbols $\Gamma^k_{ij}$ defining $\nabla$ expressed in that coordinate system vanish: $\Gamma^k_{ij}(\theta) = 0$. The coordinate system $\theta$ is called a $\nabla$-affine coordinate system. In general the parallel transport of $v \in T_p$ to $T_q$ is curve dependent: $\prod^\nabla_{c_1(t)} v \neq \prod^\nabla_{c_2(t)} v$ for smooth curves $c_1$ and $c_2$ with endpoints $c_1(0) = c_2(0) = p$ and $c_1(1) = c_2(1) = q$. One can visualize locally the presence of curvature of a connection or not at a point $p$ by considering the parallel transport of a vector $v$ along a closed infinitesimal loop $l$ encircling $p$ (with $l(0) = l(1)$): if there is an angle deficiency between $v$ and $\prod^\nabla_{l(1)} v$, then the manifold has non-zero curvature at $p$ [Nie20]. However, the parallel transport is independent of the curves linking the point $p$ to the point $q$ for flat connections. It is a fundamental result of information geometry that if $(M, \mathrm{g}, \nabla)$ is flat, then so is $(M, g, \nabla^*)$ with $\nabla^* = 2^g\nabla - \nabla$. We get the so-called dually flat spaces of information geometry [Ama16] $(M, g, \nabla, \nabla^*)$ which are special Hessian manifolds [Shi07] admitting a single chart atlas. Notice that in a dually flat space, the Levi-Civita connection is usually not flat.

In information theory, a statistical divergence like the KL divergence is loosely speaking a potentially asymmetric dissimilarity measure between probability distributions which may fail the triangle inequality of metric distances. In information geometry, a divergence (historically called a contrast function [Egu83]) is a smooth dissimilarity measure $D(\theta : \theta')$ between parameters $\theta$ and $\theta'$ that satisfies the following conditions:

1. $D(\theta : \theta') \geq 0$ for all $\theta, \theta'$ with equality iff $\theta = \theta'$.
2. $\partial_i\, D(\theta : \theta')|_{\theta'=\theta} = \partial'_j\, D(\theta : \theta')|_{\theta'=\theta} = 0$ for all $i, j$, where $\partial_l := \frac{\partial}{\partial\theta_l}$ and $\partial'_l := \frac{\partial}{\partial\theta'_l}$.
3. $-\left[\partial_i\partial'_j\, D(\theta : \theta')|_{\theta'=\theta}\right]_{ij}$ is a positive-definite matrix.

The (parameter) divergence $D(\theta : \theta')$ can also be interpreted as a function on the manifold defined by the single chart equipped with the $\theta$-coordinate system. Eguchi [Egu83] reported a method to build a dualistic structure $(M, g, \nabla, \nabla^*)$ from any divergence $D(\cdot : \cdot)$ as follows:

$$\begin{aligned} g_{ij}(\theta) &= -\partial_i\partial'_j\, D(\theta : \theta')|_{\theta'=\theta}, \\ \Gamma_{ij,k}(\theta) &= -\partial_i\partial_j\partial'_k\, D(\theta : \theta')|_{\theta'=\theta}, \\ \Gamma^*_{ij,k}(\theta) &= -\partial_k\partial'_i\partial'_j\, D(\theta : \theta')|_{\theta'=\theta}. \end{aligned}$$

It can be shown that the connections $\nabla$ and $\nabla^*$ induced respectively by $\Gamma_{ij,k}$ and $\Gamma^*_{ij,k}$ are torsion-free and dual. This geometry is called the divergence geometry of $D$ [Ama16]. Let $D^*(\theta : \theta') := D(\theta' : \theta)$ denote the dual or reverse divergence, and $(M, {}^D g, {}^D\nabla, {}^D\nabla^*)$ the information-geometric space induced by $D$. Then we have ${}^{D^*}\nabla = {}^D\nabla^*$ and ${}^{D^*}\nabla^* = {}^D\nabla$. Thus symmetric divergences $D(\theta : \theta') = D(\theta' : \theta)$ yield self-dual connections coinciding with the Levi-Civita connection. Many different divergences may yield the same divergence geometry. The divergence geometry of $f$-divergences on the $D$-dimensional probability simplex corresponds to Amari's $\alpha$-geometry for $\alpha = 3 + 2\frac{f'''(1)}{f''(1)}$, and the divergence geometry of $D_{\mathrm{Rao}}(\theta_1 : \theta_2) := \frac{1}{2}\rho^2_{\mathrm{Rao}}(p_{\theta_1}, p_{\theta_2})$ yields the 0-geometry.

In a dually flat space, we can build a canonically Fenchel-Young (non-metric) divergence $A(\theta_1 : \eta_2)$:

$$A(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top\eta_2,$$

where $F^*(\eta)$ denotes the convex conjugate obtained by the Legende-Fenchel transform:

$$F^*(\eta) = \sup_\theta\{\theta^\top\eta - F(\theta)\}.$$

Legende-Fenchel transform yields a dual coordinate system $\eta = \nabla F(\theta)$, and the Fenchel-Young inequality $F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top\eta_2$ ensures that $A(\theta_1 : \eta_2) \geq 0$ with equality iff $\eta_2 = \nabla F(\theta_1)$. This divergence is shown to be equivalent to a Bregman divergence [Ama16]: $A(\theta_1 : \eta_2) = B_F(\theta_1 : \theta_2)$

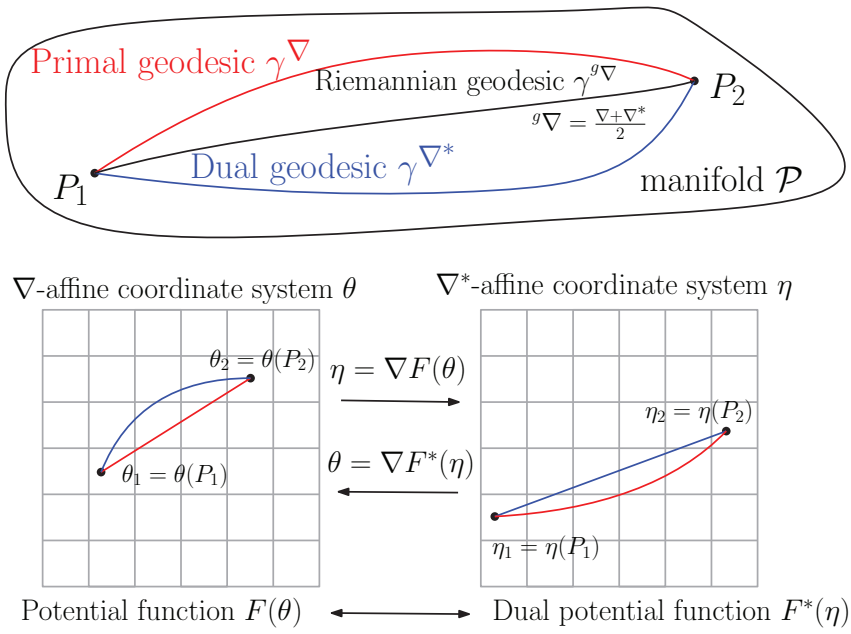



**Figure 2.** Dually flat space $(M, g, \nabla, \nabla^*)$ with $\nabla$-affine coordinate system $\theta$ and dual $\nabla^*$-affine coordinate system $\eta$. Primal geodesics $\gamma^{\nabla}$ and dual geodesics $\gamma^{\nabla^*}$ are linear when plotted in the $\theta$-coordinate system and $\eta$-coordinate system, respectively.

with $\theta_2 = \nabla F^*(\eta_2)$, where

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2)$$

is the Bregman divergence $B_F(\theta_1 : \theta_2)$ between parameters $\theta_1$ and $\theta_2$ induced by a smooth strictly convex function $F(\theta)$ and $\nabla F(\theta) := \left[\frac{\partial F(\theta)}{\partial \theta_i}\right]_i^\top$ denotes the gradient of $F(\theta)$. Bregman divergences are widely used in machine learning and originated from mathematical programming. Reciprocally, given a Bregman divergence, we can build a dually flat space (the Bregman divergence geometry) with Hessian metric [Shi07] $[g_{ij}(\theta)]_{ij} = \nabla^2 F(\theta)$ (positive-definite because $F$ is strictly convex) from its divergence geometry. The dual affine connections are $\nabla$ and $\nabla^*$ with Christoffel symbols $\Gamma_{ijk}(\theta) = 0$ and $\Gamma^{*ijk}(\eta) = 0$, respectively. It follows that in the $\theta$-coordinate system, the primal geodesics $\gamma^{\nabla}$ are linear since the $\nabla$-geodesic ODE simplifies to $\ddot{\theta}^i = 0$, and in the dual $\eta$-coordinate system, the dual geodesics are linear since the $\nabla^*$-geodesic ODE simplifies to $\ddot{\eta}^i = 0$ (Figure 2). The $\eta$-coordinate system is said to be $\nabla^*$-affine. We have $[g_{ij}(\theta)]_{ij} = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$ and the dual Riemannian metric $[g^{ij}(\eta)]_{ij} = \nabla_\eta \nabla_\eta F^*(\eta) = \nabla_\eta \theta$. It follows that $[g_{ij}(\theta)]_{ij}\,[g^{ij}(\eta)]_{ij} = I_{D\times D}$, the identity matrix. The Bregman divergence construction from a dually flat space is defined up to affine dual coordinate transformations $\theta' = A\theta + b$ and $\eta' = A^{-1}\eta + c$.

Amari's $\pm 1$-geometry for the exponential and mixture families yields dually flat spaces. Their corresponding Bregman divergences yield the following statistical divergences:

- For an exponential family [Kee10] with density $p_\theta(x) = \exp(\sum_{i=1}^D t_i(x)\theta_i - F(\theta))$, the Legendre-Fenchel conjugate function of the cumulant function $F(\theta)$ corresponds to Shannon negentropy, $F^*(\eta) = -h[P_\theta]$, and the Bregman divergence $B_F(\theta_1 : \theta_2)$ yields the reverse Kullback-Leibler divergence $D^*_{\mathrm{KL}}$ (or reverse relative entropy):

$$B_F(\theta_1 : \theta_2) = D^*_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = D_{\mathrm{KL}}[p_{\theta_2} : p_{\theta_1}].$$

- For a mixture family [Ama16, Nie20] $\mathcal{M}$ with density $m_\theta = \sum_{i=1}^D \theta_i p_i(x) + (1 - \sum_{i=1}^D \theta_i) p_0(x)$, the Bregman generator $F(\theta) = -h[m_\theta]$ is strictly convex with $\Theta = \Delta_D^\circ$, the open $D$-dimensional probability simplex. The canonical Bregman divergence amounts to calculating the Kullback-Leibler divergence [Nie20]: $B_F(\theta_1 : \theta_2) = D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}]$.

In a dually flat space $(M, g, \nabla, \nabla^*)$, a generalized Pythagorean theorem holds: Let $P$, $Q$, $R$ be three points. A primal geodesic $\gamma_{PQ}$ intersects a dual geodesic $\gamma^*_{QR}$ orthogonally w.r.t. to the metric $g$ at point $Q$, written as $\gamma_{PQ} \perp_g \gamma^*_{QR}$, iff

$$(\theta(P) - \theta(Q))^\top \times (\eta(R) - \eta(Q)) = 0.$$

In that case, the following Pythagorean equality holds:

$$B_F(\theta(P) : \theta(Q)) + B_F(\theta(Q) : \theta(R)) = B_F(\theta(P) : \theta(R)).$$

When $F(\theta) = \frac{1}{2}\theta^\top\theta$, we recover the usual Euclidean geometry (a self-dual flat space) with $\theta = \eta$ (since $\eta = \nabla F(\theta) = \theta$ and $\theta = \nabla F^*(\eta) = \eta$ with $F^*(\eta) = \frac{1}{2}\eta^\top\eta$), and we have

$B_F(\theta_1 : \theta_2) = \frac{1}{2}\|\theta_1 - \theta_2\|_2^2$, where $\|\cdot\|_2$ denotes the Euclidean norm. The canonical $\nabla$-divergence is $D_\nabla(P : Q) = B_F(\theta(P) : \theta(Q))$ and the dual $\nabla^*$-divergence is

$$\begin{aligned} D_{\nabla^*}(P : Q) &= B_{F^*}(\eta(P) : \eta(Q)) \\ &= B_F(\theta(Q) : \theta(P)) = D_\nabla(Q : P). \end{aligned}$$

A $\theta$-flat is a submanifold $M' \subset M$ such that $\theta(M') := \{\theta(P) \ : \ P \in M'\}$ is an affine subspace of $\Theta$ ($M'$ is a $\nabla$-autoparallel submanifold). Similarly, an $\eta$-flat is a submanifold $M''$ such that $\eta(M'') := \{\eta(P) \ : \ P \in M''\}$ is an affine subspace of $H := \{\nabla F(\theta) \ : \ \theta \in \theta\}$. Define the $\nabla$-projection of a point $P \in M$ onto a submanifold $M'$ as

$$\mathrm{Proj}^\nabla_{M'}(P) := \{Q \in M' \ : \ \gamma^\nabla_{PQ} \perp_g M'\}.$$

Then the $\nabla$-projection of $P$ is guaranteed to be a unique point when $M'$ is an $\eta$-flat. Moreover, we have $\mathrm{Proj}^\nabla_{M'}(P) = \arg\min_{Q\in M'} D_\nabla(Q : P)$ in a dually flat space.

These information projections can be used in statistical inference as follows. Consider the probability space $(\mathcal{X}, 2^\mathcal{X}, \mu_c)$ with finite discrete sample space $\mathcal{X} = \{1, \dots, m\}$ and $\mu_c$ the counting measure. The categorical distributions form both an exponential family and a mixture family [Ama16]. A categorical probability mass function can be viewed as a point lying on the $(m-1)$-dimensional open standard simplex $\Delta_{m-1}$.

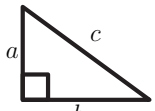
- The Maximum Likelihood Estimator (MLE) of $n$ i.i.d. observations $x_1, \dots, x_n$ sampled from an exponential family density $p_\theta \in \mathcal{E}$ is $\hat\eta = \widehat{\nabla F(\theta)} = \frac{1}{n}\sum_{i=1}^n t(x_i)$. Since the MLE is equivariant [Kee10], we have $\widehat{\nabla F(\theta)} = \nabla F(\hat\theta)$, and it follows that $\hat\theta = (\nabla F)^*(\hat\eta)$ since $(\nabla F)^{-1} = \nabla F^*$. The MLE can be interpreted as a divergence minimization problem: $\min_{\theta\in\Theta} D_{\mathrm{KL}}[p_e : p_\theta]$, where $p_e(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(x)$ is the empirical distribution with the $\delta_{x_i}$'s denoting the Dirac distributions $\delta_{x_i}(x) = 1$ iff $x = x_i$, and 0 otherwise. The MLE can be geometrically interpreted as an $m$-projection (i.e., with respect to $\nabla^m$) of $p_e$ onto the $e$-flat exponential family: $p_{\hat\theta} = \mathrm{Proj}^{\nabla^m}_{\mathcal{E}}(p_e)$. Thus the MLE $\hat\theta$ is unique since $\mathcal{E}$ is $\theta$-flat. This result holds more generally for estimations on curved exponential families $\mathcal{C} = \{p_{\theta(c)}\} \subset \mathcal{E}$ [Ama16]: for example, the family of normal distributions $p_{\mu,1+\mu^2}$ with $\theta(\mu) = (\mu, 1+\mu^2)$ is a 1D curved exponential family. By viewing the MLE as a KL divergence minimization problem, we may consider other divergence-based estimators. A divergence $D$ yields a $D$-estimator by asking to solve the minimization problem $\min_{\theta\in\Theta} D[p_e : p_\theta]$. The MLE is the $D_{\mathrm{KL}}$-estimator. Then we study the properties of various $D$-estimators. For example, the $D_\gamma$-estimator induced by the $\gamma$-divergence $D_\gamma$ for $\gamma > 0$ is proven to be robust to noise contamination [Ama16] but not the MLE which is based on the KL divergence. The $\gamma$-divergences tend to the KL divergence in the limit $\gamma \to 0$.
- The Maximum Entropy (MaxEnt) principle of E. Jaynes [Kee10] asks for the probability density $p(x)$ which maximizes the Shannon entropy under $D$ moment constraints $E_p[t_1(X)] = m_1$, …, $E_p[t_D(X)] = m_D$, i.e., $E_p[t(x)] = m$ with $t(x) = (t_1(x), \dots, t_D(x))$ and $m = (m_1, \dots, m_D)$. It can be shown that the MaxEnt distribution $p^*$ is a density belonging to an exponential family $\mathcal{E} = \{p_\theta\}$ with sufficient statistics $t(x)$. Namely, we have $p^* = p_{\theta^*_{\mathrm{ME}}}$ with $\eta^*_{\mathrm{ME}} := \nabla F(\theta^*_{\mathrm{ME}}) = m$. The MaxEnt problem can be rewritten as the following minimization problem: $\min_{p\in\mathcal{M}} D_{\mathrm{KL}}[p : u] = \min_p D^*_{\mathrm{KL}}[u : p]$, where $u$ denotes the uniform distribution on the probability simplex $\Delta_{m-1}$, and $\mathcal{M} := \{p \in \Delta_{m-1} \ : \ E_p[t(x)] = m\}$ is an $m$-flat defined by the moment constraints. By introducing any other prior density $h(x)$, we can thus generalize MaxEnt by the following minimization problem: $\min_{p\in\mathcal{M}} D_{\mathrm{KL}}[p : h]$ under the moment constraint $E_p[t(x)] = m$. The MaxEnt solution $p^*$ belongs to an exponential family $\mathcal{E} := \{p_\theta(x) = \exp(\sum_i t_i(x)\theta_i - F(\theta))h(x)\}$, and we have $p^* = p_{\theta^*_{\mathrm{ME}}}$ such that $\eta^*_{\mathrm{ME}} := \nabla F(\theta^*_{\mathrm{ME}}) = m$. We interpret the MaxEnt distribution $p_{\theta^*_{\mathrm{ME}}}$ as the unique $e$-projection point (with respect to $\nabla^e$) of $h$ onto $\mathcal{M}$ w.r.t. $\nabla^e$: $p^* = \mathrm{Proj}^{\nabla^e}_{\mathcal{M}}(h)$.

Wong [Won18] recently generalized the Legendre-Fenchel transformation used in dually flat spaces, and obtained a new kind of Pythagorean theorem expressed w.r.t. Rényi divergences.

Finally, let us mention that instead of using the invariant $f$-divergences of information geometry, we can use the theory of optimal transport [PC19] to measure the distance between any two probability measures. Optimal transport requires defining a ground distance between elements of the sample space to model the elementary cost of mass transportation, and measures the deviation between two probability measures by forward pushing one measure to another by a transportation plan. Although the optimal transport problems between discrete probability measures encountered in practice (i.e., finite weighted point sets) are computationally costly to solve (amount to solve linear programs), fast entropic-regularized methods [PC19] and various heuristics like the sliced Wasserstein distances have contributed to its huge success in machine learning and computer vision. Optimal transport does not require the probability measures to have coinciding supports, and can even measure the distance between a discrete measure and

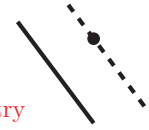
# Genesis of the Dual Structure of Information Geometry

Pythagoras of Samos
(c. 570-495 BC)
Pythagoras' theorem




$c^2 = a^2 + b^2$

Euclid
(ca 365-300 BC)
Elements, math. proof
Playfair axiom, Euclidean geometry

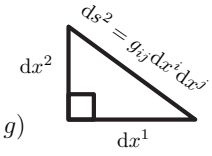


Nikolai Ivanovich Lobachevsky
(1792-1856)
Hyperbolic geometry
($\infty$-many lines passing through a point and // to another line)

Christian Felix Klein
(1849-1925)
Projective geometry & symmetry group
Erlangen program

Johann C. F. Gauss
(1777-1855)
differential geometry
of surfaces
Theorema Egregium

Georg F. B. Riemann
(1826–1866)
metric tensor (1854)
$g = g_{ij}\mathrm{d}\theta_i \otimes \mathrm{d}\theta_j$
Riemannian manifold $(M, g)$




Élie Joseph Cartan
(1869-1951)
affine connections
differential forms $\omega$

Sir Ronald Aylmer Fisher
(1890-1962)
Mathematical statistics
Fisher information, MLE
$I(\theta) = E_{p_\theta}\left[(\nabla_\theta \log p_\theta)(\nabla_\theta \log p_\theta)^\top\right]$

Sir Harold Jeffreys
(1891-1989)
Jeffreys prior $\propto \sqrt{|g|}$
$J$-divergence

Alexander P. Norden
(1904-1993)
conjugate connections wrt $g$
Affinely connected spaces

Harold Hotelling
(1895-1973)
Econometrician
Fisher metric
(1930)

Calyampudi Radhakrishna Rao
(1920-)
Fisher-Rao distance
Cramér-Rao lower bound
(1945)

Prasanta C. Mahalanobis
(1893-1972)
Distances in statistics
Mahalanobis distance
statistical field (1936)

Nikolai Nikolaevich Chentsov
(1930-1992)
statistical invariance
geometrostatistics
Category theory, connections

Solomon Kullback
(1907-1994)
Richard A. Leibler
(1914-2003)
KL divergence
$D_{\mathrm{KL}}[p : q] = \int p \log \frac{p}{q}\mathrm{d}\mu$

Imre Csiszár
(1938-)
information projections
$f$-divergences
$I_f[p : q] = \int pf(\frac{q}{p})\mathrm{d}\mu$

Claude Elwood Shannon
(1916-2001)
Information theory
Entropy:
$h(p) = -\int p \log p \mathrm{d}\mu$

Ole E. Barndorff-Nielsen
(1935-)
Exponential families
observed information geometry

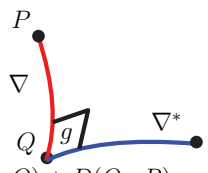
Steffen Lauritzen
(1947-)
statistical manifold $(M, g, C)$

Bradley Efron
(1938-)
statistical curvature
$E$-connection

Jean-Louis Koszul
(1921-2018)
Hirohiko Shima
homogeneous bounded domains
Hessian manifolds

Ernest Borisovich Vinberg
(1937-2020)
characteristic functions
on homogeneous cones

Shun-ichi Amari
(1936-)
Information geometry
dualistic structure$(M, g, \nabla, \nabla^*)$:
$Zg(X, Y) = g(\nabla_Z X, Y) + g\left(X, \nabla^*_Z Y\right)$
dual $\pm\alpha$-connections
$(M, g_F, \nabla^{-\alpha}, \nabla^\alpha)$




$D(P : Q) + D(Q : R) = D(P : R)$
Generalized Pythagoras' theorem
in dually flat space $(M, g, \nabla, \nabla^*)$

Information geometry
Springer journal
(2018-)

**Figure 3.** Genesis of the dual structure of information geometry.

a continuous measure. Many fruitful interactions between information geometry and optimal transport are investigated [AKO18], and counterpart notions of the FIM and Bregman divergences have been proposed in the probability density space equipped with the $L_2$-Wasserstein metric [Li21].

To summarize, the problem of geometrically modeling a family of probability distributions (the statistical model) is at the heart of information geometry. The Fisher-Rao geometry considers a Riemannian manifold equipped with the Fisher information metric, and uses the Riemannian geodesic length as a measure of dissimilarity between distributions: the Fisher-Rao distance. Amari's dual $\pm\alpha$-geometry of information geometry has revealed the dualistic structure of affine connections coupled to the Fisher metric. This key dualistic structure is purely geometric and therefore can be used beyond the realm of statistics (for example, when studying optimization algorithms with convex barrier functions [Ama16]). Dually flat spaces are Hessian manifolds [Shi07] with a single-chart atlas where the Legendre-Fenchel transformation plays an essential role to define dual coordinate systems and dual potential functions. Dually flat spaces generalize Euclidean geometry and enjoy a generalized Pythagorean theorem [Ama16]. Many pioneers have contributed to the now well-established classical dual structure of information geometry: Figure 3 displays historically the main actors who contributed to the genesis of the dual structure of information geometry with achieved milestones.

Recent advances in information geometry studies the geometry of deformed exponential families and their use in thermostatistics [Nau11], the geometry of nonparametric models, the quantum information geometry, the Lie group thermodynamics, and the interactions of geometric mechanics with information geometry via symplectic and contact structures. Information geometry has found many applications beyond statistics. We refer to the textbook [Ama16] for applications in signal processing, data science, and machine learning. To conclude with an application in machine learning, consider training a neural network $y = \mathrm{NN}_\theta(x)$ parameterized by weights $\theta$. The neural network is typically trained by using the method of gradient descent to minimize a loss function

$$L(\theta) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathrm{NN}_\theta(x_i))^2$$

defined by a supervised training set of $n$ labeled pairs $\{(x_i, y_i)\}$, where $y_i$ denotes the label of $x_i$: initialize $\theta_0$ and iteratively update $\theta_{t+1} := \theta_t - \beta\nabla L(\theta_t)$, where $\beta$ denotes the step size. The ordinary gradient $\nabla_\theta L(\theta)$ depends on the chosen parameterization, i.e., $\nabla_\theta L(\theta) \neq \nabla_\eta L(\theta(\eta))$ for a smooth invertible parameter transformation $\eta = \theta(\eta)$. A better parameter-invariant gradient has been proposed in information geometry for optimization on Riemannian manifolds $(M, g)$: the natural gradient [Ama16] $\tilde{\nabla}_\theta L(\theta) := [g_{ij}(\theta)]^{-1}_{ij}\nabla_\theta L(\theta)$. The natural gradient ensures that $\tilde{\nabla}_\theta L(\theta) = \tilde{\nabla}_\eta L(\theta(\eta))$. The natural gradient descent is used to train stochastic neural networks with parameter space modeled as a Fisher-Rao manifold, called a neuromanifold [Ama16]. Since 2018, an eponymous journal devoted to information geometry (INGE, `https://www.springer.com/journal/41884`) is published by Springer which reports the latest advances in the field.

## References


[Ama16] Shun-ichi Amari, *Information geometry and its applications*, Applied Mathematical Sciences, vol. 194, Springer, [Tokyo], 2016, DOI 10.1007/978-4-431-55978-8. MR3495836

[Ama21] Shun-ichi Amari, *Information geometry*, Jpn. J. Math. **16** (2021), no. 1, 1–48, DOI 10.1007/s11537-020-1920-5. MR4206647

[AKO18] Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi, *Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem*, Inf. Geom. **1** (2018), no. 1, 13–37, DOI 10.1007/s41884-018-0002-8. MR3974671

[AJLS17] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer, *Information geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 64, Springer, Cham, 2017, DOI 10.1007/978-3-319-56478-4. MR3701408

[Čen82] Nikolai Nikolaevich Čencov, *Statistical decision rules and optimal inference*, Translations of Mathematical Monographs, vol. 53, American Mathematical Soc., 1981.

[Efr75] Bradley Efron, *Defining the curvature of a statistical problem (with applications to second order efficiency)*, Ann. Statist. **3** (1975), no. 6, 1189–1242. MR428531

[Egu83] Shinto Eguchi, *Second order efficiency of minimum contrast estimators in a curved exponential family*, Ann. Statist. **11** (1983), no. 3, 793–803. MR707930

[GN14] Leonor Godinho and José Natário, *An introduction to Riemannian geometry: With applications to mechanics and relativity*, Universitext, Springer, Cham, 2014, DOI 10.1007/978-3-319-08666-8. MR3289090

[Kee10] Robert W. Keener, *Theoretical statistics: Topics for a core course*, Springer Texts in Statistics, Springer, New York, 2010, DOI 10.1007/978-0-387-93839-4. MR2683126

[Li21] Wuchen Li, *Transport information Bregman divergences*, `arXiv:2101.01162`, 2021.

[Nau11] Jan Naudts, *Generalised thermostatistics*, Springer-Verlag London, Ltd., London, 2011, DOI 10.1007/978-0-85729-355-8. MR2777415

[Nie20] Frank Nielsen, *An elementary introduction to information geometry*, Entropy **22** (2020), no. 10, Paper No. 1100, 61, DOI 10.3390/e22101100. MR4221069

[PC19] Gabriel Peyré and Marco Cuturi, *Computational optimal transport: With applications to data science*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.
[Shi07] Hirohiko Shima, *The geometry of Hessian structures*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007, DOI 10.1142/9789812707536. MR2293045
[Won18] Ting-Kam Leonard Wong, *Logarithmic divergences from optimal transport and Rényi geometry*, Inf. Geom. **1** (2018), no. 1, 39–78, DOI 10.1007/s41884-018-0012-6. MR4010746

Frank Nielsen

### Credits

The opening image is courtesy of ermess via Getty.
Figures 1–3 are courtesy of the author.
Photo of the author is courtesy of Maryse Beaumont.