WHAT IS...

# an Information Projection?

*Frank Nielsen*

*Communicated by Cesar E. Silva*

## Orthogonal Projections as Distance Minimizers

In Euclidean geometry, the orthogonal projection $p_S$ of a vector $p$ onto a subset $S$ as in Figure 1 can be defined as the point(s) $q$ of $S$ minimizing the distance $D(p,q)$ from $p$ to $q$. In general, the projection may not be unique: for example, projecting the center of a unit ball onto its boundary sphere yields the full boundary sphere. However, the projection $p_S$ is always guaranteed to be unique when $S$ is an affine subspace.

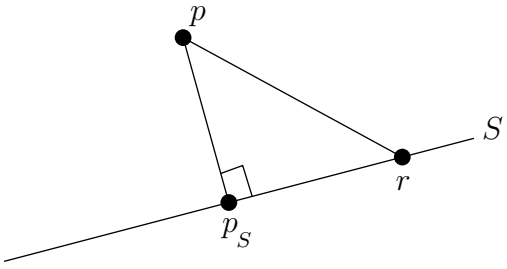


**Figure 1. In Euclidean geometry, the orthogonal projection $p_S$ of $p$ onto $S$ can be defined as the minimizer of the Euclidean distance: it is unique when $S$ is affine.**

We shall describe how these notions generalize to the dual geometry of information spaces.

## Dual Divergences and Information Projections

In information theory [2], we prefer to use a dissimilarity measure $D(p,q)$ between probability distributions $p(x)$ and $q(x)$ instead of the Euclidean distance. Since those distortion measures are often asymmetric, $D(p,q) \neq D(q,p)$, we use the notation $D(p : q)$ to highlight the asymmetric property of information distances and call $D(p : q)$ a *divergence*, assumed to be infinitely differentiable.

Here the word "divergence" is not to be confused with the divergence operator from calculus. Similar to the Euclidean case, an information projection of $p \in M$ onto $S \subset M$ can be defined by minimizing the divergence $D(q : p)$ for $q \in S$. Since the divergence is asymmetric, we define a dual divergence $D^*(p : q) = D(q : p)$.

## Information Monotonicity, Invariant Divergence, and Invariant Metric

For example, consider the space $M$ of Gaussian distributions on $\mathcal{X}$ the real line with

$$p(x;\xi) = p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

parameterized by $\xi = (\mu,\sigma)$. There exist many statistical distances $D(p(x;\xi_1) : p(x;\xi_2))$ for measuring the distortion between any two distributions of the statistical manifold $M$. However, assume we apply a mapping $y = k(x)$ (not necessarily one-to-one) and define the distributions

$$p'(y;\xi) = \int_{\{x:k(x)=y\}} p(x;\xi)\mathrm{d}x.$$

Then we would like to have

$$D(p'(y;\xi_1) : p'(y;\xi_2)) \leq D(p(x;\xi_1), p(x;\xi_2)),$$

with equality when $y = k(x)$ is one-to-one or when $k(x)$ is a sufficient statistic. This property of divergences is called the information monotonicity. Members of the class of statistical $f$-divergences

$$I_f(p : q) = \int_{\mathcal{X}} p(x) f(q(x)/p(x))\mathrm{d}x$$

defined for a convex function $f(u)$ satisfying $f(1) = 0$ have this property and are called invariant divergences. These include all divergences represented as sums or integrals of elementary scalar divergences that satisfy the information monotonicity (except for divergences on binary alphabets, with $\mathcal{X} = \{0,1\}$). Since $I_g(p : q) =$

*Frank Nielsen is professor of computer science at École Polytechnique, France, and senior researcher at Sony Computer Science Laboratories, Japan. His email address is* Frank.Nielsen@acm.org.



DOI: http://dx.doi.org/10.1090/noti1647

$I_f(p:q)$ for $g(u) = f(u) + c(u-1)$ with $c \in \mathbb{R}$, we may assume that $f'(1) = 0$. Furthermore, since $I_{\lambda f}(p:q) = \lambda I_f(p:q)$ for $\lambda > 0$, we define the standard $f$-divergences for $f''(1) = 1$. The dual $f$-divergence $I_f^*(p:q) = I_f(q:p)$ of a standard $f$-divergence $I_f(p:q)$ is a standard $f$-divergence obtained for the convex generator $f^\diamond(u) = uf(1/u)$.

Any standard $f$-divergence induces a Riemannian geometry $(M, g)$ given by a certain "Fisher information matrix." This metric is called the Fisher metric and allows one to define the Fisher orthogonality of vectors.

## Dual Geodesic Projections and Dual Pythagorean Theorems

The most fundamental distance or divergence in information theory is the Kullback-Leibler invariant divergence, commonly called $I$-divergence for short,

$$I(p:q) = I_f(p:q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x,$$

obtained for $f(u) = -\log u$.

Recall that in Euclidean geometry the line segment $[pp_S]$ meets the subset $S$ orthogonally at the projected point $p_S$ (that is, $[pp_S] \perp S$) and that the projection $p_S$ is guaranteed to be unique when $S$ is affine. Information geometry extends these results by revealing the dual nature of the $I$-divergence geometry using the framework of differential geometry. Consider $M$ as a smooth manifold of a family of distributions. When the family belongs to the exponential families (e.g., Gaussian distributions), the density can be written canonically as

$$p(x;\theta) = \exp(\langle \theta, t(x) \rangle - F(\theta)),$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product, $F(\theta) = \log \int_{\mathcal{X}} \exp(\langle \theta, t(x) \rangle) \mathrm{d}x$ is a $C^\infty$ (convex) "Legendre" function that ensures normalization to a probability distribution, and $\theta$ is the natural parameter belonging to the parameter space $\Theta = \{\theta \ : \ \int_{\mathcal{X}} \exp(\langle \theta, t(x) \rangle) \mathrm{d}x < \infty\}$. Any Legendre function $F(\theta)$ has a conjugate Legendre function [1] $F^\star(\eta)$ defined by

$$F^\star(\eta) = \sup_{\theta \in \Theta} \{\langle \theta, \eta \rangle - F(\theta)\}.$$

The parameter $\eta = \eta(\theta)$ is called the expectation parameter since $E_{x \sim p(x;\theta)}[t(x)] = \eta$. For the univariate Gaussian family, we get $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ and $\eta = (\mu, \mu^2 + \sigma^2)$ (with $t(x) = (x, x^2)$).

Thus a distribution of an exponential family can be indexed by either its natural parameter $\theta$ or its expectation parameter $\eta$: $p(x;\theta) = p(x;\eta)$, with the conversion $\theta = \nabla F^\star(\eta)$ and $\eta = \nabla F(\theta)$, where $\nabla$ denotes the gradient operator.

It turns out that the $I$-divergence between two distributions of the same exponential family is equivalent to a Bregman divergence:

$$I(p(x;\theta_1) : p(x;\theta_2)) = B_F(\theta_2 : \theta_1),$$

where

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

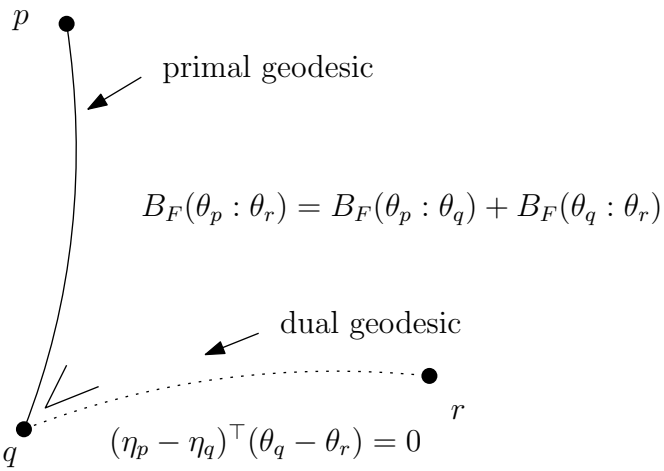


**Figure 2. The Pythagorean theorem in an information space.**

We can also express the information divergence using the $\eta$-parameters as $I(p(x;\eta_1) : p(x;\eta_2)) = B_{F^\star}(\eta_1 : \eta_2)$.

To connect two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ on the exponential family manifold $M = \{p(x,\theta) \ : \ \theta \in \Theta\}$, we may consider the path $\gamma_e(\theta_1, \theta_2, a) = p(x;\theta(a))$ with $\theta(a) = (1-a)\theta_1 + a\theta_2$ for $a \in [0,1]$. This path forms a 1D exponential family, and we can rewrite it by taking the logarithm as

$$\log p(x,\theta(a)) = (1-a) \log p(x;\theta_1) + a \log p(x;\theta_2) - F(\theta(a)).$$

This is a linear interpolation on the logarithmic scale, hence its name $e$-geodesic $\gamma_e(\theta_1, \theta_2) = \{\gamma_e(\theta_1, \theta_2, a) \ : \ a \in [0,1]\}$, which stands for exponential geodesic. Or, we can alternatively connect the two distributions using the path $\gamma_m(\eta_1, \eta_2, a) = p(x;\eta(a)) = p(x;(1-a)\eta_1 + a\eta_2)$. For discrete probability distributions $p_1$ and $p_2$ of the probability simplex, we get the mixture distribution $p(a) = (1-a)p_1 + ap_2$, hence its name $m$-geodesic $\gamma_m(\eta_1, \eta_2) = \{\gamma_m(\eta_1, \eta_2, a) \ : \ a \in [0,1]\}$, which stands for mixture geodesic. The $e$-geodesic and $m$-geodesic are visualized as straight line segments in the $\theta$- and $\eta$-coordinate systems, respectively. Let us define an $e$-flat subspace ($e$-flat for short) as an affine subspace in the $\theta$-coordinate system and an $m$-flat subspace ($m$-flat for short) as an affine subspace in the $\eta$-coordinate system.

The $e$-projection $p_S^e$ of $p$ onto $S$ is defined by minimizing $I(q:p)$ for $q \in S$ and is unique when $S$ is $m$-flat. The $m$-projection $p_S^m$ of $p$ onto $S$ is defined by minimizing $I(p:q)$ for $q \in S$ and is unique when $S$ is $e$-flat.

Similar to the Euclidean case, the proof of the uniqueness of the dual information geodesic projections follows from the dual Pythagorean theorems of Bregman divergences (Figure 2): When the triangle $pqr$ is such that $\gamma_m(\eta_p, \eta_q) \perp \gamma_e(\theta_q, \theta_r)$ (dual geodesics perpendicular at $q$), we have

$$B_F(\theta_p : \theta_r) = B_F(\theta_p : \theta_q) + B_F(\theta_q : \theta_r).$$

The orthogonality implies that $(\eta_p - \eta_q)^\top (\theta_q - \theta_r) = 0$.

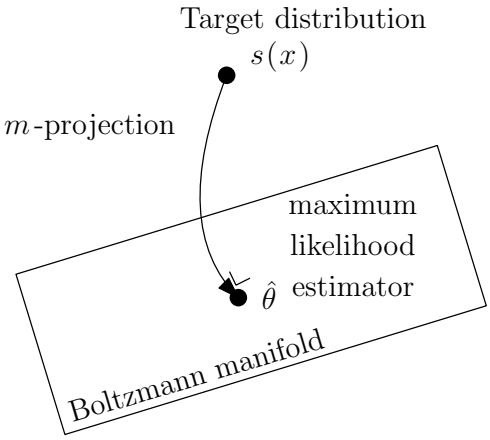


**Figure 3. The maximum likelihood estimator $\hat{\theta}$ of a Boltzmann machine is the unique $m$-projection of the target distribution $s(x)$ onto the Boltzmann manifold.**

When the triangle $pqr$ is such that $\gamma_e(\theta_p, \theta_q) \perp \gamma_m(\eta_q, \eta_r)$ (dual geodesics perpendicular at $q$), we have

$$B_{F^\star}(\eta_p : \eta_r) = B_{F^\star}(\eta_p : \eta_q) + B_{F^\star}(\eta_q : \eta_r).$$

The orthogonality implies that $(\theta_p - \theta_q)^\top(\eta_q - \eta_r) = 0$.

To illustrate the geodesic information projections, let us consider the following two examples: In machine learning [3], a Boltzmann machine is a fully interconnected network of $n$ stochastic units that defines an exponential family distribution on $\mathcal{X} = \{0,1\}^n$ by

$$p(x;\theta) = \exp\left(\sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j - F(\theta)\right),$$

where the $\theta_{ij}$'s are the weights connecting unit $i$ to unit $j$ and the $\theta_i$'s are the bias parameters. Boltzmann machines are universal approximators: they can represent any smooth distribution within any prescribed accuracy. The set of all machines $M = \{p(x;\theta) \ : \ \theta \in \Theta\}$ defines the Boltzmann $e$-flat manifold. Given a target distribution $s(x)$ that we wish to learn from, the Maximum Likelihood Estimator (MLE) $\hat{\theta}$ is characterized by the unique $m$-projection of $s(x)$ onto $M$ as in Figure 3.

The Maximum Entropy (MaxEnt) principle yields a distribution $p(x)$ maximizing the Shannon entropy under a set of $D$ moment constraints $E_X[t_i(X)] = m_i$, $i \in \{1,\dots,D\}$. It can be shown that the MaxEnt distribution belongs to an exponential family and is the unique $e$-projection of the uniform distribution on the $m$-flat manifold $\{X \ : \ E[t_1(X)] = m_1, \dots, E_X[t_D(X)] = m_D\}$.

A geodesic information projection $\min_{q\in S} D(q : p)$ can be rewritten as a point-set divergence $D(S : p)$. Consider two submanifolds $U$ and $V$ of $S$, and define

$$D(U : V) = \min_{u\in U, v\in V} D(u : v) = D(u^*, v^*),$$

where $u^*$ and $v^*$ form a closest pair between $U$ and $V$. We approximate a closest pair between the submanifolds by the alternating minimization algorithm: Begin with $v_1 \in V$, minimize $D(u : v_1)$ by an information projection to get $u_1$, and minimize $D(u_1 : V)$ to get $v_2$ by a dual information projection, etc. This alternating projection algorithm always converges, since the divergence decreases and is lower bounded by 0. Moreover, there is a unique closest pair when $V$ is flat and $U$ is dually flat as in Figure 4.

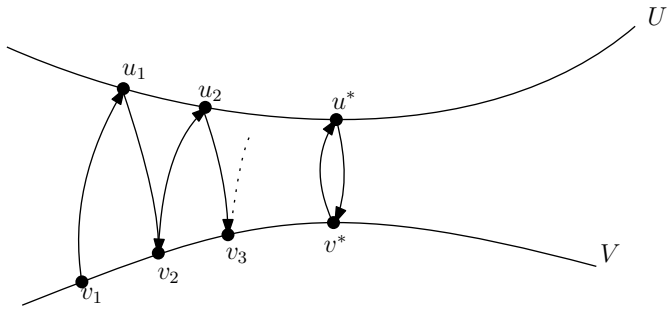


**Figure 4. The alternating geodesic projection algorithm for computing the divergence between submanifolds always converges.**

## Dual Geometry of Information Projections

Information projections are a core concept of information sciences that are met whenever minimizing divergences [2]. Depending on whether the minimization is carried out on the left argument of the divergence $D(\cdot : \cdot)$ or on its right argument (that is, the left argument of the dual divergence $D^*$), we end up with an information projection or a dual information projection. The geometric nature of information projections is elucidated using the dual geodesics of information geometry. In differential geometry, the notion of a geodesic $\gamma(p,q)$ passing through two points $p, q \in M$ depends on a connection. A connection $\prod_{p\to q}$ indicates how to transport vectors from one tangent plane $T_p$ to any other tangent plane $T_q$. A geodesic is then defined as an auto-parallel curve satisfying $\nabla_{\dot\gamma}\dot\gamma = 0$, where $\nabla$ is the covariant derivative associated to the connection. An affine connection $\nabla$ is defined by its Christoffel symbols. The fundamental structure of information geometry is a pair of torsion-free affine connections $\nabla$ and $\nabla^*$ that are coupled to a Riemannian metric tensor $g$ with $g^* = g$ and $\frac{\nabla+\nabla^*}{2} = \nabla^g$, the Levi-Civita metric connection. This dualistic structure $(M, g, \nabla, \nabla^*)$ can be built from any divergence $D(\cdot : \cdot)$ and generalizes Euclidean geometry. A manifold is called $\nabla$-flat if the Christoffel coefficients of $\nabla$ vanish in some coordinate system.

A dually flat geometry can be built from any smooth strictly convex function via the Legendre transformation and the corresponding dual Bregman divergences: those geometries are said to be dually flat because their primal and dual geodesics can be expressed as straight lines in the primal and dual affine coordinate systems, respectively.

Bregman divergences are the canonical divergences of dually flat manifolds: any dually flat manifold is induced by a corresponding Bregman divergence. Many stochastic neuronal network models (like the stochastic multilayer perceptrons [3] or the Boltzmann machines) handle exponential families in disguise and can thus be

studied using the method of information geometry and its dual information projections.

## References


[1] HEINZ H. BAUSCHKE and YVES LUCET, What is a Fenchel conjugate? *Notices Amer. Math. Soc.* **59** (2012), no. 1, 44–46. MR 2908159

[2] THOMAS M. COVER and JOY A. THOMAS, *Elements of Information Theory*, Second edition, John Wiley & Sons, 2006. MR 2239987

[3] IAN GOODFELLOW, YOSHUA BENGIO, and AARON COURVILLE, *Deep Learning*, MIT Press, 2016. MR 3617773


## Image Credits

Figures 1–4 courtesy of Frank Nielsen.
Photo of Frank Nielsen courtesy of Maryse Beaumont.

**Frank Nielsen**

### ABOUT THE AUTHOR

When not sitting on a transcontinental airplane, **Frank Nielsen** enjoys walking along the coastal paths of the French Riviera.



### *From the January 2018 Electronic Newsletter of the International Mathematical Union*[1]

The Committee for Women in Mathematics funded 10 proposals, most devoted to developing regional networks for Women in Mathematics, in Africa, Latin America, and Asia. Often the initiatives take the form of a meeting with both a mathematical part and a career development part. This is the case for two regional meetings of the African Women in Mathematics Association, one in Addis Ababa (Ethiopia) for East Africa and one in Ibadan (Nigeria) for West Africa, and also for the second Central Asia Women in Mathematics Association meeting in Uzbekistan. There are other meetings in India, Macedonia, El Salvador, and Uruguay. The African Women in Mathematics Association will also be writing portraits of African women mathematicians. Two further events are taking place in Europe, an ICTP school in Trieste (Italy) on Dynamical Systems, with all female organizers and lecturers, and the European Women in Mathematics General Meeting in Graz (Austria). In both cases the grant will be used to support the attendance of women from developing countries. The remaining part of the budget will be used to support $(WM)^2$, the first World Meeting for Women in Mathematics taking place on 31 July 2018 as a satellite event of ICM Rio.[2] In particular, women from all over the world who are supported by the Open Arms program have been invited to attend $(WM)^2$, with no registration fee. Submissions of scientific mathematical posters and thematic posters on women in mathematics to $(WM)^2$ is possible until 30 March.

[1] www.mathunion.org/organization/imu-net

[2] www.worldwomeninmaths.org/