An elementary introduction to information geometry

Frank Nielsen

Sony Computer Science Laboratories Inc, Japan

@FrnkNlsn

Slides: FrankNielsen.github.com

Shannon kicked off information sciences with his mathematical theory of communication in 1948: birth of **Information Theory** (IT)

- More generally, **Information Sciences** (IS) study “communication” between data and models (assuming a family of models *a priori*)

- Information sciences include statistics, information theory, (statistical) machine learning (ML), AI (neural networks), etc.

Inference

\[ \hat{\theta}_n(x_1, \ldots, x_n) \]: Extract information (= model parameter) from data about model via **estimators**

- Handle uncertainty due to limited/potentially corrupted data: **Wald framework**: Minimize risk in decision and find best decision rule
What is information geometry? A broad definition

\[ \text{IG} = \text{Geometry of decision making (incl. model fitting)} \]

- **Goodness-of-fit distance** between data and inferred models via data-model estimators (model fitting): statistics (likelihood), machine learning (classifier with loss functions), mathematical programming (constraints with objective functions)

But also

- **Distance between models**:
  - Decide between models (hypothesis testing/classification)
  - Model inference is a decision problem too: Decide which (parametric) model in a family models via decision rules (Wald)
Outline of this talk

1. Minimal introduction to differential geometry
   Manifold with an affine connection
   \((M, g, \nabla)\) = manifold \(M\) equipped with a metric tensor \(g\) and an affine connection \(\nabla\) defining \(\nabla\)-geodesics, parallel transport and covariant derivatives

2. Theory: Dual information-geometric structure \((M, g, \nabla, \nabla^*)\) and the fundamental theorem in information geometry

3. Statistical manifolds \((M, g, C)\)

4. Examples of dually metric-coupled connection geometry:
   A. Dual geometry induced by a divergence
   B. Dually flat Pythagorean geometry (from Bregman divergences)
   C. Expected \(\alpha\)-geometry (from invariant statistical \(f\)-divergences)

5. Some applications of these structures and a quick overview of some principled divergences

6. Summary: IG and its uses in statistics
1. Basics of Riemannian and non-Riemannian differential geometry
Manifold: Geometry and local coordinates

- **Geometric objects/entities** defined on smooth manifolds: tensors (fields), differential operators
- Geometric objects can be expressed using **local coordinate systems** using an **atlas** of overlapping **charts** covering the manifold
- Geometric calculations are **coordinate-free**, does not depend on any chosen coordinate system (unbiased). Coordinate-free & local coordinate equations.
Two essential geometric concepts on a manifold: $g$ and $\nabla$

- **Metric tensor** (field) $g$:
  allows to measure on tangent planes **angles** between vectors and vector **magnitudes** (“lengths”)

- **Connection** $\nabla$:
  a differential operator (hence the ’gradient’ notation $\nabla$) that allows to
  
  1. calculate the **covariant derivative** $Z = \nabla_Y X := \nabla(X, Y)$ of a vector field $X$ by another vector field $Y$
  2. “**parallel transport**” vectors between different tangent planes along smooth curves (path dependent)
  3. define $\nabla$-**geodesics** as autoparallel curves
Smooth vector fields $X \in \mathfrak{X}(M)$

- **Tangent plane** $T_p$ linearly approximates the manifold $M$ at point $p$
- **Tangent bundle** $TM = \bigcup_p T_p = \{(p, v), p \in M, v \in T_p\}$
  (More generally, fiber bundles like tensor bundles)
- A **tangent vector** $v$ plays the role of a **directional derivative**: $vf$ means derivative of smooth function $f \in \mathcal{C}(M)$ along the direction $v$
- Technically, a smooth **vector field** $X$ is defined as a “cross-section” for the tangent bundle: $X \in \mathfrak{X}(M) = \Gamma(TM)$

- In **local coordinates**, $X = \sum_i X^i e_i \equiv X^i e_i$ using Einstein convention on dummy indices
  $(X)_B = (X^i)$ with $(X)_B$ the **contravariant vector components** in the **natural basis** $B = \{e_i = \partial_i\}_i$ with $\partial_i := \frac{\partial}{\partial x_i}$ in chart $(\mathcal{U}, x)$.
  (Visualize coordinate lines in a chart)
Reciprocal basis and contravariant/covariant components

A tangent plane (=vector space) equipped with an inner product $\langle \cdot, \cdot \rangle$ induces a reciprocal basis $B^* = \{ e^i = \partial^i \}_i$ of $B = \{ e_i = \partial_i \}_i$ so that vectors can also be expressed using the covariant vector component in the natural reciprocal basis.

Primal/reciprocal basis are mutually orthogonal by construction.

\[
\langle e_i, e^j \rangle = \delta^j_i \]

\[
\langle e^i, e^j \rangle = \delta^j_i \]

\[
v^i = \langle v, e^i \rangle, \quad v_i = \langle v, e_i \rangle
\]

\[
 g_{ij} = \langle e_i, e_j \rangle, \quad g^{ij} = \langle e^i, e^j \rangle, \quad G^* = G^{-1}, \quad e^i \equiv \sum g^{ij} e_j, \quad e_i \equiv \sum g_{ij} e^j
\]
Metric tensor field \( g \) ("metric" \( g \) for short)

- **Symmetric positive-definite bilinear form**
  For \( u, v \in T_p \), \( g(u, v) \geq 0 \in \mathbb{R} \).
  Also written as \( g_p(u, v) = \langle u, v \rangle_g = \langle u, v \rangle_{g(p)} = \langle u, v \rangle \)

- **Orthogonal** vectors: \( u \perp v \iff \langle u, v \rangle = 0 \)

- **Vector length from induced norm**:
  \( \| u \|_p = \sqrt{\langle u, u \rangle_{g(p)}} \)

- In local coordinates of chart \((U, x)\), using matrix/vector algebra

  \[
  g(u, v) = (u)^T_B \times G_x(p) \times (v)_B = (u)^T_{B^*} \times G_x(p)^{-1} \times (v)_{B^*}
  \]

  (with primal/reciprocal basis, so that \( G \times G^{-1} = I \))

- **Geometry is conformal** when \( \langle \cdot, \cdot \rangle_p = \kappa(p) \langle \cdot, \cdot \rangle_{\text{Euclidean}} \)
  (angles/orthogonality without deformation)

- Technically, \( g \equiv \sum g_{ij} d x_i \otimes d x_j \) is a 2-covariant tensor field interpreted as a bilinear function (\( \rightarrow \) eating two contravariant vectors)
Defining an affine connection $\nabla$ via Christoffel symbols $\Gamma$

- $\nabla$ is a **differential operator** that defines the **derivative** $Z$ of a vector field $Y$ by another vector field $X$ (or tensors):

$$Z = \nabla_X Y := \nabla(X, Y)$$

- For a $D$-dimensional manifold, define $D^3$ **smooth functions** called **Christoffel symbols** (of the second kind) that induces the connection $\nabla$:

$$\Gamma^k_{ij} := (\nabla \partial_i \partial_j)^k \in \mathcal{F}(\mathcal{M}) \Rightarrow \nabla$$

where $(\cdot)^k$ defines the $k$-th component of vector

- The Christoffel symbols are **not** tensors because the transformation rules under a **change of basis** do not obey either of the tensor contravariant/covariant rules
Parallel transport $\prod^\nabla_c$ along a smooth curve $c(t)$

- Affine connection $\nabla$ defines a way to “translate” vectors = “**parallel transport**” vectors on tangent planes along a smooth curve $c(t)$. (Manifolds are not embedded)

- Parallel transport of vectors between tangent planes $T_p$ and $T_q$ linked by $c$ (defined by $\nabla$), with $c(0) = p$ and $c(1) = q$:

\[
\forall v \in T_p, \quad v_t = \prod_{c(0) \to c(t)}^\nabla v \in T_{c(t)}
\]
\(\nabla\)-geodesics

- \(\nabla\)-geodesics are **autoparallel curves** = curves \(\gamma\) such that their **velocity vector** \(\dot{\gamma} = \frac{d}{dt} \gamma(t)\) is moving on the curve (via \(\nabla\)) parallel to itself: \(\nabla\)-geodesics are "straight" lines.

\[
\nabla \dot{\gamma} \dot{\gamma} = 0
\]

- In local coordinates, \(\gamma(t) = (\gamma^k(t))\), the autoparallelism amounts to solve second-order Ordinary Differential Equations (ODEs) via the Euler-Lagrange equations:

\[
\frac{d^2}{dt^2} \gamma^k(t) + \Gamma^k_{ij} \frac{d}{dt} \gamma^i(t) \frac{d}{dt} \gamma^j(t) = 0, \quad \gamma'(t) = x' \circ \gamma(t)
\]

- In general, \(\nabla\)-geodesics are not available in closed-form (and require numerical schemes like Schild/pole ladders, etc.)

When \(\Gamma^k_{ij} = 0\), easy to solve: \(\rightarrow\) **straight lines** in the local chart!
Affine connection $\nabla$ compatible with the metric tensor $g$

- **Definition of metric compatibility** $g$ of an affine connection $\nabla$:
  - Geometric coordinate-free definition:

$$X\langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$$

- Definition using local coordinates with natural basis $\{\partial_i\}$:

$$\partial_k g_{ij} = \langle \nabla_{\partial_k} \partial_i, \partial_j \rangle + \langle \partial_i, \nabla_{\partial_k} \partial_j \rangle$$

- That is, the parallel transport is compatible with the metric $\langle \cdot, \cdot \rangle_g$:

$$\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \rightarrow c(t)} \nabla u, \prod_{c(0) \rightarrow c(t)} \nabla v \right\rangle_{c(t)} \quad \forall t.$$ 

  → preserves angles/lengths of vectors in tangent planes when transported along any smooth curve.
Fundamental theorem of Riemannian geometry

Theorem (Levi-Civita connection from metric tensor)

There exists a unique torsion-free affine connection compatible with the metric called the **Levi-Civita connection**: $^{\text{LC}}\nabla$

- Christoffel symbols of the Levi-Civita connection can be expressed from the metric tensor

\[
^{\text{LC}}\Gamma^k_{ij} \equiv \frac{1}{2} g^{kl} \left( \partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij} \right)
\]

where $g^{ij}$ denote the matrix elements of the inverse matrix $g^{-1}$.

- “Geometric” equation defining the Levi-Civita connection is given by the **Koszul formula**:

\[
2g(\nabla_X Y, Z) = X(g(Y, Z)) + Y(g(X, Z)) - Z(g(X, Y)) + g([X, Y], Z) - g([X, Z], Y) - g([Y, Z], X).
\]
Curvature tensor $R$ of an affine connection $\nabla$

- Riemann-Christoffel 4D curvature tensor $R (R^i_{jkl})$: Geometric equation:

$$R(X, Y)Z = \nabla_X \nabla_Y X - \nabla_Y \nabla_X Z - \nabla_{[X, Y]} Z$$

where $[X, Y](f) = X(Y(f)) - Y(X(f)) \ (\forall f \in \mathcal{F}(M))$ is the Lie bracket of vector fields

- Manifold is flat = $\nabla$-flat (i.e., $R = 0$) when there exists a coordinate system $x$ such that $\Gamma^k_{ij}(x) = 0$, i.e., all connection coefficients vanish

- Manifold is torsion-free when connection is symmetric: $\Gamma^k_{ij} = \Gamma^k_{ji}$ Geometric equation:

$$\nabla_X Y - \nabla_Y X = [X, Y]$$
Curvature and parallel transport along a loop (closed curve)

- In general, parallel transport is **path-dependent**
- The angle defect of a vector transported on a loop is related to the curvature.
- But for a **flat connection, it is path-independent**
  Eg., Riemannian cylinder manifold is (locally) flat.
2. Information-geometric structure:

Dual connections compatible with the metric and dual parallel transport
Conjugate connections \( \{ \nabla, \nabla^* \}_g \) and dual parallel transport \( \{ \prod^\nabla, \prod^{\nabla^*} \}_g \)

- **Definition of a conjugate connection** \( \nabla^* \) with respect to the metric tensor \( g \) (coupled with \( g \)): \( \nabla^* := C_g \nabla \)

\[
X\langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla^*_X Z \rangle
\]

- **Involution**: \( \nabla^{**} = C_g \nabla^* = \nabla \).

- **Dual parallel transport** of vectors is **metric compatible**:

\[
\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)} u, \prod_{c(0)\to c(t)} v \right\rangle_{c(t)}.
\]

- First studied by Norden (1945, relative geometry), studied also used in affine differential geometry, and in information geometry.
A 1D family of dualistic structures of information geometry

- Given a dualistic structure \((M, g, \nabla, \nabla^*)\) with \(\{\nabla, \nabla^*\}_g\), we have 
  \[ \tilde{\nabla} := \frac{\nabla + \nabla^*}{2} = \nabla^{LC} \nabla \] 
  coincides with the Levi-Civita connection induced by \(g\).

- Define the **totally symmetric cubic tensor** \(C\) called **Amari-Chentsov tensor**:
  \[ C_{ijk} = \Gamma^*_i j - \Gamma^j_i = C_{\sigma(i) \sigma(j) \sigma(k)} \]
  \[ C(X, Y, Z) = \langle \nabla_X Y - \nabla^*_X Y, Z \rangle, \quad C(\partial_i, \partial_j, \partial_k) = \langle \nabla_{\partial_i} \partial_j - \nabla^*_{\partial_i} \partial_j, \partial_k \rangle \]

- Then define a 1-**family of connections** \(\{\nabla^\alpha\}_{\alpha \in \mathbb{R}}\) dually coupled to the metric with \(\nabla^0 = \tilde{\nabla} = \nabla^{LC}\) (with \(\nabla^1 = \nabla\) and \(\nabla^{-1} = \nabla^*\)):
  \[ \Gamma^\alpha_{ijk} = \Gamma^0_{ijk} - \frac{\alpha}{2} C_{ijk}, \quad \Gamma^{-\alpha}_{ijk} = \Gamma^0_{ijk} + \frac{\alpha}{2} C_{ijk}, \]
  where \(\Gamma_{kij} \equiv \Gamma^l_{ij} g_{lk}\)

- **Theorem:** \((M, g, \{\nabla^{-\alpha}, \nabla^\alpha\}_g)\) is an information-geometric dualistic structure (with dual parallel transport) \(\forall \alpha \in \mathbb{R}\).
The fundamental theorem of information geometry

Theorem (Dually constant curvature manifolds)

If a torsion-free affine connection $\nabla$ has constant curvature $\kappa$ then its conjugate torsion-free connection $\nabla^* = C_g \nabla$ has the same constant curvature $\kappa$

Corollary I: Dually $\alpha$-flat manifolds

$\nabla^\alpha$-flat $\iff$ $\nabla^{-\alpha}$-flat

$(M, g, \nabla^\alpha)$ flat $\iff (M, g, \nabla^{-\alpha} = (\nabla^\alpha)^*)$ flat

Corollary II: Dually flat manifolds ($\alpha = \pm 1$)

1-flat $\iff$ $-1$-flat

$(M, g, \nabla)$ flat $\iff (M, g, \nabla^*)$ flat

$$R^\alpha(X, Y)Z - R^{-\alpha}(X, Y)Z = \alpha (R(X, Y)Z - R^*(X, Y)Z)$$

Divergence $D$ induces structure $(M, Dg, D\nabla^\alpha, D\nabla^{-\alpha}) \equiv (M, Dg, D\alpha C)$
Convex potential $F$ induces structure $(M, Fg, F\nabla^\alpha, F\nabla^{-\alpha}) \equiv (M, Fg, F\alpha C)$
(via Bregman divergence $B_F$)
Dually flat manifold

= Bregman dual Pythagorean geometry
Dually flat manifold

- Consider a strictly convex smooth function $F$, called a potential function.
- Associate its Bregman divergence (parameter divergence):

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$$

- The induced information-geometric structure is

$$(M, F_g, F_C) := (M, B_F g, B_F C)$$

with:

$$F_g := B_F g = - \left[ \partial_i \partial_j B_F(\theta : \theta') \big|_{\theta' = \theta} \right] = \nabla^2 F(\theta)$$

$$F_\Gamma := B_F \Gamma_{ijk}(\theta) = 0$$

$$F_C_{ijk} := B_F C_{ijk} = \partial_i \partial_j \partial_k F(\theta)$$

- The manifold is $F^\nabla$-flat.
- Levi-Civita connection $^{LC} \nabla$ is obtained from the metric tensor $F_g$ (usually not flat), and we get the conjugate connection $(F^\nabla)^* = F^\nabla^{-1}$ from $(M, F_g, F_C)$.
The Legendre-Fenchel transformation gives the **convex conjugate** $F^*$, the **dual potential function**: $F^*(\eta) := \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\}$

- Dual affine coordinate systems: $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$
  (Function convexity does not change by affine transformation)

- Crouzeix identity $\nabla^2 F(\theta) \nabla^2 F^*(\eta) = I$ reveals that $\{\partial_i\}_i/\{\partial^j\}_j$ are reciprocal basis.

- Bregman divergence reinterpreted using Young-Fenchel (in)equality as the **canonical divergence**:

  $$B_F(\theta : \theta') = A_{F,F^*}(\theta : \eta') = \boxed{F(\theta) + F^*(\eta') - \theta^\top \eta'} = A_{F^*,F}(\eta' : \theta)$$

- The **dual Bregman divergence**

  $$B_{F^*}(\theta : \theta') := B_F(\theta' : \theta) = B_{F^*}(\eta : \eta')$$

  yields

  $$F g^{ij}(\eta) = \partial^i \partial^j F^*(\eta), \quad \partial^l := \frac{\partial}{\partial \eta^l}$$

  $$F \Gamma^{ijk}_{*}(\eta) = 0, \quad F C^{ijk} = \partial^i \partial^j \partial^k F^*(\eta)$$

- The manifold is both $F \nabla$-flat and $F \nabla^*$-flat = **Dually flat manifold**
Dual Pythagorean theorems

Any pair of points can either be linked using the $\nabla$-geodesic ($\theta$-straight $[PQ]$) or the $\nabla^*$-geodesic ($\eta$-straight $[PQ]^*$).

$\Rightarrow$ There are $2^3 = 8$ geodesic triangles.

$\gamma^*(P, Q) \perp_F \gamma(Q, R) \Leftrightarrow (\eta(P) - \eta(Q))^\top \theta(Q) - \theta(R) = 0$

$\gamma(P, Q) \perp_F \gamma^*(Q, R) \Leftrightarrow (\theta(P) - \theta(Q))^\top \eta(Q) - \eta(R) = 0$

$D(P : R) = D(P : Q) + D(Q : R)$

$D^*(P : R) = D^*(P : Q) + D^*(Q : R)$
Dual Bregman projection: A uniqueness theorem

- A submanifold \( S \subset M \) is said \( \nabla\)-flat (\( \nabla^*\)-flat) if it corresponds to an affine subspace in \( \theta \) (in \( \eta \), respectively)

- In a dually flat space, there is a canonical (Bregman) divergence \( D \)

- The \( \nabla\)-projection \( P_S \) of \( P \) on \( S \) is unique if \( S \) is \( \nabla^*\)-flat and minimizes the \( \nabla\)-divergence \( D(\theta(P) : \theta(Q)) \):

\[
\nabla\text{-projection: } P_S = \arg \min_{Q \in S} D(\theta(P) : \theta(Q))
\]

- The dual \( \nabla^*\)-projection \( P_S^* \) is unique if \( M \subset S \) is \( \nabla\)-flat and minimizes the \( \nabla^*\)-divergence \( D^*(\theta(P) : \theta(Q)) = D(\theta(Q) : \theta(P)) \):

\[
\nabla^*\text{-projection: } P_S^* = \arg \min_{Q \in S} D(\theta(Q) : \theta(P))
\]
Expected $\alpha$-Geometry from Fisher information metric and skewness cubic tensor of a parametric family of distributions
Fisher Information Matrix (FIM), positive semi-definite

- **Parametric family** of probability distributions $\mathcal{P} := \{p_\theta(x)\}_\theta$ for $\theta \in \Theta$

- Likelihood function $L(\theta; x)$ and log-likelihood function $l(\theta; x) := \log L(\theta; x)$

- **Score vector** $s_\theta = \nabla_\theta l = (\partial_i l)^i$ indicates the **sensitivity** of the likelihood $\partial_i l := \frac{\partial}{\partial \theta_i} l(\theta; x)$

- **Fisher information matrix** (FIM) of $D \times D$ for $\dim(\Theta) = D$:

  \[
  \mathcal{P} I(\theta) := E_\theta \left[ \partial_i l \partial_j l \right]_{ij} \succeq 0
  \]

- Cramér-Rao lower bound\(^1\) on the variance of any unbiased estimator

  \[
  V_\theta[\hat{\theta}_n(X)] \succeq \frac{1}{n} \mathcal{P} I^{-1}(\theta)
  \]

- FIM invariant by reparameterization of the sample space $\mathcal{X}$, and **covariant** by reparameterization of the parameter space $\Theta$.

---

\(^1\)F. Nielsen. “Cramér-Rao lower bound and information geometry”. In: Connected at Infinity II. Springer, 2013.
Some examples of Fisher information matrices

- **FIM of an exponential family**: Include Gaussian, Beta, \( \Delta_D \), etc.

\[
\mathcal{E} = \left\{ p_\theta(x) = \exp \left( \sum_{i=1}^{D} t_i(x)\theta_i - F(\theta) + k(x) \right) \text{ such that } \theta \in \Theta \right\}
\]

\( F \) is the strictly convex cumulant function

\[
\mathcal{E} I(\theta) = \text{Cov}_\theta[t(X)] = \nabla^2 F(\theta) = \nabla^2 F^*(\eta)^{-1} \succeq 0
\]

- **FIM of a mixture family**: Include statistical mixtures, \( \Delta_D \)

\[
\mathcal{M} = \left\{ p_\theta(x) = \sum_{i=1}^{D} \theta_i F_i(x) + C(x) \text{ such that } \theta \in \Theta \right\}
\]

with \( \{F_i(x)\}_i \) linearly independent on \( \mathcal{X} \), \( \int F_i(x)d\mu(x) = 0 \) and \( \int C(x)d\mu(x) = 1 \)

\[
\mathcal{M} I(\theta) = E_\theta \left[ \frac{F_i(x)F_j(x)}{p_\theta^2(x)} \right] = \int_{\mathcal{X}} \frac{F_i(x)F_j(x)}{p_\theta(x)} d\mu(x) \succeq 0
\]
Expected $\alpha$-geometry from expected dual $\alpha$-connections

- Fisher "information metric" tensor from FIM (regular models)
  \[ \mathcal{P} g(u, v) := (u)_{\theta}^{\top} \mathcal{P} l_{\theta}(\theta)(v)_{\theta} \succeq 0 \]

- Exponential connection and mixture connection:
  \[ \mathcal{P} \nabla^e := E_{\theta} \left[ (\partial_i \partial_j l)(\partial_k l) \right], \quad \mathcal{P} \nabla^m := E_{\theta} \left[ (\partial_i \partial_j l + \partial_i l \partial_j l)(\partial_k l) \right] \]

- Dualistic structure ($\mathcal{P}, \mathcal{P} g, \mathcal{P} \nabla^-, \mathcal{P} \nabla^+$) with cubic skewness tensor:
  \[ C_{ijk} = E_{\theta} \left[ \partial_i l \partial_j l \partial_k l \right] \]

- It follows a one-family of $\alpha$-CCMs: \( \{ (\mathcal{P}, \mathcal{P} g, \mathcal{P} \nabla^{-\alpha}, \mathcal{P} \nabla^{+\alpha}) \} \) with
  \[ \mathcal{P} \Gamma_{ij}^{\alpha k} := -\frac{1 + \alpha}{2} C_{ijk} = E_{\theta} \left[ \left( \partial_i \partial_j l + \frac{1 - \alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right] \]

  \[ \mathcal{P} \nabla^{-\alpha} + \mathcal{P} \nabla^{\alpha} = \frac{L \mathcal{C} \nabla}{2} = \mathcal{P} \nabla := L \mathcal{C} \nabla (\mathcal{P} g) \]

- In case of an exponential family $\mathcal{E}$ or a mixture family $\mathcal{M}$, we get
  Dually Flat Manifolds (Bregman geometry/ std $f$-divergence)

  \[ \mathcal{E} \Gamma = \mathcal{M} \Gamma = \mathcal{E} \Gamma = \mathcal{M} \Gamma = 0 \]
Statistical invariance criteria

Questions:

- Which metric tensors $g$ make statistical sense?
- Which affine connections $\nabla$ make statistical sense?
- Which statistical divergences make sense?
- Invariant metric $g$ shall preserve the inner product under important statistical mappings, called Markov mappings.

Theorem (Uniqueness of Fisher metric)

*Fisher metric is unique up to a scaling constant.*

---

Divergence: Information monotonicity and $f$-divergences

- A divergence satisfies the **information monotonicity** iff
  \[ D(\theta_{\tilde{A}} : \theta'_{\tilde{A}}) \leq D(\theta : \theta') \]
  for any **coarse-grained partition** $\mathcal{A}$ ($\mathcal{A}$-lumping)

\[
\begin{array}{cccccccc}
  p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p \\
\end{array}
\]

\[
\begin{array}{cccccccc}
  p_1 + p_2 & p_3 + p_4 + p_5 & p_6 & p_7 + p_8 & p_A \\
\end{array}
\]

- The only invariant and **decomposable** divergences are $f$-divergences\(^3\) (when $D > 2$, curious binary case):

  \[
  I_f(\theta : \theta') := \sum_i \theta_i f \left( \frac{\theta'_i}{\theta_i} \right) \geq f(1), \quad f(1) = 0
  \]

- **Standard $f$-divergences** are defined for $f$-generators satisfying
  \[
  f'(1) = 0 \text{ (fix $f_\lambda(u) := f(u) + \lambda(u - 1)$ since $I_{f_\lambda} = I_f$), and $f''(u) = 1$ (fix scale)}
  \]

---

Properties of Csiszár-Ali-Silvey $f$-divergences

- Statistical $f$-divergences are **invariant**\(^4\) under one-to-one/sufficient statistic transformations $y = t(x)$ of the sample space: $p(x; \theta) = q(y(x); \theta)$.

\[
I_f[p(x; \theta) : p(x; \theta')] = \int_X p(x; \theta) f \left( \frac{p(x; \theta')}{p(x; \theta)} \right) \, d\mu(x)
\]
\[
= \int_Y q(y; \theta) f \left( \frac{q(y; \theta')}{q(y; \theta)} \right) \, d\mu(y)
\]
\[
= I_f[q(y; \theta) : q(y; \theta')]
\]

- **Dual $f$-divergences** for reference duality:

\[
I_f^*[p(x; \theta) : p(x; \theta')] = I_f[p(x; \theta') : p(x; \theta)] = I_{f^\diamond}[p(x; \theta) : p(x; \theta')]
\]

for the standard **conjugate $f$-generator** (diamond $f^\diamond$ generator):

\[
f^\diamond(u) := uf \left( \frac{1}{u} \right)
\]

---

Some common examples of $f$-divergences

- The family of $\alpha$-divergences:

$$I_{\alpha}[p : q] := \frac{4}{1 - \alpha^2} \left(1 - \int p^{\frac{1-\alpha}{2}}(x) q^{1+\alpha}(x) d\mu(x)\right)$$

obtained for $f(u) = \frac{4}{1-\alpha^2}(1 - u^{\frac{1+\alpha}{2}})$, and include

- Kullback-Leibler $\text{KL}[p : q] = \int p(x) \log \frac{p(x)}{q(x)} d\mu(x)$ for $f(u) = -\log u$ ($\alpha = 1$)
- reverse Kullback-Leibler $\text{KL}^*[p : q] = \int q(x) \log \frac{q(x)}{p(x)} d\mu(x)$ for $f(u) = u \log u$ ($\alpha = -1$)
- squared Hellinger divergence $H^2[p : q] = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x)$ for $f(u) = (\sqrt{u} - 1)^2$ ($\alpha = 0$)
- Pearson and Neyman chi-squared divergences

- Jensen-Shannon divergence (bounded):

$$\frac{1}{2} \int \left(p(x) \log \frac{2p(x)}{p(x) + q(x)} + q(x) \log \frac{2q(x)}{p(x) + q(x)}\right) d\mu(x)$$

for $f(u) = -(u + 1) \log \frac{1+u}{2} + u \log u$

- Total Variation (metric) $\frac{1}{2} \int |p(x) - q(x)| d\mu(x)$ for $f(u) = \frac{1}{2} |u - 1|$
Invariant $f$-divergences yields Fisher metric/$\alpha$-connections

- Invariant (separable) standard $f$-divergences related infinitesimally to the Fisher metric:

\[
I_f[p(x; \theta) : p(x; \theta + d\theta)] = \int p(x; \theta)f\left(\frac{p(x; \theta + d\theta)}{p(x; \theta)}\right)\,d\mu(x)
\]

\[
\sum 1 = \frac{1}{2} Fg_{ij}(\theta) d\theta^i d\theta^j
\]

- A statistical parameter divergence on a parametric family of distributions yield an equivalent parameter divergence:

\[
\mathcal{P} D(\theta : \theta') := D_\mathcal{P}[p(x; \theta) : p(x; \theta')]
\]

Thus we can build the information-geometric structure induced by this parameter divergence $\mathcal{P} D(\cdot : \cdot)$

- For $\mathcal{P} D(\cdot : \cdot) = I_f[\cdot : \cdot]$, the induced $\pm 1$-divergence connections $\mathcal{P} I_f \nabla := \mathcal{P} I_f \nabla$ and $\mathcal{P} I_f^* \nabla := \mathcal{P} I_f^* \nabla$ are the expected $\pm \alpha$-connections with $\alpha = 2f'''(1) + 3$. $+1$-connection $= e/m$-connection
4.

Some applications of the information-geometric manifolds

\((M, g, \nabla, \nabla^*) \equiv (M, g, C)\)

\((M, g, \nabla^{-\alpha}, \nabla^{\alpha}) \equiv (M, g, \alpha C)\)
Broad view of applications of Information Geometry

- **Statistics:**
  Asymptotic inference, Expectation-Maximization (EM/em), time series ARMA models

- **Statistical machine learning:**
  Restricted Boltzmann machines (RBMs), **neuromanifolds** (deep learning) and natural gradient

- **Signal processing:**
  PCA, ICA, Non-negative Matrix Factorization (NMF)

- **Mathematical programming:**
  Barrier function of interior point methods

- **Game theory:**
  Score functions

---


Hypothesis testing in the dually flat exponential family manifold
Bayesian Hypothesis Testing/Binary classification

Given two distributions $P_0$ and $P_1$, classify observations $X_{1:n}$ (= decision problem) as either iid sampled from $P_0$ or from $P_1$? For example, $P_0 =$ signal, $P_1 =$ noise (assume same prior $w_1 = w_2 = \frac{1}{2}$)

Assume $P_0 \sim P_{\theta_0}$, $P_1 \sim P_{\theta_1} \in \mathcal{E}$ belong to an **exponential family manifold** $\mathcal{E} = \{ P_{\theta} \}$ with dually flat structure $(\mathcal{E}, \varepsilon g, \varepsilon \nabla^e, \varepsilon \nabla^m)$. This structure can also be derived from a **divergence manifold structure** $(\mathcal{E}, \text{KL}^*)$.

Therefore $\text{KL}[P_{\theta} : P_{\theta'}]$ amounts to a Bregman divergence (for the cumulant function of the exponential family):

$$\text{KL}[P_{\theta} : P_{\theta'}] = B_F(\theta' : \theta)$$
Hypothesis Testing (HT)/Binary classification

The best exponent error of the best Maximum A Priori (MAP) decision rule is found by minimizing the Bhattacharyya distance to get the Chernoff information:

\[ C[P_1, P_2] = -\log \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)d\mu(x) \geq 0, \]

On \( \mathcal{E} \), the Bhattacharyya distance amounts to a skew Jensen parameter divergence:

\[ J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1 - \alpha) F(\theta_2) - F(\theta_1 + (1 - \alpha)\theta_2), \]

**Theorem**: Chernoff information = Bregman divergence for exponential families at the optimal exponent value:

\[ C[P_{\theta_1} : P_{\theta_2}] = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)}) \]

---


10 Nielsen, “An Information-Geometric Characterization of Chernoff Information”. 

Geometry\textsuperscript{11} of the best error exponent: binary hypothesis on the exponential family manifold

\[ P^* = P_{\theta_{12}^*} = Ge(P_1, P_2) \cap Bi_m(P_1, P_2) \]

\[ C(\theta_1 : \theta_2) = B(\theta_1 : \theta_{12}^*) \]

**Synthetic information geometry** ("Hellinger arc"): Exact characterization but not necessarily closed-form formula

\textsuperscript{11}Nielsen, “An Information-Geometric Characterization of Chernoff Information”.
Information geometry of multiple hypothesis testing\textsuperscript{13}

Bregman Voronoi diagrams\textsuperscript{12} on $E$ of $P_1, \ldots, P_n \in E$

\begin{itemize}
  \item $\eta$-coordinate system
  \item Chernoff distribution between natural neighbours
\end{itemize}


\textsuperscript{13} F. Nielsen. “Hypothesis Testing, Information Divergence and Computational Geometry”. In: GSI. 2013, pp. 241–248; Pham, Boyer, and Nielsen, “Computational Information Geometry for Binary Classification of High-Dimensional Random Tensors”.
Clustering statistical $\psi$-mixtures on the dually flat mixture family manifold $M$
Dually flat mixture family manifold\textsuperscript{14}

Consider $D + 1$ prescribed component distributions $\{p_0, \ldots, p_D\}$ and form the mixture manifold $\mathcal{M}$ by considering all their convex weighted combinations $\mathcal{M} = \{m_\theta(x) = \sum_{i=0}^{D} w_i p_i(x)\}$

Information-geometric structure: $(\mathcal{M}, \mathcal{M}g, \mathcal{M}\nabla^{-1}, \mathcal{M}\nabla^1)$ is dually flat and equivalent to $(M_\theta, \text{KL})$. That is, the KL between two mixtures with prescribed components is equivalent to a Bregman divergence.

$$\text{KL}[m_\theta : m_\theta'] = B_F(\theta : \theta'), F(\theta) = -h(m_\theta) = \int m_\theta(x) \log m_\theta(x) d\mu(x)$$

Clustering on the dually flat \( w \)-mixture family manifold

Apply Bregman \( k \)-means\(^{15} \) on **Monte Carlo dually flat spaces**\(^{16} \) of \( w \)-GMMs (Gaussian Mixture Models)

Because \( F = -h(m(x; \theta)) \) is the **negative differential entropy** of a mixture (not available in closed form\(^{17} \)), we approximate \( F \) by Monte-Carlo convex \( \tilde{F}_\chi \) and get a sequence of tractable dually flat manifolds converging to the ideal mixture manifold

---

\(^{15}\) F. Nielsen and R. Nock. “Sided and symmetrized Bregman centroids”. In: *IEEE transactions on Information Theory* 55.6 (2009).


Divergences and geometry: Not one-to-one in general
Principled classes of distances and divergences: Axiomatic approach with exhaustivity characteristics, conformal

divergences\textsuperscript{18}, projective divergences\textsuperscript{19}, etc.


The **fundamental information-geometric structure** \((M, g, \nabla, \nabla^*)\) is defined by a pair of conjugate affine connections coupled to the metric tensor. It yields a dual parallel transport that is metric-compatible.

- Not related to any application field! (but can be realized by statistical models\(^{20}\))
- Not related to any distance/divergence (but can be realized by divergences or canonical divergences can be defined)
- Not unique since we can always get a **1D family of \(\alpha\)-geometry**

Conjugate flat connections have geodesics corresponding to straight lines in dual affine coordinates (Legendre convex conjugate potential functions), and a **dual Pythagorean theorem** characterizes uniqueness of **information projections**\(^{21}\) (\(\rightarrow\) Hessian manifolds)

These “pure” **geometric structures** \((M, g, \nabla, \nabla^*) \equiv (M, g, C)\) can be applied in information sciences (Statistics, Mathematical Programming, etc.). But is it fit to the application field?

---


In statistics, **Fisher metric is unique invariant metric tensor** (Markov morphisms).

In statistics, **expected $\alpha$-connections arise from invariant $f$-divergences**, and at infinitesimal scale standard $f$-divergences expressed using Fisher information matrix.

**Divergences** get an underlying information geometry: conformal divergences$^{22}$, projective$^{23}$ divergences, etc.

A Riemannian distance is never a divergence (not smooth), and a (symmetrized) divergence $J = \frac{D + D^*}{2}$ cannot always be powered $J^\beta$ ($\beta > 0$) to yield a metric distance (although $\sqrt{JS}$ is a metric).

Information geometry has a major role to play in machine learning and deep learning by studying **neuromanifolds**

---

$^{22}$Nock, Nielsen, and Amari, “On Conformal Divergences and Their Population Minimizers”.
$^{23}$Nielsen, Sun, and Marchand-Maillet, “On Hölder Projective Divergences”.