arXiv:1808.08271v2 [cs.LG] 6 Sep 2020

An elementary introduction to information geometry

Frank Nielsen
Sony Computer Science Laboratoties Inc
Tokyo, Japan

Abstract

In this survey, we describe the fundamental differential-geometric structures of information manifolds, state the fundamental theorem of information geometry, and illustrate some use cases of these information manifolds in information sciences. The exposition is self-contained by concisely introducing the necessary concepts of differential geometry, but proofs are omitted for brevity.

Keywords: Differential geometry; metric tensor; affine connection; metric compatibility; conjugate connections; dual metric-compatible parallel transport; information manifold; statistical manifold; curvature and flatness; dually flat manifolds; Hessian manifolds; exponential family; mixture family; statistical divergence; parameter divergence; separable divergence; Fisher-Rao distance; statistical invariance; Bayesian hypothesis testing; mixture clustering; αembeddings; gauge freedom

1 Introduction

1.1 Overview of information geometry

We present a concise and modern view of the basic structures lying at the heart of *Information Geometry* (IG), and report some applications of those information-geometric manifolds (herein termed "information manifolds") in statistics (Bayesian hypothesis testing) and machine learning (statistical mixture clustering).

By analogy to *Information Theory* (IT) (pioneered by Claude Shannon in his celebrated 1948's paper [119]) which considers primarily the communication of messages over noisy transmission channels, we may define *Information Sciences* (IS) as the fields that study "communication" between (noisy/imperfect) data and families of models (postulated as *a priori* knowledge). In short, information sciences seek methods to *distill* information from data to models. Thus information sciences encompass information theory but also include the fields of Probability & Statistics, Machine Learning (ML), Artificial Intelligence (AI), Mathematical Programming, just to name a few.

We review some key milestones of information geometry and report some definitions of the field by its pioneers in §5.2. Professor Shun-ichi Amari, the founder of modern information geometry, defined information geometry in the preface of his latest textbook [8] as follows: "Information geometry is a method of exploring the world of information by means of modern geometry." In short, information geometry geometrically investigates information sciences. It is a mathematical endeavour to define and bound the term geometry itself as geometry is open-ended. Often, we start by studying the invariance of a problem (eg., invariance of distance between probability distributions) and get as a result a novel geometric structure (eg., a "statistical manifold"). However, a geometric structure is "pure" and thus may be applied to other application areas beyond the scope of the original problem (eg, use of the dualistic structure of statistical manifolds in mathematical programming [57]): the method of geometry [9] thus yields a pattern of abduction [103, 115].

A narrower definition of information geometry can be stated as the field that studies the *geometry of decision making*. This definition also includes *model fitting* (inference) which can be interpreted as a decision problem as illustrated in Figure 1: Namely, deciding which model parameter to choose from a family of parametric models. This framework was advocated by Abraham Wald [131, 132, 36] who considered all statistical problems as statistical decision problems. Dissimilarities (also loosely called distances among

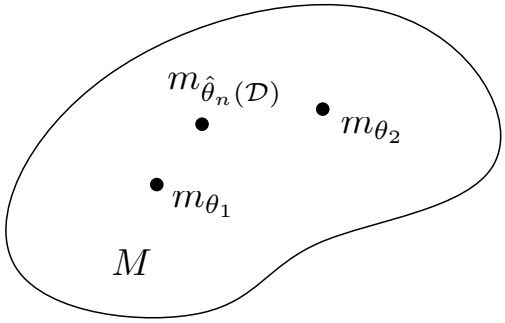


Figure 1: The parameter inference $\hat{\theta}$ of a model from data $\mathcal{D}$ can also be interpreted as a decision making problem: Decide which parameter of a parametric family of models $M = \{m_\theta\}_{\theta \in \Theta}$ suits the "best" the data. Information geometry provides a differential-geometric structure on manifold M which useful for designing and studying statistical decision rules.

others) play a crucial role not only for measuring the *goodness-of-fit* of data to model (say, likelihood in statistics, classifier loss functions in ML, objective functions in mathematical programming or operations research, etc.) but also for measuring the discrepancy (or deviance) between models.

One may ponder why adopting a geometric approach? Geometry allows one to study *invariance* of "figures" in a coordinate-free framework. The *geometric language* (e.g., line, ball or projection) also provides affordances that help us reason intuitively about problems. Note that although figures can be visualized (i.e., plotted in coordinate charts), they should be thought of as purely abstract objects, namely, geometric figures.

Geometry also allows one to study *equivariance*: For example, the centroid $c(T)$ of a triangle is equivariant under any affine transformation A: $c(A.T) = A.c(T)$. In Statistics, the Maximum Likelihood Estimator (MLE) is equivariant under a monotonic transformation g of the model parameter θ: $\widehat{(g(\theta))} = g(\hat{\theta})$, where the MLE of θ is denoted by $\hat{\theta}$.

1.2 Outline of the survey

This survey is organized as follows:

In the first part (§2), we start by concisely introducing the necessary background on differential geometry in order to define a manifold structure (M, g, ∇), ie., a manifold M equipped with a metric tensor field g and an affine connection ∇. We explain how this framework generalizes the Riemannian manifolds (M, g) by stating the fundamental theorem of Riemannian geometry that defines a unique torsion-free metric-compatible Levi-Civita connection which can be derived from the metric tensor.

Bruh I hope they explain that

In the second part (§3), we explain the dualistic structures of information manifolds: We present the conjugate connection manifolds (M, g, ∇, ∇^*), the statistical manifolds (M, g, C) where C denotes a cubic tensor, and show how to derive a family of information manifolds $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ for $\alpha \in \mathbb{R}$ provided any given pair $(\nabla = \nabla^{-1}, \nabla^* = \nabla^1)$ of conjugate connections. We explain how to get conjugate connections ∇ and ∇^* coupled to the metric g from any smooth (potentially asymmetric) distances (called divergences), present the dually flat manifolds obtained when considering Bregman divergences, and define, when dealing with parametric family of probability models, the exponential connection ${}^e\nabla$ and the mixture connection ${}^m\nabla$ that are dual connections coupled to the Fisher information metric. We discuss the concept of statistical invariance for the metric tensor and the notion of information monotonicity for statistical divergences [30, 8]. It follows that the Fisher information metric is the unique invariant metric (up to a scaling factor), and that the f-divergences are the unique separable invariant divergences.

In the third part (§4), we illustrate how to use these information-geometric structures in simple applications: First, we described the natural gradient descent method in §4.1 and its relationships with the

Riemannian gradient descent and the Bregman mirror descent. Second, we consider two applications in dually flat spaces in §4.2: In the first application, we consider the problem of Bayesian hypothesis testing and show how Chernoff information (which defines the best error exponent) can be geometrically characterized on the dually flat structure of an exponential family manifold. In the second application, we show how to cluster statistical mixtures sharing the same component distributions on the dually flat mixture family manifold.

Finally, we conclude in §5 by summarizing the important concepts and structures of information geometry, and by providing further references and textbooks [25, 8] for further readings to more advanced structures and applications of information geometry. We also mention recent studies of generic classes of principled distances and divergences.

In the Appendix §A, we show how to estimate the statistical f-divergences between two probability distributions in order to ensure that the estimates are non-negative in §B, and report the canonical decomposition of the multivariate Gaussian family, an example of exponential family which admits a dually flat structure.

At the beginning of each part, we start by outlining its contents. A summary of the notations used throughout this survey is provided page 47.

2 Prerequisite: Basics of differential geometry

In §2.1, we review the very basics of Differential Geometry (DG) for defining a manifold (M, g, ∇) equipped with both a metric tensor field g and an affine connection ∇. We explain these two *independent* metric/connection structures in §2.2 and in §2.3, respectively. From an affine connection ∇, we show how to derive the notion of covariant derivative in §2.3.1, parallel transport in §2.3.2 and geodesics in §2.3.3. We further explain the *intrinsic curvature and torsion* of manifolds induced by the connection in §2.3.4, and state the fundamental theorem of Riemannian geometry in §2.4: The existence of a unique torsion-free Levi-Civita connection ${}^{\mathrm{LC}}\nabla$ compatible with the metric (metric connection) that can be derived from the metric tensor g. Thus the Riemannian geometry (M, g) is obtained as a special case of the more general manifold structure $(M, g, {}^{\mathrm{LC}}\nabla)$: $(M, g) \equiv (M, g, {}^{\mathrm{LC}}\nabla)$. Information geometry shall further consider a dual structure (M, g, ∇^*) associated to (M, g, ∇), and the pair of dual structures shall form an information manifold (M, g, ∇, ∇^*).

2.1 Overview of differential geometry: Manifold (M, g, ∇)

Informally speaking, a *smooth D-dimensional manifold* M is a topological space that locally behaves like the D-dimensional Euclidean space $\mathbb{R}^D$. Geometric objects (e.g., points, balls, and vector fields) and entities (e.g., functions and differential operators) live on M, and are *coordinate-free* but can conveniently be expressed in *any* local coordinate system of an atlas $\mathcal{A} = \{(\mathcal{U}_i, x_i)\}_i$ of charts$(\mathcal{U}_i, x_i)$'s (fully covering the manifold) for calculations. Historically, René Descartes (1596-1650) allegedly invented the global Cartesian coordinate system while wondering how to locate a fly on the ceiling from his bed. In practice, we shall use the most expedient coordinate system to facilitate calculations. In information geometry, we usually handle a single chart fully covering the manifold.

A C^k manifold is obtained when the *change of chart transformations* are C^k. The manifold is said smooth when it is C^∞. At each point $p \in M$, a tangent plane T_p locally best linearizes the manifold. On any smooth manifold M, we can define two *independent* structures:

1. a metric tensor g, and
2. an affine connection ∇.

tangent plane Tp is like the local flat approximation to the manifold at a point

a metric tensor (or simply metric) is an additional structure on a manifold M (such as a surface) that allows defining distances and angles, just as the inner product on a Euclidean space allows defining distances and angles there.

The metric tensor g induces on each tangent plane T_p an *inner product space* that allows one to measure vector magnitudes (vector "lengths") and angles/orthogonality between vectors. The affine connection ∇ is a differential operator that allows one to define:

1. the *covariant derivative operator* which provides a way to calculate differentials of a vector field Y with respect to another vector field X: Namely, the covariant derivative $\nabla_X Y$,

2. the *parallel transport* $\prod_c^{\nabla}$ which defines a way to transport vectors between tangent planes along any smooth curve c, How to move between points on a manifold

3. the notion of ∇-geodesics γ_∇ which are defined as autoparallel curves, thus extending the ordinary notion of Euclidean straightness,

4. the intrinsic curvature and torsion of the manifold.

2.2 Metric tensor fields g

The *tangent bundle* of M is defined as the "union" of all tangent spaces:

$$TM := \cup_p T_p = \{(p, v), \quad p \in M, v \in T_p\}. \tag{1}$$

Thus the tangent bundle TM of a D-dimensional manifold M is of dimension $2D$. (The tangent bundle is a particular example of a fiber bundle with base manifold M.)

Informally speaking, a *tangent vector* v plays the role of a directional derivative, with vf informally meaning the derivative of a smooth function f (belonging to the space of smooth functions $\mathfrak{F}(M)$) along the direction v. Since the manifolds are abstract and not embedded in some Euclidean space, we do not view a vector as an "arrow" anchored on the manifold. Rather, vectors can be understood in several ways in differential geometry like directional derivatives or equivalent class of smooth curves at a point. That is, tangent spaces shall be considered as the manifold abstract too.

A smooth *vector field* X is defined as a "cross-section" of the tangent bundle: $X \in \mathfrak{X}(M) = \Gamma(TM)$, where $\mathfrak{X}(M)$ or $\Gamma(TM)$ denote the space of smooth vector fields. A basis $B = \{b_1, \ldots, b_D\}$ of a finite D-dimensional vector space is a *maximal linearly independent set of vectors*: A set of vectors $B = \{b_1, \ldots, b_D\}$ is linearly independent if and only if $\sum_{i=1}^D \lambda_i b_i = 0$ iff $\lambda_i = 0$ for all $i \in [D]$. That is, in a linearly independent vector set, no vector of the set can be represented as a linear combination of the remaining vectors. A vector set is linearly independent maximal when we cannot add another linearly independent vector. Tangent spaces carry algebraic structures of vector spaces. Furthermore, to any vector space V, we can associate a dual covector space V^* which is the vector space of real-valued linear mappings. We do not enter into details here to preserve this gentle introduction to information geometry with as little intricacy as possible. Using local coordinates on a chart $(\mathcal{U}, x)$, the vector field X can be expressed as $X = \sum_{i=1}^D X^i e_i \overset{\Sigma}{=} X^i e_i$ using Einstein summation convention on dummy indices (using notation $\overset{\Sigma}{=}$), where $(X)_B := (X^i)$ denotes the *contravariant vector components* (manipulated as "column vectors" in algebra) in the *natural basis* $B = \{e_1 = \partial_1, \ldots, e_D = \partial_D\}$ with $\partial_i :=: \frac{\partial}{\partial x_i}$. A tangent plane (vector space) equipped with an *inner product* $\langle \cdot, \cdot \rangle$ yields an *inner product space*. We define 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 components* in the natural reciprocal basis. The primal and reciprocal basis are *mutually orthogonal* by construction as illustrated in Figure 2.

For any vector v, its contravariant components v^i's (superscript notation) and its covariant components v_i's (subscript notation) can be retrieved from v using the inner product with the use of the reciprocal and primal basis, respectively:

$$v^i = \langle v, e^{*i} \rangle, \tag{2}$$
$$v_i = \langle v, e_i \rangle. \tag{3}$$

The inner product defines a *metric tensor* g and a *dual metric tensor* g^*:

$$g_{ij} := \langle e_i, e_j \rangle, \tag{4}$$
$$g^{*ij} := \langle e^{*i}, e^{*j} \rangle. \tag{5}$$

Technically speaking, the metric tensor $g_p : T_pM \times T_pM \to \mathbb{R}$ is a 2-covariant tensor field:

$$g \overset{\Sigma}{=} g_{ij} \mathrm{d}x_i \otimes \mathrm{d}x_j, \tag{6}$$

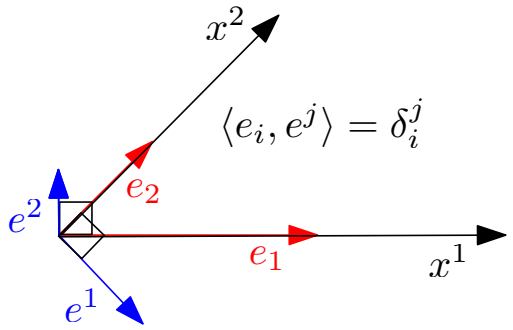


Figure 2: Primal basis (red) and reciprocal basis (blue) of an inner product $\langle \cdot, \cdot \rangle$ space. The primal/reciprocal basis are mutually orthogonal: e^1 is orthogonal to e_2, and e_1 is orthogonal to e^2.

where $\otimes$ is the dyadic tensor product performed on pairwise covector basis $\{\mathrm{d}x_i\}_i$ (the covectors corresponding to the reciprocal vector basis). We do not describe tensors in details for sake of brevity. A tensor is a geometric entity of a tensor space that can also be interpreted as a multilinear map. A contravariant vector lives in a vector space while a covariant vector lives in the dual covector space. We recommend the textbook [66] for a concise and well-explained description of tensors.

Let $G = [g_{ij}]$ and $G^* = [g^{*ij}]$ denote the $D \times D$ matrices It follows by construction of the reciprocal basis that $G^* = G^{-1}$. The reciprocal basis vectors e^{*i}'s and primal basis vectors e_i's can be expressed using the dual metric g^* and metric g on the primal basis vectors e_j's and reciprocal basis vectors e^{*j}'s, respectively:

$$e^{*i} \overset{\Sigma}{=} g^{*ij} e_j, \tag{7}$$

$$e_i \overset{\Sigma}{=} g_{ij} e^{*j}. \tag{8}$$

The *metric tensor field* g ("metric tensor" or "metric" for short) defines a smooth symmetric positive-definite *bilinear form* on the tangent bundle so that for $u, v \in T_p, \quad g(u, v) \geq 0 \in \mathbb{R}$. We can also write equivalently $g_p(u, v) := \langle u, v \rangle_p := \langle u, v \rangle_{g(p)} := \langle u, v \rangle$. Two vectors u and v are said orthogonal, denoted by $u \perp v$, iff $\langle u, v \rangle = 0$. The length of a vector is induced from the *norm* $\|u\|_p := \|u\|_{g(p)} = \sqrt{\langle u, u \rangle_{g(p)}}$. Using local coordinates of a chart $(\mathcal{U}, x)$, we get the vector contravariant/covariant components, and compute the metric tensor using matrix algebra (with column vectors by convention) as follows:

$$g(u, v) = (u)_B^\top \times G_{x(p)} \times (v)_B = (u)_{B^*}^\top \times G_{x(p)}^{-1} \times (v)_{B^*}, \tag{9}$$

since it follows from the primal/reciprocal basis that $G \times G^* = I$, the identity matrix. Thus on any tangent plane T_p, we get a *Mahalanobis distance*:

$$M_G(u, v) := \|u - v\|_G = \sqrt{\sum_{i=1}^{D} \sum_{j=1}^{D} G_{ij} (u^i - v^i)(u^j - v^j)}. \tag{10}$$

The inner product of two vectors u and v is a scalar (a 0-tensor) that can be equivalently calculated as:

$$\langle u, v \rangle := g(u, v) \overset{\Sigma}{=} u^i v_i \overset{\Sigma}{=} u_i v^i. \tag{11}$$

A metric tensor g of manifold M is said *conformal* when $\langle \cdot, \cdot \rangle_p = \kappa(p) \langle \cdot, \cdot \rangle_{\text{Euclidean}}$. That is, when the inner product is a scalar function $\kappa(\cdot)$ of the Euclidean dot product. More precisely, we define the notion of a metric g' conformal to another metric g when these metrics define the same angles between vectors u and v of a tangent plane T_p:

$$\frac{g'_p(u, v)}{\sqrt{g'_p(u, u)} \sqrt{g'_p(v, v)}} = \frac{g_p(u, v)}{\sqrt{g_p(u, u)} \sqrt{g_p(v, v)}}. \tag{12}$$

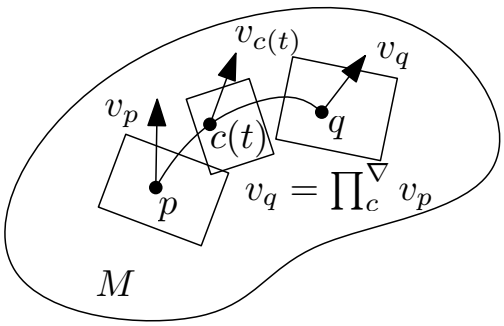


Figure 3: Illustration of the parallel transport of vectors on tangent planes along a smooth curve. For a smooth curve c, with $c(0) = p$ and $c(1) = q$, a vector $v_p \in T_p$ is parallel transported smoothly to a vector $v_q \in T_q$ such that for any $t \in [0, 1]$, we have $v_{c(t)} \in T_{c(t)}$.

Usually g' is chosen as the Euclidean metric. In conformal geometry, we can measure angles between vectors in tangent planes as if we were in an Euclidean space, without any deformation. This is handy for checking orthogonality in charts. For example, Poincaré disk model of hyperbolic geometry is conformal but Klein disk model is not conformal (except at the origin), see [89].

2.3 Affine connections ∇

An affine connection ∇ is a differential operator defined on a manifold that allows us to define (1) a covariant derivative of vector fields, (2) a parallel transport of vectors on tangent planes along a smooth curve, and (3) geodesics. Furthermore, an affine connection fully characterizes the curvature and torsion of a manifold.

2.3.1 Covariant derivatives $\nabla_X Y$ of vector fields

A connection defines a *covariant derivative* operator that tells us how to differentiate a vector field Y according to another vector field X. The covariant derivative operator is denoted using the traditional gradient symbol ∇. Thus a covariate derivative ∇ is a function:

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M), \tag{13}$$

that has its own special subscript notation $\nabla_X Y =: \nabla(X, Y)$ for indicating that it is differentiating a vector field Y according to another vector field X.

By prescribing D^3 smooth functions $\Gamma_{ij}^k = \Gamma_{ij}^k(p)$, called the *Christoffel symbols of the second kind*, we define the unique *affine connection* ∇ that satisfies in local coordinates of chart $(\mathcal{U}, x)$ the following equations:

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k. \tag{14}$$

The Christoffel symbols can also be written as $\Gamma_{ij}^k := (\nabla_{\partial_i} \partial_j)^k$, where $(\cdot)^k$ denote the k-th coordinate. The k-th component $(\nabla_X Y)^k$ of the covariant derivative of vector field Y with respect to vector field X is given by:

$$(\nabla_X Y)^k \overset{\Sigma}{=} X^i (\nabla_i Y)^k \overset{\Sigma}{=} X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma_{ij}^k Y^j \right). \tag{15}$$

The Christoffel symbols are *not* tensors (fields) because the transformation rules induced by a change of basis do not obey the tensor contravariant/covariant rules.

2.3.2 Parallel transport $\prod_c^\nabla$ along a smooth curve c

Since the manifold is not embedded[1] in a Euclidean space, we cannot add a vector $v \in T_p$ to a vector $v' \in T_{p'}$ as the tangent vector spaces are unrelated to each others without a connection.[2] Thus a *connection* ∇ defines how to associate vectors between infinitesimally close tangent planes T_p and $T_{p+\mathrm{d}p}$. Then the connection allows us to smoothly *transport* a vector $v \in T_p$ by sliding it (with infinitesimal moves) along a smooth curve $c(t)$ (with $c(0) = p$ and $c(1) = q$), so that the vector $v_p \in T_p$ "corresponds" to a vector $v_q \in T_q$: This is called the *parallel transport*. This mathematical prescription is necessary in order to study dynamics on manifolds (e.g., study the motion of a particle[3] on the manifold). We can express the parallel transport along the smooth curve c as:

$$\forall v \in T_p, \forall t \in [0,1], \quad v_{c(t)} = \prod_{c(0)\to c(t)}^{\nabla} v \in T_{c(t)} \tag{16}$$

The parallel transport is schematically illustrated in Figure 3.

2.3.3 ∇-geodesics γ_∇: Autoparallel curves

A connection ∇ allows one to define ∇-*geodesics* as autoparallel curves, that are curves γ such that we have:

$$\nabla_{\dot\gamma}\dot\gamma = 0. \tag{17}$$

That is, the *velocity vector* $\dot\gamma$ is moving along the curve parallel to itself (and all tangent vectors on the geodesics are mutually parallel): In other words, ∇-geodesics generalize the notion of "straight Euclidean" lines. In local coordinates $(\mathcal{U}, x)$, $\gamma(t) = (\gamma^k(t))_k$, the autoparallelism amounts to solve the following second-order Ordinary Differential Equations (ODEs):

$$\ddot\gamma(t) + \Gamma_{ij}^k \dot\gamma(t)\dot\gamma(t) = 0, \quad \gamma^l(t) = x^l \circ \gamma(t), \tag{18}$$

where Γ_{ij}^k are the *Christoffel symbols of the second kind*, with:

$$\Gamma_{ij}^k \overset{\Sigma}{=} \Gamma_{ij,l} g^{lk}, \quad \Gamma_{ij,k} \overset{\Sigma}{=} g_{lk}\Gamma_{ij}^l, \tag{19}$$

where $\Gamma_{ij,l}$ the *Christoffel symbols of the first kind*. Geodesics are 1D autoparallel submanifolds and ∇-hyperplanes are defined similarly as autoparallel submanifolds of dimension $D-1$. We may specify in subscript the connection that yields the geodesic γ: γ_∇.

The geodesic equation $\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0$ may be either solved as an *Initial Value Problem* (IVP) or as a *Boundary Value Problem* (BVP):

- Initial Value Problem (IVP): fix the conditions $\gamma(0) = p$ and $\dot\gamma(0) = v$ for some vector $v \in T_p$.
- Boundary Value Problem (BVP): fix the geodesic extremities $\gamma(0) = p$ and $\gamma(1) = q$.

2.3.4 Curvature and torsion of a manifold

An affine connection ∇ defines a 4D[4] *curvature tensor* R (expressed using components R_{jkl}^i of a $(1,3)$-tensor). The coordinate-free equation of the curvature tensor is given by:

$$R(X,Y)Z := \nabla_X\nabla_Y X - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z, \tag{20}$$

[1] Whitney embedding theorem states that any D-dimensional Riemannian manifold can be embedded into $\mathbb{R}^{2D}$.

[2] When embedded, we can implicitly use the ambient Euclidean connection ${}^{\mathrm{Euc}}\nabla$, see [2].

[3] Elie Cartan introduced the notion of affine connections [27, 3] in the 1920's motivated by the *principle of inertia* in mechanics: A point particle, without any force acting on it, shall move along a straight line with constant velocity.

[4] It follows from symmetry constraints that the number of independent components of the Riemann tensor is $\frac{D^2(D^2-1)}{12}$ in D dimensions.

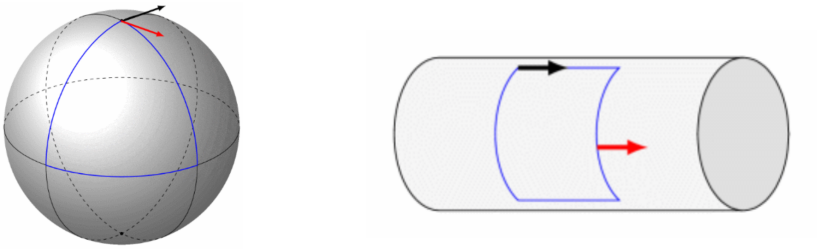


Figure 4: Parallel transport with respect to the metric connection: Curvature effect can be visualized as the angle defect along the parallel transport on smooth (infinitesimal) loops. For a sphere manifold, a vector parallel-transported along a loop does not coincide with itself, while it always conside with itself for a (flat) manifold. Drawings are courtesy of © CNRS, `http://images.math.cnrs.fr/Visualiser-la-courbure.html`

where $[X,Y](f) = X(Y(f)) - Y(X(f))$ $(\forall f \in \mathfrak{F}(M))$ is the *Lie bracket* of vector fields. When the connection is the metric Levi-Civita, the curvature is called *Riemann-Christoffel curvature tensor*. In a local coordinate system, we have:

$$R(\partial_j, \partial_k)\partial_i \overset{\Sigma}{=} R^l_{jki}\partial_l. \tag{21}$$

Informally speaking, the curvature tensor as defined in Eq. 20 quantifies the amount of non-commutativity of the covariant derivative.

A manifold M equipped with a connection ∇ is said *flat* (meaning ∇-flat) when $R = 0$. This holds in particular when finding a *particular*[5] coordinate system x of a chart $(\mathcal{U}, x)$ such that $\Gamma^k_{ij} = 0$, i.e., when all connection coefficients vanish.

A manifold is *torsion-free* when the connection is *symmetric*. A symmetric connection satisfies the following coordinate-free equation:

$$\nabla_X Y - \nabla_Y X = [X,Y]. \tag{22}$$

Using local chart coordinates, this amounts to check that $\Gamma^k_{ij} = \Gamma^k_{ji}$. The torsion tensor is a $(1,2)$-tensor defined by:

$$T(X,Y) := \nabla_X Y - \nabla_Y X - [X,Y]. \tag{23}$$

For a torsion-free connection, we have the first Bianchi identity:

$$R(X,Y)Z + R(Z,X)Y + R(Y,Z)X = 0, \tag{24}$$

and the second Bianchi identity:

$$(\nabla_V R)(X,Y)Z + (\nabla_X R)(Y,V)Z + (\nabla_Y R)(V,X)Z = 0. \tag{25}$$

In general, the parallel transport is *path-dependent*. The *angle defect* of a vector transported on an *infinitesimal closed loop* (a smooth curve with coinciding extremities) is related to the curvature. However for a *flat connection*, the parallel transport does not depend on the path, and yields *absolute parallelism geometry* [133]. Figure 4 illustrates the parallel transport along a curve for a curved manifold (the sphere manifold) and a flat manifold (the cylinder manifold[6]).

An affine connection is a torsion-free linear connection. Figure 5 summarizes the various concepts of differential geometry induced by an affine connection ∇ and a metric tensor g.

[5] For example, the Christoffel symbols vanish in a rectangular coordinate system of a plane but not in the polar coordinate system of it.

[6] The Gaussian curvature at of point of a manifold is the product of the minimal and maximal sectional curvatures: $\kappa_G := \kappa_{\min}\kappa_{\max}$. For a cylinder, since $\kappa_{\min} = 0$, it follows that the Gaussian curvature of a cylinder is 0. Gauss's Theorema Egregium (meaning "remarkable theorem") proved that the Gaussian curvature is intrinsic and does not depend on how the surface is embedded into the ambient Euclidean space.

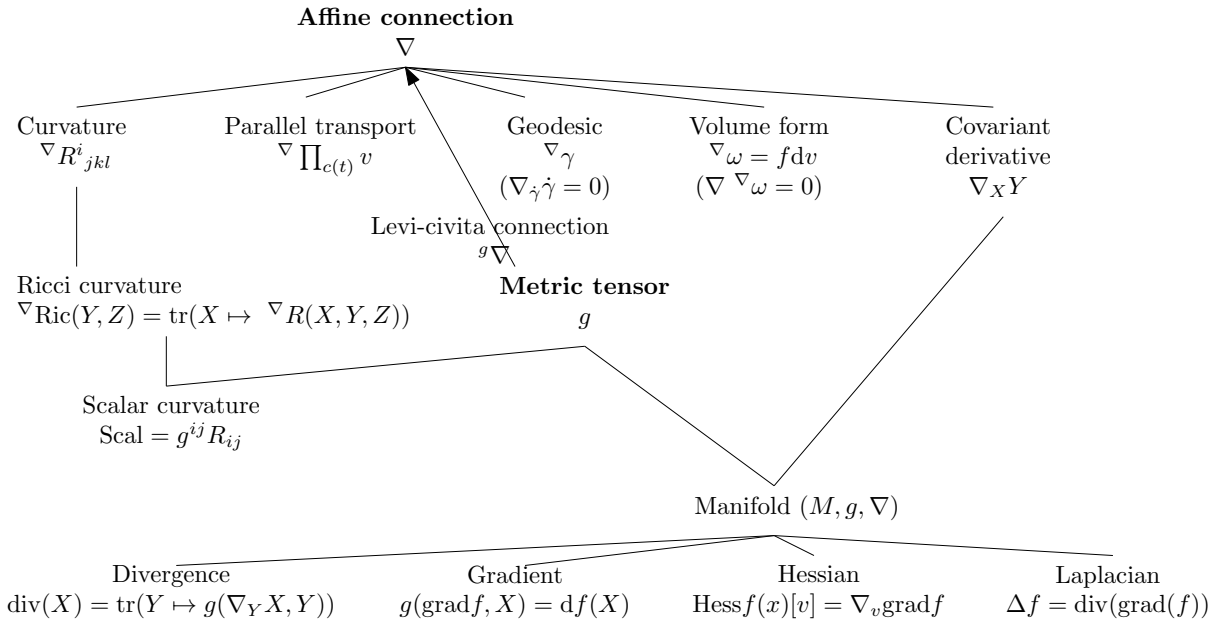


Figure 5: Differential-geometric concepts associated to an affine connection ∇ and a metric tensor g.

Curvature is a fundamental concept inherent to geometry [22]: There are several notions of curvatures: scalar curvature, sectional curvature, Gaussian curvature of surfaces to Riemannian-Christoffel 4-tensor, Ricci symmetric 2-tensor, synthetic Ricci curvature in Alexandrov geometry, etc.

2.4 The fundamental theorem of Riemannian geometry: The Levi-Civita metric connection

By definition, an affine connection ∇ is said *metric compatible* with g when it satisfies for any triple (X, Y, Z) of vector fields the following equation:

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \tag{26}$$

which can be written equivalently as:

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z) \tag{27}$$

Using local coordinates and natural basis $\{\partial_i\}$ for vector fields, the metric-compatibility property amounts to check that we have:

$$\partial_k g_{ij} = \langle \nabla_{\partial_k}\partial_i, \partial_j\rangle + \langle \partial_i, \nabla_{\partial_k}\partial_j\rangle \tag{28}$$

A property of using a metric-compatible connection is that the parallel transport $\prod^{\nabla}$ of vectors preserve the metric:

$$\langle u, v\rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla} v \right\rangle_{c(t)} \quad \forall t. \tag{29}$$

That is, the parallel transport preserves angles (and orthogonality) and lengths of vectors in tangent planes when transported along a smooth curve.

The fundamental theorem of Riemannian geometry states the existence of a unique torsion-free metric compatible connection:

Theorem 1 (Levi-Civita metric connection)**.** *There exists a unique torsion-free affine connection compatible with the metric called the* Levi-Civita connection ${}^{\mathrm{LC}}\nabla$.

The Christoffel symbols of the Levi-Civita connection can be expressed from the metric tensor g as follows:

$$ {}^{\mathrm{LC}}\Gamma_{ij}^{k} \overset{\Sigma}{=} \frac{1}{2} g^{kl} \left(\partial_i g_{il} + \partial_j g_{il} - \partial_l g_{ij} \right), \tag{30}$$

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

The Levi-Civita connection can also be defined coordinate-free with the *Koszul formula*:

$$ 2g(\nabla_X Y, Z) = X(g(Y,Z)) + Y(g(X,Z)) - Z(g(X,Y)) + g([X,Y],Z) - g([X,Z],Y) - g([Y,Z],X). \tag{31}$$

There exists metric-compatible connections with torsions studied in theoretical physics. See for example the flat Weitzenböck connection [15].

The metric tensor g induces the torsion-free metric-compatible Levi-Civita connection that determines the *local structure* of the manifold. However, the metric g does not fix the *global topological structure*: For example, although a cone and a cylinder have locally the same flat Euclidean metric, they exhibit different global structures.

2.5 Preview: Information geometry versus Riemannian geometry

In information geometry, we consider a pair of conjugate affine connections ∇ and ∇^* (often but not necessarily torsion-free) that are coupled to the metric g: The structure is conventionally written as (M, g, ∇, ∇^*). The key property is that those conjugate connections are metric compatible, and therefore the induced dual parallel transport preserves the metric:

$$ \langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)}. \tag{32}$$

Thus the Riemannian manifold (M, g) can be interpreted as the self-dual information-geometric manifold obtained for $\nabla = \nabla^* = {}^{\mathrm{LC}}\nabla$ the unique torsion-free Levi-Civita metric connection: $(M, g) \equiv (M, g, {}^{\mathrm{LC}}\nabla, {}^{\mathrm{LC}}\nabla^* = {}^{\mathrm{LC}}\nabla)$. However, let us point out that for a pair of self-dual Levi-Civita conjugate connections, the information-geometric manifold does not induce a distance. This contrasts with the Riemannian modeling (M, g) which provides a Riemmanian metric distance $D_\rho(p, q)$ defined by the length of the geodesic γ connecting the two points $p = \gamma(0)$ and $q = \gamma(1)$:

$$ D_\rho(p,q) := \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \mathrm{d}t = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))} \mathrm{d}t, \tag{33}$$

$$ = \int_0^1 \sqrt{\dot\gamma(t)^\top g_{\gamma(t)} \dot\gamma(t)} \mathrm{d}t. \tag{34}$$

This geodesic length distance $D_\rho(p, q)$ can also be interpreted as the shortest path linking point p to point q: $D_\rho(p,q) = \inf_\gamma \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \mathrm{d}t$ (with $p = \gamma(0)$ and $q = \gamma(1)$).

Usually, this Riemannian geodesic distance is not available in closed-form (and need to be approximated or bounded) because the geodesics cannot be explicitly parameterized (see geodesic shooting methods [11]).

We are now ready to introduce the key geometric structures of information geometry.

3 Information manifolds

3.1 Overview

In this part, we explain the *dualistic structures* of manifolds in information geometry. In §3.2, we first present the core *Conjugate Connection Manifolds* (CCMs) (M, g, ∇, ∇^*), and show how to build *Statistical Manifolds*

(SMs) (M, g, C) from a CCM in §3.3. From any statistical manifold, we can build a 1-parameter family $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ of CCMs, the information α-manifolds. We state the fundamental theorem of information geometry in §3.5. These CCMs and SMs structures are not related to any distance *a priori* but require at first a pair (∇, ∇^*) of conjugate connections coupled to a metric tensor g. We show two methods to build an initial pair of conjugate connections. A first method consists in building a pair of conjugate connections $({}^{D}\nabla, {}^{D}\nabla^*)$ from any divergence D in §3.6. Thus we obtain self-conjugate connections when the divergence is symmetric: $D(\theta_1 : \theta_2) = D(\theta_2 : \theta_1)$. When the divergences are Bregman divergences (i.e., $D = B_F$ for a strictly convex and differentiable Bregman generator), we obtain Dually Flat Manifolds (DFMs) $(M, \nabla^2 F, {}^{F}\nabla, {}^{F}\nabla^*)$ in §3.7. DFMs nicely generalize the Euclidean geometry and exhibit Pythagorean theorems. We further characterize when orthogonal ${}^{F}\nabla$-projections and dual ${}^{F}\nabla^*$-projections of a point on submanifold a is unique.[7] A second method to get a pair of conjugate connections $({}^{e}\nabla, {}^{m}\nabla)$ consists in defining these connections from a regular parametric family of probability distributions $\mathcal{P} = \{p_\theta(x)\}_\theta$. In that case, these 'e'xponential connection ${}^{e}\nabla$ and 'm'ixture connection ${}^{m}\nabla$ are coupled to the Fisher information metric ${}_{\mathcal{P}}g$. A statistical manifold $(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}C)$ can be recovered by considering the skewness Amari-Chentsov cubic tensor ${}_{\mathcal{P}}C$, and it follows a 1-parameter family of CCMs, $(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha})$, the statistical expected α-manifolds. In this parametric statistical context, these information manifolds are called *expected information manifolds* because the various quantities are expressed from statistical expectations $E.[\cdot]$. Notice that these information manifolds can be used in information sciences in general, beyond the traditional fields of statistics. In statistics, we motivate the choice of the connections, metric tensors and divergences by studying statistical invariance criteria, in §3.10. We explain how to recover the expected α-connections from standard f-divergences that are the only separable divergences that satisfy the property of information monotonicity. Finally, in §3.11, the recall the Fisher-Rao expected Riemannian manifolds that are Riemannian manifolds $(\mathcal{P}, {}_{\mathcal{P}}g)$ equipped with a geodesic metric distance called the Fisher-Rao distance, or Rao distance for short.

3.2 Conjugate connection manifolds: (M, g, ∇, ∇^*)

We begin with a definition:

Definition 1 (Conjugate connections)**.** *A connection ∇^* is said to be conjugate to a connection ∇ with respect to the metric tensor g if and only if we have for any triple (X, Y, Z) of smooth vector fields the following identity satisfied:*

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle, \quad \forall X, Y, Z \in \mathfrak{X}(M). \tag{35}$$

We can notationally rewrite Eq. 35 as:

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z), \tag{36}$$

and further explicit that for each point $p \in M$, we have:

$$X_p g_p(Y_p, Z_p) = g_p((\nabla_X Y)_p, Z_p) + g_p(Y_p, (\nabla^*_X Z)_p). \tag{37}$$

We check that the right-hand-side is a scalar and that the left-hand-side is a directional derivative of a real-valued function, that is also a scalar.

Conjugation is an involution: $(\nabla^*)^* = \nabla$.

Definition 2 (Conjugate Connection Manifold)**.** *The structure of the Conjugate Connection Manifold (CCM) is denoted by (M, g, ∇, ∇^*), where (∇, ∇^*) are conjugate connections with respect to the metric g.*

A remarkable property is that the dual parallel transport of vectors preserves the metric. That is, for any smooth curve $c(t)$, the inner product is conserved when we transport one of the vector u using the primal parallel transport $\prod_c^{\nabla}$ and the other vector v using the dual parallel transport $\prod_c^{\nabla^*}$.

[7] In Euclidean geometry, the orthogonal projection of a point p onto an affine subspace S is proved to be unique using the Pythagorean theorem.

$$\langle u, v\rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)}. \tag{38}$$

Property 1 (Dual parallel transport preserves the metric)**.** *A pair (∇, ∇^*) of conjugate connections preserves the metric g if and only if:*

$$\forall t \in [0,1], \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)} = \langle u, v\rangle_{c(0)}. \tag{39}$$

Property 2. *Given a connection ∇ on (M, g) (i.e., a structure (M, g, ∇)), there exists a unique conjugate connection ∇^* (i.e., a dual structure (M, g, ∇^*)).*

We consider a manifold M equipped with a pair of conjugate connections ∇ and ∇^* that are coupled with the metric tensor g so that the dual parallel transport preserves the metric. We define the mean connection $\bar{\nabla}$:

$$\bar{\nabla} = \frac{\nabla + \nabla^*}{2}, \tag{40}$$

with corresponding Christoffel coefficients denoted by $\bar{\Gamma}$. This mean connection coincides with the Levi-Civita metric connection:

$$\bar{\nabla} = {}^{\mathrm{LC}}\nabla. \tag{41}$$

Property 3. *The mean connection $\bar{\nabla}$ is self-conjugate, and coincide with the Levi-Civita metric connection.*

3.3 Statistical manifolds: (M, g, C)

Lauritzen introduced this corner structure [62] of information geometry in 1987. Beware that although it bears the name "statistical manifold," it is a purely geometric construction that may be used outside of the field of Statistics. However, as we shall mention later, we can always find a *statistical model* $\mathcal{P}$ corresponding to a statistical manifold [128]. We shall see how we can convert a conjugate connection manifold into such a statistical manifold, and how we can subsequently derive an infinite family of CCMs from a statistical manifold. In other words, once we have a pair of conjugate connections, we will be able to build a family of pairs of conjugate connections.

We define a *totally symmetric*[8] cubic $(0,3)$-tensor (i.e., 3-covariant tensor) called the *Amari-Chentsov tensor*:

$$C_{ijk} := \Gamma_{ij}^{k} - \Gamma^{*}{}_{ij}^{k}, \tag{42}$$

or in coordinate-free equation:

$$C(X, Y, Z) := \langle \nabla_X Y - \nabla_X^* Y, Z\rangle. \tag{43}$$

Using the local basis, this cubic tensor can be expressed as:

$$C_{ijk} = C(\partial_i, \partial_j, \partial_k) = \langle \nabla_{\partial_i}\partial_j - \nabla^*_{\partial_i}\partial_j, \partial_k\rangle \tag{44}$$

Definition 3 (Statistical manifold [62])**.** *A statistical manifold (M, g, C) is a manifold M equipped with a metric tensor g and a totally symmetric cubic tensor C.*

[8] This means that $C_{ijk} = C_{\sigma(i)\sigma(j)\sigma(k)}$ for any permutation σ. The metric tensor is totally symmetric.

3.4 A family $\{(M, g, \nabla^{-\alpha}, \nabla^{\alpha} = (\nabla^{-\alpha})^*)\}_{\alpha\in\mathbb{R}}$ of conjugate connection manifolds

For any pair (∇, ∇^*) of conjugate connections, we can define a 1-*parameter family of connections* $\{\nabla^{\alpha}\}_{\alpha\in\mathbb{R}}$, called the α-*connections* such that $(\nabla^{-\alpha}, \nabla^{\alpha})$ are dually coupled to the metric, with $\nabla^0 = \bar{\nabla} = {}^{\mathrm{LC}}\nabla$, $\nabla^1 = \nabla$ and $\nabla^{-1} = \nabla^*$. By observing that the scaled cubic tensor αC is also a totally symmetric cubic 3-covariant tensor, we can derive the α-connections from a statistical manifold (M, g, C) as:

$$\Gamma^{\alpha}_{ij,k} = \Gamma^0_{ij,k} - \frac{\alpha}{2} C_{ij,k}, \tag{45}$$

$$\Gamma^{-\alpha}_{ij,k} = \Gamma^0_{ij,k} + \frac{\alpha}{2} C_{ij,k}, \tag{46}$$

where $\Gamma^0_{ij,k}$ are the Levi-Civita Christoffel symbols, and $\Gamma_{ki,j} \overset{\Sigma}{=} \Gamma^l_{ij} g_{lk}$ (by index juggling).

The α-connection ∇^{α} can also be defined as follows:

$$g(\nabla^{\alpha}_X Y, Z) = g({}^{\mathrm{LC}}\nabla_X Y, Z) + \frac{\alpha}{2} C(X, Y, Z), \forall X, Y, Z \in \mathfrak{X}(M). \tag{47}$$

Theorem 2 (Family of information α-manifolds)**.** *For any $\alpha \in \mathbb{R}$, $(M, g, \nabla^{-\alpha}, \nabla^{\alpha} = (\nabla^{-\alpha})^*)$ is a conjugate connection manifold.*

The α-connections ∇^{α} can also be constructed directly from a pair (∇, ∇^*) of conjugate connections by taking the following weighted combination:

$$\Gamma^{\alpha}_{ij,k} = \frac{1+\alpha}{2}\Gamma_{ij,k} + \frac{1-\alpha}{2}\Gamma^*_{ij,k}. \tag{48}$$

3.5 The fundamental theorem of information geometry: ∇ κ-curved $\Leftrightarrow$ ∇^* κ-curved

We now state the fundamental theorem of information geometry and its corollaries:

Theorem 3 (Dually constant curvature manifolds)**.** *If a torsion-free affine connection ∇ has constant curvature κ then its conjugate torsion-free connection ∇^* has necessarily the* same *constant curvature κ.*

The proof is reported in [25] (Proposition 8.1.4, page 226).

A statistical manifold (M, g, C) is said α-*flat* if its induced α-connection is flat. It can be shown that $R^{\alpha} = -R^{-\alpha}$.

We get the following two corollaries:

Corollary 1 (Dually α-flat manifolds)**.** *A manifold $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ is ∇^{α}-flat if and only if it is $\nabla^{-\alpha}$-flat.*

Corollary 2 (Dually flat manifolds ($\alpha = \pm 1$))**.** *A manifold (M, g, ∇, ∇^*) is ∇-flat if and only if it is ∇^*-flat.*

(See Theorem 3.3 of [9])

Let us now define the notion of constant curvature of a statistical structure [46]:

Definition 4 (Constant curvature κ)**.** *A statistical structure (M, g, ∇) is said of constant curvature κ when*

$$R^{\nabla}(X, Y)Z = \kappa\{g(Y, Z)X - g(X, Z)Y\}, \quad \forall X, Y, Z \in \Gamma(TM),$$

where $\Gamma(TM)$ denote the space of smooth vector fields.

It can be proved that the Riemann-Christoffel (RC) 4-tensors of conjugate α-connections [25] are related as follows:

$$g\left(R^{(\alpha)}(X, Y)Z, W\right) + g\left(Z, R^{(-\alpha)}(X, Y)W\right) = 0. \tag{49}$$

Thus we have $g\left(R^{\nabla^*}(X, Y)Z, W\right) = -g\left(Z, R^{\nabla}(X, Y)W\right)$.

Thus once we are given a pair of conjugate connections, we can always build a 1-parametric family of manifolds. Manifolds with constant curvature κ are interesting from the computational viewpoint as dual geodesics have simple closed-form expressions.

3.6 Conjugate connections from divergences: $(M, D) \equiv (M,\ {}^{D}g,\ {}^{D}\nabla,\ {}^{D}\nabla^* = {}^{D^*}\nabla)$

Loosely speaking, a divergence $D(\cdot : \cdot)$ is a smooth distance [138], potentially asymmetric. In order to define precisely a divergence, let us first introduce the following handy notations: $\partial_{i,\cdot} f(x,y) = \frac{\partial}{\partial x^i} f(x,y)$, $\partial_{\cdot,j} f(x,y) = \frac{\partial}{\partial y^j} f(x,y)$, $\partial_{ij,k} f(x,y) = \frac{\partial^2}{\partial x^i \partial x^j} \frac{\partial}{\partial y^k} f(x,y)$ and $\partial_{i,jk} f(x,y) = \frac{\partial}{\partial x^i} \frac{\partial^2}{\partial y^j \partial y^k} f(x,y)$, etc.

Definition 5 (Divergence)**.** *A* divergence $D : M \times M \to [0, \infty)$ *on a manifold* M *with respect to a local chart* $\Theta \subset \mathbb{R}^D$ *is a* C^3*-function satisfying the following properties:*

1. $D(\theta : \theta') \geq 0$ *for all* $\theta, \theta' \in \Theta$ *with equality holding iff* $\theta = \theta'$ *(law of the indiscernibles),*
2. $\partial_{i,\cdot} D(\theta : \theta')|_{\theta=\theta'} = \partial_{\cdot,j} D(\theta : \theta')|_{\theta=\theta'} = 0$ *for all* $i, j \in [D]$*,*
3. $-\partial_{\cdot,i}\partial_{\cdot,j} D(\theta : \theta')|_{\theta=\theta'}$ *is positive-definite.*

The *dual divergence* is defined by swapping the arguments:

$$D^*(\theta : \theta') := D(\theta' : \theta), \tag{50}$$

and is also called the *reverse divergence* (reference duality in information geometry). Reference duality of divergences is an involution: $(D^*)^* = D$.

The Euclidean distance is a metric distance but not a divergence. The squared Euclidean distance is a non-metric symmetric divergence. The metric tensor g yields Riemannian metric distance D_ρ but it is never a divergence.

From any given divergence D, we can define a conjugate connection manifold following the construction of Eguchi [42, 43] (1983):

Theorem 4 (Manifold from divergence)**.** $(M, {}^{D}g, {}^{D}\nabla, {}^{D^*}\nabla)$ *is an information manifold with:*

$$\begin{aligned} {}^{D}g &:= -\partial_{i,j} D(\theta : \theta')|_{\theta=\theta'} = {}^{D^*}g, & (51)\\ {}^{D}\Gamma_{ijk} &:= -\partial_{ij,k} D(\theta : \theta')|_{\theta=\theta'}, & (52)\\ {}^{D^*}\Gamma_{ijk} &:= -\partial_{k,ij} D(\theta : \theta')|_{\theta=\theta'}. & (53) \end{aligned}$$

The associated statistical manifold is $(M, {}^{D}g, {}^{D}C)$ with:

$${}^{D}C_{ijk} = {}^{D^*}\Gamma_{ijk} - {}^{D}\Gamma_{ijk}. \tag{54}$$

Since $\alpha^D C$ is a totally symmetric cubic tensor for any $\alpha \in \mathbb{R}$, we can derive a one-parameter family of conjugate connection manifolds:

$$\left\{ (M, {}^{D}g, {}^{D}C^{\alpha}) \equiv (M, {}^{D}g, {}^{D}\nabla^{-\alpha}, ({}^{D}\nabla^{-\alpha})^* = {}^{D}\nabla^{\alpha}) \right\}_{\alpha \in \mathbb{R}}. \tag{55}$$

In the remainder, we use the shortcut (M, D) to denote the divergence-induced information manifold $(M, {}^{D}g, {}^{D}\nabla, {}^{D}\nabla^*)$. Notice that it follows from construction that:

$${}^{D}\nabla^* = {}^{D^*}\nabla. \tag{56}$$

3.7 Dually flat manifolds (Bregman geometry): $(M, F) \equiv (M,\ {}^{B_F}g,\ {}^{B_F}\nabla,\ {}^{B_F}\nabla^* = {}^{B_{F^*}}\nabla)$

We consider dually flat manifolds that satisfy asymmetric Pythagorean theorems. These flat manifolds can be obtained from a canonical Bregman divergence.

Consider a *strictly convex smooth function* $F(\theta)$ called a *potential function*, with $\theta \in \Theta$ where Θ is an open convex domain. Notice that the function convexity does not change by an affine transformation. We associate to the potential function F a corresponding *Bregman divergence* (parameter divergence):

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta'). \tag{57}$$

We write also the Bregman divergence between point P and point Q as $D(P:Q) := B_F(\theta(P):\theta(Q))$, where $\theta(P)$ denotes the coordinates of a point P.

The information-geometric structure induced by a Bregman generator[9] is $(M, {}^F g, {}^F C) := (M, {}^{B_F} g, {}^{B_F} C)$ with:

$$
\begin{aligned}
{}^F g &:= {}^{B_F} g = -\left[\partial_i \partial_j B_F(\theta:\theta')|_{\theta'=\theta}\right] = \nabla^2 F(\theta), && (58)\\
{}^F \Gamma &:= {}^{B_F}\Gamma_{ij,k}(\theta) = 0, && (59)\\
{}^F C_{ijk} &:= {}^{B_F} C_{ijk} = \partial_i \partial_j \partial_k F(\theta). && (60)
\end{aligned}
$$

Since all coefficients of the Christoffel symbols vanish (Eq. 59), the information manifold is ${}^F\nabla$-*flat*. The Levi-Civita connection ${}^{\mathrm{LC}}\nabla$ 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 yields the *convex conjugate* F^* that is interpreted as the *dual potential function*:

$$F^*(\eta) := \sup_{\theta\in\Theta}\{\theta^\top \eta - F(\theta)\}. \quad (61)$$

Theorem 5 (Fenchel-Moreau biconjugation [52])**.** *If F is a lower semicontinuous[10] and convex function, then its Legendre-Fenchel transformation is involutive: $(F^*)^* = F$ (biconjugation).*

In a dually flat manifold, there exists two global dual affine coordinate systems $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$, and therefore the manifold can be covered by a single chart. Thus if a probability family belongs to an exponential family then its natural parameters cannot belong to, say, a spherical space (that requires at least two charts).

We have the Crouzeix [32] identity relating the Hessians of the potential functions:

$$\nabla^2 F(\theta)\nabla^2 F^*(\eta) = I, \quad (62)$$

where I denote the $D \times D$ identity matrix. This Crouzeix identity reveals that $B = \{\partial_i\}_i$ and $B^* = \{\partial^j\}_j$ are the primal and reciprocal basis, respectively.

The Bregman divergence can be reinterpreted using Young-Fenchel (in)equality as the *canonical divergence* A_{F,F^*} [12]:

$$B_F(\theta:\theta') = A_{F,F^*}(\theta:\eta') = F(\theta) + F^*(\eta') - \theta^\top \eta' = A_{F^*,F}(\eta':\theta). \quad (63)$$

The *dual Bregman divergence* ${B_F}^*(\theta:\theta') := B_F(\theta':\theta) = B_{F^*}(\eta:\eta')$ yields

$$
\begin{aligned}
{}^F g^{ij}(\eta) &= \partial^i \partial^j F^*(\eta), \quad \partial^l :=: \frac{\partial}{\partial \eta^l} && (64)\\
{}^F {\Gamma^*}^{ijk}(\eta) &= 0, \quad {}^F C^{ijk} = \partial^i \partial^j \partial^k F^*(\eta) && (65)
\end{aligned}
$$

Thus the information manifold is both ${}^F\nabla$-flat and ${}^F\nabla^*$-flat: This structure is called a *dually flat manifold* (DFM). In a DFM, we have two global affine coordinate systems $\theta(\cdot)$ and $\eta(\cdot)$ related by the Legendre-Fenchel transformation of a pair of potential functions F and F^*. That is, $(M, F) \equiv (M, F^*)$, and the dual atlases are $\mathcal{A} = \{(M, \theta)\}$ and $\mathcal{A}^* = \{(M, \eta)\}$.

In a dually flat manifold, any pair of points P and Q can either be linked using the ∇-geodesic (that is θ-straight) or the ∇^*-geodesic (that is η-straight). In general, there are $2^3 = 8$ types of *geodesic triangles* in a dually flat manifold.

[9] Here, we define a Bregman generator as a proper, lower semi-continuous, and strictly convex and C^3 differentiable real-valued function.

[10] A function f is lower semicontinous (lsc) at x_0 iff $f(x_0) \leq \lim_{x\to x_0} \inf f(x)$. A function f is lsc if it is lsc at x for all x in the function domain.

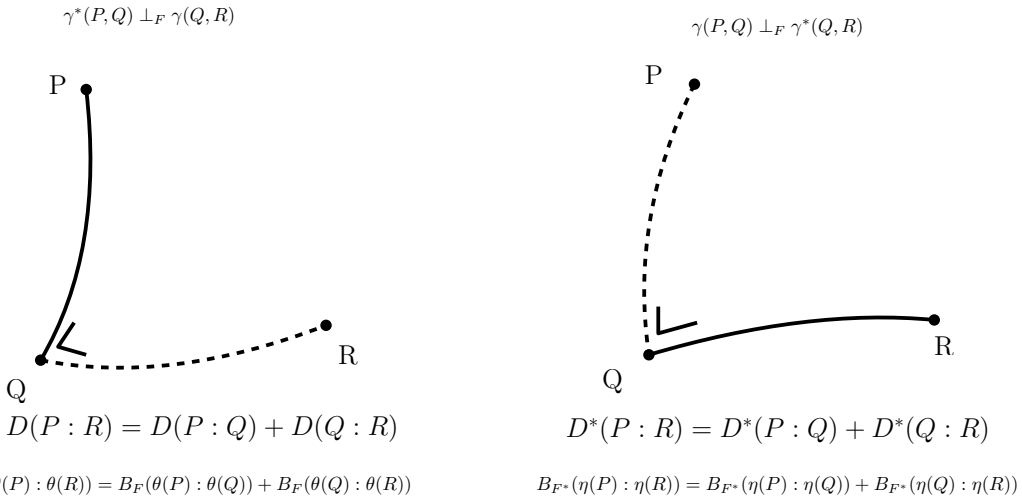


Figure 6: Dual Pythagorean theorems in a dually flat space.

on a Bregman manifold, the primal parallel transport of a vector does not change the contravariant vector components, and the dual parallel transport does not change the covariant vector components. Because the dual connections are flat, the dual parallel transports are path-independent.

Moreover, the dual Pythagorean theorems [76] illustrated in Figure 6 holds. Let $\gamma(P,Q) = \gamma_\nabla(P,Q)$ denote the ∇-geodesic passing through points P and Q, and $\gamma^*(P,Q) = \gamma_{\nabla^*}(P,Q)$ denote the ∇^*-geodesic passing through points P and Q. Two curves γ_1 and γ_2 are orthogonal at point $p = \gamma_1(t_1) = \gamma_2(t_2)$ with respect to the metric tensor g when $g(\dot{\gamma}_1(t_1), \dot{\gamma}_2(t_2)) = 0$.

Theorem 6 (Dual Pythagorean identities)**.**

$$\begin{aligned}
\gamma^*(P,Q) \perp \gamma(Q,R) \quad &\Leftrightarrow \quad (\eta(P) - \eta(Q))^\top (\theta(Q) - \theta(R)) \overset{\Sigma}{=} (\eta_i(P) - \eta_i(Q))(\theta_i(Q) - \theta_i(R)) = 0, \\
\gamma(P,Q) \perp \gamma^*(Q,R) \quad &\Leftrightarrow \quad (\theta(P) - \theta(Q))^\top (\eta(Q) - \eta(R)) \overset{\Sigma}{=} (\theta_i(P) - \theta_i(Q))^\top (\eta_i(Q) - \eta_i(R)) = 0.
\end{aligned}$$

We can define dual Bregman projections and characterize when these projections are unique: A *submanifold* $S \subset M$ is said ∇-flat (∇^*-flat) iff. it corresponds to an *affine subspace* in the θ-coordinate system (in the η-coordinate system, respectively).

Theorem 7 (Uniqueness of projections)**.** *The* ∇-projection P_S *of* P *on* S *is* unique *if* S *is* ∇^*-flat *and minimizes the divergence* $D(\theta(P):\theta(Q))$*:*

$$\nabla\text{-projection:} \quad P_S = \arg\min_{Q \in S} D(\theta(P):\theta(Q)). \tag{66}$$

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

$$\nabla^*\text{-projection:} \quad P_S^* = \arg\min_{Q \in S} D(\theta(Q):\theta(P)). \tag{67}$$

Let $S \subset M$ and $S' \subset M$, then we define the divergence between S and S' as

$$D(S:S') := \min_{s \in S, s' \in S'} D(s:s'). \tag{68}$$

When S is a ∇-flat submanifold and S' ∇^*-flat submanifold, the divergence $D(S:S')$ between submanifold S and submanifold S' can be calculated using the method of alternating projections [8]. Let us remark

that Kurose [61] reported a Pythagorean theorem for dually constant curvature manifolds that generalizes the Pythagorean theorems of dually flat spaces.

We shall concisely explain the *space of Bregman spheres* explained in details in [20]. Let D denote the dimension of Θ. We define the lifting of primal coordinates θ to the primal potential function $\mathcal{F} = \{\hat{\theta} = (\theta, \theta_{D+1} = F(\theta)) \ : \ \theta \in \Theta\}$ using an extra dimension θ_{D+1}. A Bregman ball Σ

$$\mathrm{Ball}_F(C : r) := \{P \text{ such that } F(\theta(P)) + F^*(\eta(C)) - \langle \theta(P), \eta(C) \rangle \leq r\} \tag{69}$$

can then be lifted to $\mathcal{F}$: $\hat{\Sigma} = \{\hat{\theta}(P) \ : \ P \in \sigma\}$. The boundary Bregman sphere $\sigma = \partial\Sigma$ is lifted to $\partial\hat{\Sigma} = \hat{\sigma}$, and the lifted points are all supported by a supporting $(D+1)$-dimensional hyperplane (of dimension D):

$$H_{\hat{\sigma}} : \theta_{D+1} = \langle \theta - \theta(C), \eta(C) \rangle + F(\theta(C)) + r. \tag{70}$$

Let $H_{\hat{\sigma}}^-$ denotes the halfspaces bounded by $H_{\hat{\sigma}}$ and containing $\hat{\theta}(C) = (\theta(C), F(\theta(C)))$. A point P belongs to a Bregman ball Σ iff $\hat{\theta(P)} \in H_{\hat{\sigma}}^-$, see [20]. Reciprocally, a $(D+1)$-dimensional hyperplane $H : \theta_{D+1} = \langle \theta, \eta_a \rangle + b$ cutting the potential function $\mathcal{F}$ yields a Bregman sphere σ_H of center C with $\theta(C) = \nabla F^*(\eta_a)$ and radius $r = \langle \nabla F^*(\eta_a), \eta_a \rangle - F(\theta_a) + b = F^*(\eta_a) + b$, where $\theta_a = \nabla F^*(\eta_a)$. It follows that the intersection of k Bregman balls is a $(D-k)$-dimensional Bregman ball, and that a Bregman sphere can be defined by $D+1$ points in general position since an hyperplane in the augmented space is defined by $D+1$ points. We can test whether a point P belongs to a Bregman ball with bounding Bregman sphere passing through $D+1$ points $P_1, \ldots, P_{D+1}$ or not by checking the sign of a $(D+2) \times (D+2)$ determinant:

$$\mathrm{InBregmanBall}_F(P_1, \ldots, P_{d+1}; P) := \mathrm{sign} \begin{pmatrix} \left| \begin{array}{cccc} 1 & \ldots & 1 & 1 \\ \theta(P_1) & \ldots & \theta(P_{D+1}) & \theta(P) \\ F(\theta(P_1)) & \ldots & F(\theta(P_{D+1})) & F(\theta(P)) \end{array} \right| \end{pmatrix}. \tag{71}$$

We have:

$$\mathrm{InBregmanBall}_F(P_1, \ldots, P_{d+1}; P) : \begin{cases} = -1 & \Leftrightarrow \quad P \in \mathrm{InBregmanBall}_F^\circ(P_1, \ldots, P_{D+1}; P) \\ = 0 & \Leftrightarrow \quad P \in \partial\mathrm{InBregmanBall}_F(P_1, \ldots, P_{D+1}; P) \\ = +1 & \Leftrightarrow \quad P \notin \mathrm{InBregmanBall}_F(P_1, \ldots, P_{D+1}; P) \end{cases} \tag{72}$$

Similarly, a dual-type Bregman ball Σ^* can be defined by

$$\mathrm{Ball}_F^*(C : r) := \{P \text{ such that } F(\theta(C)) + F^*(\eta(P)) - \langle \theta(C), \eta(P) \rangle \leq r\}, \tag{73}$$

and be lifted to the dual potential function $\mathcal{F}^*$. Notice that $\mathrm{Ball}_F^*(C : r) = \mathrm{Ball}_{F^*}(C : r)$.

In general, we have the following quadrilateral relation for Bregman divergences:

Property 4 (Bregman 4-parameter property [37])**.** *For any four points P_1, P_2, Q_1, Q_2, we have the following identity:*

$$\begin{aligned} B_F(\theta(P_1) : \theta(Q_1)) + B_F(\theta(P_2) : \theta(Q_2)) \qquad & -B_F(\theta(P_1) : \theta(Q_2)) - B_F(\theta(P_2) : \theta(Q_1)) \\ & -(\theta(P_2) - \theta(P_1))^\top (\eta(Q_1) - \eta(Q_2)) = 0. \end{aligned} \tag{74}$$

In summary, to define a dually flat space, we need a convex Bregman generator. When the α-geometries are neither dually flat (eg., Cauchy manifolds [79], we may still build a dually flat structure on the manifold by considering some Bregman generator (eg., Bregman-Tsallis generator for the dually flat Cauchy manifold [79]). The dually flat geometry can be investigated under the wider scope of *Hessian manifolds* [120] which consider *locally* potential functions. In general, a dually flat space can be built from any smooth strictly convex generator F. For example, a dually flat geometry can be built on homogeneous cones with the characteristic function F of the cone [120]. Figure 7 illustrates several common constructions of dually flat spaces.

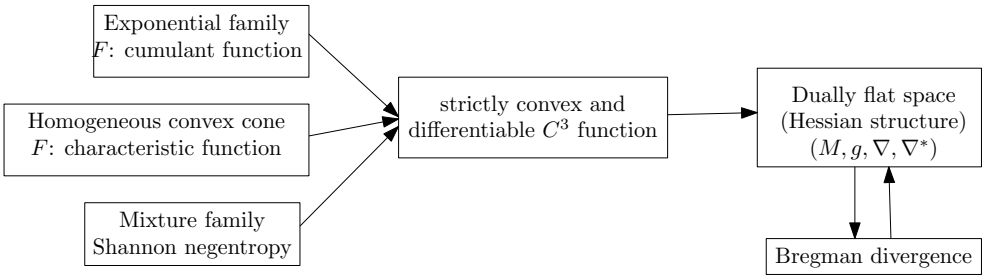


Figure 7: Common dually flat spaces associated to smooth and strictly convex generators.

3.8 Hessian α-geometry: $(M, F, \alpha) \equiv (M, {}^{F}g, {}^{F}\nabla^{-\alpha}, {}^{F}\nabla^{\alpha}$

The dually flat manifold is also called a manifold with a *Hessian structure* [120] induced by a convex potential function F. Since we built two dual affine connections ${}^{B_F}\nabla = {}^{F}\nabla$ and ${}^{B_F}\nabla^* = {}^{F}\nabla^* = {}^{F^*}\nabla$, we can build a family of α-geometry as follows:

$$ {}^{F}g_{ij}(\theta) = \partial_i \partial_j F(\theta), \quad {}^{F}g^{ij}(\eta) = \partial^i \partial^j F(\eta), \tag{75}$$

and

$$ {}^{F}\Gamma^{\alpha}_{ijk}(\theta) = \frac{1-\alpha}{2}\partial_i\partial_j\partial_k F(\theta), \quad {}^{F}\Gamma^{\alpha}_{ijk}{}^{*}(\eta) = {}^{F^*}\Gamma^{\alpha}_{ijk}(\eta) = \frac{1+\alpha}{2}\partial^i\partial^j\partial^k F^*(\eta). \tag{76}$$

Thus when $\alpha = \pm 1$, the Hessian α-geometry is dually flat.

We now consider information manifolds induced by parametric statistical models.

3.9 Expected α-manifolds of a family of parametric probability distributions: $(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{\alpha})$

Informally speaking, an *expected manifold* is an information manifold built on a regular parametric family of distributions. It is sometimes called "expected" manifold or "expected" geometry in the literature [140] because the components of the metric tensor g and the Amari-Chentsov cubic tensor C are expressed using statistical expectations $E_{\cdot}[\cdot]$.

Let $\mathcal{P}$ be a *parametric family* of probability distributions:

$$ \mathcal{P} := \{p_\theta(x)\}_{\theta \in \Theta}, \tag{77}$$

with θ belonging to the open parameter space Θ. The order of the family is the dimension of its parameter space. We define the likelihood function[11] $L(\theta; x) := p_\theta(x)$ as a function of θ, and its corresponding *log-likelihood* function:

$$ l(\theta; x) := \log L(\theta; x) = \log p_\theta(x). \tag{78}$$

The *score vector*:

$$ s_\theta = \nabla_\theta l = (\partial_i l)_i, \tag{79}$$

indicates the sensitivity of the likelihood $\partial_i l =: \frac{\partial}{\partial \theta_i} l(\theta; x)$.

The *Fisher information matrix* (FIM) of $D \times D$ for $\dim(\Theta) = D$ is defined by:

$$ {}_{\mathcal{P}}I(\theta) := E_\theta \left[\partial_i l \partial_j l\right]_{ij} \succeq 0, \tag{80}$$

where $\succeq$ denotes the Löwner order. That is, for two symmetric positive-definite matrices A and B, $A \succeq B$ if and only if matrix $A - B$ is positive semidefinite. For regular models [25], the FIM is positive definite: ${}_{\mathcal{P}}I(\theta) \succ 0$, where $A \succ B$ if and only if matrix $A - B$ is positive-definite.

[11] The likelihood function is an *equivalence class* of functions defined modulo a positive scaling factor.

The FIM is invariant by reparameterization of the sample space $\mathcal{X}$, and covariant by reparameterization of the parameter space Θ, see [25]. That is, let $\bar{p}(x;\eta) = p(\theta(\eta);x)$. Then we have:

$$\bar{I}(\eta) = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]^\top_{ij} I(\theta(\eta)) \left[\frac{\partial \theta_i}{\partial \eta_j}\right]_{ij}. \tag{81}$$

Matrix $J_{ij} = \left[\frac{\partial \theta_i}{\partial \eta_j}\right]_{ij}$ is the Jacobian matrix.

Example 1. *For example, consider the family*

$$\mathcal{N} = \left\{ p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})) \ : \ (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_{++} \right\} \tag{82}$$

of univariate normal distributions. The 2D parameter vector is $\lambda = (\mu,\sigma)$ with μ denoting the mean and σ the standard deviation. Another common parameterization of the normal family is $\lambda' = (\mu,\sigma^2)$. The λ' parameterization extends naturally to d-variance normal distributions with $\lambda' = (\mu,\Sigma)$, where Σ denotes the covariance matrix (with $\Sigma = \sigma^2$ when $d = 1$). For multivariate normal distributions, the λ-parameterization can be interpreted as $\lambda = (\mu, L^\top)$ where $L^\top$ is the upper triangular matrix in the Cholesky decomposition (when $d = 1$, $L^\top = \sigma$). We have the following Fisher information matrices in the λ-parameterization and λ'-parameterization:

$$I_\lambda(\lambda) = \begin{bmatrix} \frac{1}{\lambda_2^2} & 0 \\ 0 & \frac{2}{\lambda_2^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \tag{83}$$

and

$$I_{\lambda'}(\lambda') = \begin{bmatrix} \frac{1}{\lambda_2} & 0 \\ 0 & \frac{1}{2\lambda_2^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \tag{84}$$

Since the FIM is covariant, we have the following the change of transformation:

$$I_{\lambda'}(\lambda') = J^\top_{\lambda,\lambda'} I_\lambda(\lambda(\lambda')) J_{\lambda,\lambda'}, \tag{85}$$

with

$$J_{\lambda',\lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} \tag{86}$$

Thus we check that

$$I_\lambda(\lambda) = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \tag{87}$$

Notice that the infinitesimal length elements are invariant: $\mathrm{d}s_\lambda = \mathrm{d}s_{\lambda'}$.

As a corollary, notice that we can recognize the Euclidean metric in any other coordinate system if the metric tensor g can be written $J^\top_{\lambda,\lambda'} J_{\lambda,\lambda'}$. For example, the Riemannian geometry induced by a dually flat space with a separable potential function is Euclidean [49].

In statistics, the FIM plays a role in the attainable precision of unbiased estimators. For any unbiased estimator, the Cramér-Rao lower bound [71] on the variance of the estimator is:

$$\mathrm{Var}_\theta[\hat{\theta}_n(X)] \succeq \frac{1}{n} {}_{\mathcal{P}} I^{-1}(\theta). \tag{88}$$

Figure 8 illustrates the Cramér-Rao lower bound (CRLB) for the univariate distributions: At regular grid locations (μ,σ) of the upper space of normal parameters, we repeat 200 runs (trials) of estimating the normal parameters $\widehat{(\mu,\sigma)}$ using the MLE on 100 iid samples $x_1,\ldots,x_n \sim N(\mu,\sigma)$. The sample mean and the sample covariance matrix are calculated for the number of trials and displayed as back ellipses. The Fisher

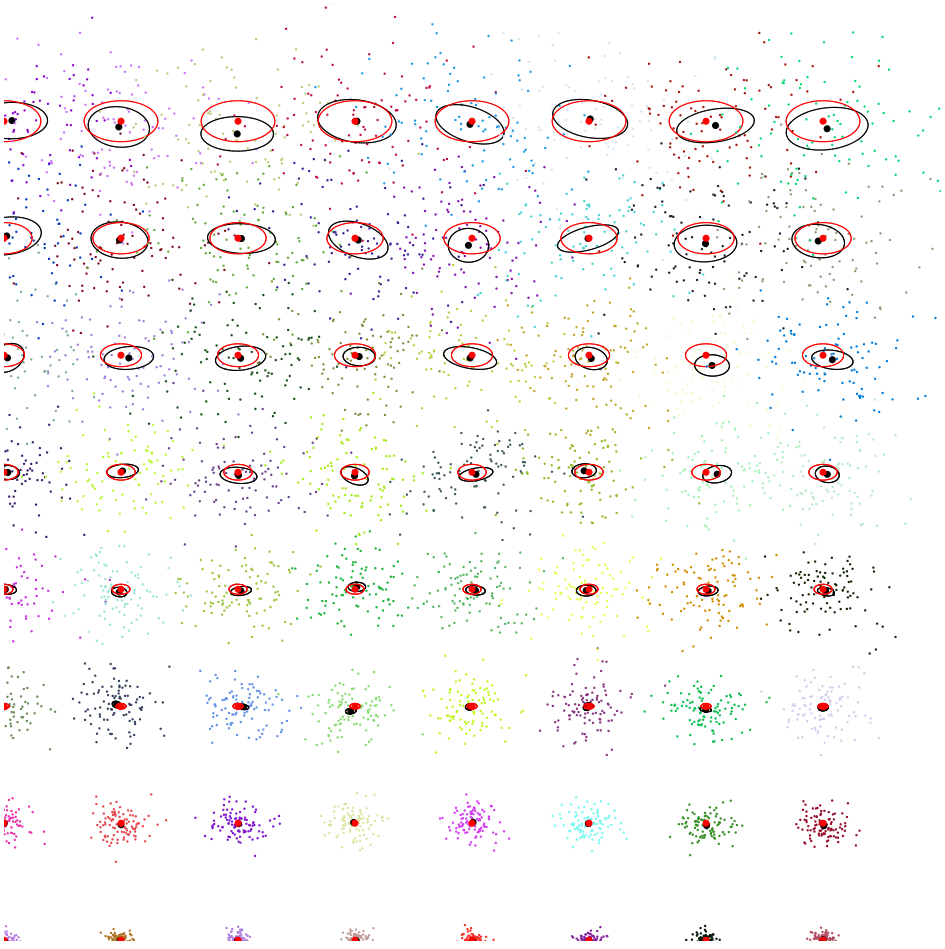


Figure 8: Visualizing the Cramér-Rao lower bound: The red ellipses display the Fisher information matrix of normal distributions at grid locations. The black ellipses are sample covariance matrices centered at the sample means calculated by repeating 200 runs of sampling 100 iid variates for the normal parameters of the grid.

information matrix is plotted as red ellipses at the grid locations: The red ellipses have semi-axis parallel to the coordinate system since the parameters μ and σ are orthogonal (diagonal FIM). This is not true anymore for the sample covariance matrix of the MLE estimator, and the centers of the sample covariance matrices deviate from the grid locations.

We report the expression of the FIM for two important generic parametric family of probability distributions: (1) an exponential family (with its prominent multivariate normal family), and (2) a mixture family.

Example 2 (FIM of an exponential family $\mathcal{E}$)**.** *An* exponential family *[84] $\mathcal{E}$ is defined for a sufficient statistic vector $t(x) = (t_1(x), \ldots, t_D(x))$, and an auxiliary carrier measure $k(x)$ by the following canonical density:*

$$\mathcal{E} = \left\{ p_\theta(x) = \exp\left(\sum_{i=1}^{D} t_i(x)\theta_i - F(\theta) + k(x) \right) \text{ such that } \theta \in \Theta \right\}, \tag{89}$$

where F is the strictly convex cumulant function (also called log-normalizer, and log partition function or free energy in statistical mechanics). Exponential families include the Gaussian family, the Gamma and Beta families, the probability simplex Δ, etc. The FIM of an exponential family is given by:

$$_{\mathcal{E}}I(\theta) = \mathrm{Cov}_{X\sim p_\theta(x)}[t(x)] = \nabla^2 F(\theta) = (\nabla^2 F^*(\eta))^{-1} \succ 0. \tag{90}$$

Natural parameters beyond vector types can also be used in the canonical decomposition of the density of an exponential family: For example, we may use a matrix type for defining the zero-centered multivariate Gaussian family or the Wishart family, a complex numbers for defining the complex-valued Gaussian distribution family, etc. We then replace the term $\sum_{i=1}^{D} t_i(x)\theta_i$ in Eq. 89 by an inner product defined for the natural parameter type (e.g., dot product for vectors, matrix product trace for matrices, etc). Furthermore, natural parameters can be of compound types: For example, the multivariate Gaussian distribution can be written using $\theta = (\theta_v, \theta_M)$ where θ_v is a vector part and θ_M a matrix part, see [84].

Let $\Sigma = [\sigma_{ij}]$ denote the covariance matrix *and $\Sigma^{-1} = [\sigma^{ij}]$ the* precision matrix *of a multivariate normal distribution. The Fisher information matrix of the multivariate Gaussian [114, 121] $N(\mu, \Sigma)$ is given by*

$$I(\mu, \Sigma) = \begin{array}{c} \begin{array}{cc} \mu & \Sigma = [\sigma_{ij}] \end{array} \\ \begin{bmatrix} \sigma^{ij} & 0 \\ 0 & \sigma^{il}\sigma^{jk} + \sigma^{ik}\sigma^{jl} \end{bmatrix} \end{array} \begin{array}{c} \mu \\ \Sigma = [\sigma_{kl}] \end{array} \tag{91}$$

Notice that the lower right block matrix is a 4D tensor of dimension $d \times d \times d \times d$. The zero subblock matrices in the FIM indicate that the parameters μ and Σ are orthogonal to each other. In particular, when $d = 1$, since $\sigma^{11} = \frac{1}{\sigma^2}$, we recover the Fisher information matrix of the univariate Gaussian:

$$I(\mu, \Sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \tag{92}$$

We refer to [63] for the FIM of a Gaussian distribution using other canonical parameterizations (natural/expectation parameters of exponential family).

Example 3 (FIM of a mixture family $\mathcal{M}$)**.** *A mixture family is defined for $D+1$ functions $F_1, \ldots, F_D$ and C as:*

$$\mathcal{M} = \left\{ p_\theta(x) = \sum_{i=1}^{D} \theta_i F_i(x) + C(x) \text{ such that } \theta \in \Theta \right\}, \tag{93}$$

where the functions $\{F_i(x)\}_i$ are linearly independent on the common support $\mathcal{X}$ and satisfying $\int F_i(x)\mathrm{d}\mu(x) = 0$. Function C is such that $\int C(x)\mathrm{d}\mu(x) = 1$. Mixture families include statistical mixtures with prescribed component distributions and the probability simplex Δ. The FIM of a mixture family is given by:

$$_{\mathcal{M}}I(\theta) = E_{X\sim p_\theta(x)}\left[\frac{F_i(x)F_j(x)}{(p_\theta(x))^2} \right] = \int_{\mathcal{X}} \frac{F_i(x)F_j(x)}{p_\theta(x)} \mathrm{d}\mu(x) \succ 0. \tag{94}$$

The family of Gaussian mixture model (GMM) with prescribed component distributions (ie., convex weight combinations of $D+1$ Gaussian densities) form a mixture family [93].

Notice that the probability simplex of discrete distributions can be *both* modeled as an exponential family or a mixture family [8].

The *expected α-geometry* is built from the *expected dual $\pm\alpha$-connections*. The Fisher "*information metric*" tensor is built from the FIM as follows:

$$ {}_{\mathcal{P}}g(u,v) := (u)_\theta^\top \, {}_{\mathcal{P}}I(\theta) \, (v)_\theta \tag{95} $$

The expected *exponential connection* and expected *mixture connection* are given by

$$ {}^{e}_{\mathcal{P}}\nabla := E_\theta \left[(\partial_i \partial_j l)(\partial_k l) \right], \tag{96} $$

$$ {}^{m}_{\mathcal{P}}\nabla := E_\theta \left[(\partial_i \partial_j l + \partial_i l \partial_j l)(\partial_k l) \right]. \tag{97} $$

The dualistic structure is denoted by $(\mathcal{P}, {}_{\mathcal{P}}g, {}^{m}_{\mathcal{P}}\nabla, {}^{e}_{\mathcal{P}}\nabla)$ with Amari-Chentsov cubic tensor called the *skewness tensor*:

$$ C_{ijk} := E_\theta \left[\partial_i l \partial_j l \partial_k l \right]. \tag{98} $$

It follows that we can build a one-family of expected information α-manifolds:

$$ \left\{ (\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{+\alpha}) \right\}_{\alpha \in \mathbb{R}}, \tag{99} $$

with

$$ {}_{\mathcal{P}}\Gamma^{\alpha}{}_{ij,k}(\theta) := E_\theta \left[\partial_i \partial_j l \partial_k l \right] + \frac{1-\alpha}{2} C_{ijk}(\theta), \tag{100} $$

$$ = E_\theta \left[\left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right]. \tag{101} $$

The Levi-Civita metric connection is recovered as follows:

$$ {}_{\mathcal{P}}\bar{\nabla} = \frac{{}_{\mathcal{P}}\nabla^{-\alpha} + {}_{\mathcal{P}}\nabla^{\alpha}}{2} = {}^{\mathrm{LC}}_{\mathcal{P}}\nabla := {}^{\mathrm{LC}}\nabla({}_{\mathcal{P}}g) \tag{102} $$

The α-Riemann-Christoffel curvature tensor is:

$$ {}_{\mathcal{P}}R_{ijkl} = \partial_i \Gamma^{\alpha}_{jk,l} - \partial_j \Gamma^{\alpha}_{ik,l} + g^{rs} \left(\Gamma^{\alpha}_{ik,r} \Gamma^{\alpha}_{js,l} - \Gamma^{\alpha}_{jk,r} \Gamma^{\alpha}_{is,l} \right), \tag{103} $$

with $R^{\alpha}_{ijkl} = -R^{-\alpha}_{ijlk}$. We check that the expected $\pm\alpha$-connections are coupled with the metric: $\partial_i g_{jk} = \Gamma^{\alpha}_{ij,k} + \Gamma^{-\alpha}_{ik,j}$.

In case of an exponential family $\mathcal{E}$ or a mixture family $\mathcal{M}$ equipped with the dual exponential/mixture connection, we get *dually flat manifolds* (Bregman geometry).

Indeed, for the exponential/mixture families, it is easy to check that the Christoffel symbols of ∇^e and ∇^m vanish:

$$ {}^{e}_{\mathcal{M}}\Gamma = {}^{m}_{\mathcal{M}}\Gamma = {}^{e}_{\mathcal{E}}\Gamma = {}^{m}_{\mathcal{E}}\Gamma = 0. \tag{104} $$

3.10 Criteria for statistical invariance

So far we have explained how to build an information manifold (or information α-manifold) from a pair of conjugate connections. Then we reported two ways to obtain such a pair of conjugate connections: (1) from a parametric divergence, or (2) by using the predefined expected exponential/mixture connections. We now ask the following question: Which information manifold makes sense in Statistics? We can refine the question as follows:

- Which metric tensors g make sense in statistics?

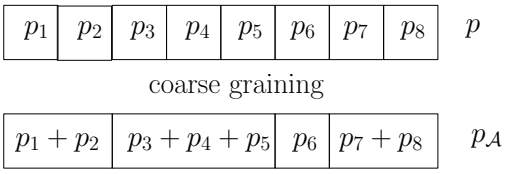


Figure 9: A divergence satisfies the property of information monotonicity iff $D(\theta_{\bar{\mathcal{A}}} : \theta'_{\bar{\mathcal{A}}}) \leq D(\theta : \theta')$. Here, parameter θ represents a discrete distribution.

- Which affine connections ∇ make sense in statistics?
- Which statistical divergences make sense in statistics (from which we can get the metric tensor and dual connections)?

By definition, an *invariant metric tensor* g shall preserve the inner product under important *statistical mappings* called Markov embeddings. Informally, we embed Δ_D into $\Delta_{D'}$ with $D' > D$ and the induced metric should be preserved (see [8], page 62).

Theorem 8 (Uniqueness of Fisher information metric [26, 129])**.** *The Fisher information metric is the unique invariant metric tensor under Markov embeddings up to a scaling constant.*

A D-dimensional parameter (discrete) divergence satisfies the *information monotonicity*[12] if and only if:

$$D(\theta_{\bar{\mathcal{A}}} : \theta'_{\bar{\mathcal{A}}}) \leq D(\theta : \theta') \tag{105}$$

for any *coarse-grained partition* $\mathcal{A} = \{\mathcal{A}_i\}_{i=1}^E$ of $[D] = \{1, \ldots, D\}$ ($\mathcal{A}$-lumping [34]) with $E \leq D$, where $\theta^i_{\bar{\mathcal{A}}} = \sum_{j \in \mathcal{A}_i} \theta^j$ for $i \in [E]$. This concept of coarse-graining is illustrated in Figure 9.

A *separable divergence* $D(\theta_1 : \theta_2)$ is a divergence that can be expressed as the sum of elementary *scalar divergences* $d(x : y)$:

$$D(\theta_1 : \theta_2) := \sum_i d(\theta_1^i : \theta_2^j). \tag{106}$$

For example, the squared Euclidean distance $D(\theta_1 : \theta_2) = \sum_i (\theta_1^i - \theta_2^i)^2$ is a separable divergence for the scalar Euclidean divergence $d(x : y) = (x - y)^2$. The Euclidean distance $D_E(\theta_1, \theta_2) = \sqrt{\sum_i (\theta_1^i - \theta_2^i)^2}$ is *not* separable because of the square root operation.

The only invariant and *decomposable* divergences when $D > 1$ are f-divergences [56] defined for a convex functional generator f:

$$I_f(\theta : \theta') := \sum_{i=1}^D \theta_i f\left(\frac{\theta'_i}{\theta_i}\right) \geq f(1), \quad f(1) = 0 \tag{107}$$

The *standard f-divergences* are defined for f-generators satisfying $f'(1) = 0$ (choose $f_\lambda(u) := f(u) + \lambda(u - 1)$ since $I_{f_\lambda} = I_f$), and $f''(u) = 1$ (scale fixed).

Statistical f-divergences are *invariant* [108] under one-to-one/sufficient statistic transformations $y = t(x)$ of sample space: $p(x; \theta) = q(y(x); \theta)$:

$$\begin{aligned} I_f[p(x;\theta) : p(x;\theta')] &= \int_{\mathcal{X}} p(x;\theta) f\left(\frac{p(x;\theta')}{p(x;\theta)}\right) \mathrm{d}\mu(x), \\ &= \int_{\mathcal{Y}} q(y;\theta) f\left(\frac{q(y;\theta')}{q(y;\theta)}\right) \mathrm{d}\mu(y), \\ &= I_f[q(y;\theta) : q(y;\theta')]. \end{aligned}$$

[12] This property could be renamed as the "distance coarse-binning inequality property."

The dual f-divergences for reference duality is

$$I_f{}^*[p(x;\theta) : p(x;\theta')] = I_f[p(x;\theta') : p(x;\theta)] = I_{f^\diamond}[p(x;\theta) : p(x;\theta')] \tag{108}$$

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

$$f^\diamond(u) := uf\left(\frac{1}{u}\right). \tag{109}$$

One can check that $f^\diamond$ is a standard f-generator when f is standard.

Let us report some common examples of f-divergences:

- The family of α-*divergences*:

$$I_\alpha[p : q] := \frac{4}{1-\alpha^2}\left(1 - \int p^{\frac{1-\alpha}{2}}(x) q^{\frac{1+\alpha}{2}}(x) \mathrm{d}\mu(x)\right), \tag{110}$$

  obtained for $f(u) = \frac{4}{1-\alpha^2}(1 - u^{\frac{1+\alpha}{2}})$. The α-divergences include:

  - the *Kullback-Leibler* when $\alpha \to 1$:

$$\mathrm{KL}[p : q] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x), \tag{111}$$

    for $f(u) = -\log u$.
  - the *reverse Kullback-Leibler* $\alpha \to -1$:

$$\mathrm{KL}^*[p : q] = \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}\mu(x) = \mathrm{KL}[q : p], \tag{112}$$

    for $f(u) = u \log u$.
  - the symmetric squared *Hellinger divergence*:

$$H^2[p : q] = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x), \tag{113}$$

    for $f(u) = (\sqrt{u} - 1)^2$ (corresponding to $\alpha = 0$)
  - the Pearson and Neyman chi-squared divergences [90], etc.

- the *Jensen-Shannon divergence*:

$$\mathrm{JS}[p : q] = \frac{1}{2} \int \left(p(x) \log \frac{2p(x)}{p(x) + q(x)} + q(x) \log \frac{2q(x)}{p(x) + q(x)} \right) \mathrm{d}\mu(x), \tag{114}$$

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

- the *Total Variation*

$$\mathrm{TV}[p : q] = \frac{1}{2} \int |p(x) - q(x)| \mathrm{d}\mu(x), \tag{115}$$

  for $f(u) = \frac{1}{2}|u - 1|$. The total variation distance is the only metric f-divergence.

The f-topology is the topology generated by open f-balls, open balls with respect to f-divergences. Atopology$T$ is saidstrongerthan atopologyT' if T contains all the open sets of T'. Csiszar's theorem [33] states that when $|\alpha| < 1$, the α-topology is equivalent to the topology induced by the total variation metric distance. Otherwise, the α-topology is stronger than the TV topology.

Let us state an important feature of f divergences:

Theorem 9. *The f-divergences are invariant by diffeomorphisms $m(x)$ of the sample space $\mathcal{X}$: Let $Y = m(X)$, and $X_i \sim p_i$ with $Y_i = m(X_i) \sim q_i$. Then we have $I_f[q_1 : q_2] = I_f[p_1 : p_2]$.*

Example 4. *Consider the exponential distributions and the Rayleigh distributions which are related by:*

$$X \sim \text{Exponential}(\lambda) \Leftrightarrow Y = m(X) = \sqrt{X} \sim \text{Rayleigh}\left(\sigma = \frac{1}{\sqrt{2\lambda}}\right).$$

The densities of the exponential distributions are defined by

$$p_\lambda(x) = \lambda \exp(-\lambda x) \text{ with support } \mathcal{X} = [0, \infty),$$

and the densities of the Rayleigh distributions are defined by

$$q_\sigma(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \text{ with support } \mathcal{X} = [0, \infty).$$

We have

$$D_{\mathrm{KL}}\left[q_{\sigma_1} : q_{\sigma_2}\right] = \log\left(\frac{\lambda_2^2}{\lambda_1^2}\right) + \frac{\sigma_1^2 - \sigma_2^2}{\sigma_2^2}.$$

It follows that

$$\begin{aligned} D_{\mathrm{KL}}\left[p_{\lambda_1} : p_{\lambda_2}\right] &= D_{\mathrm{KL}}\left[q\frac{1}{\sqrt{2\lambda_1}} : q\frac{1}{\sqrt{2\lambda_1}}\right] \\ &= \log\frac{2\lambda_1}{2\lambda_2} + 2\lambda_2\left(\frac{1}{2\lambda_1} - \frac{1}{\lambda_2}\right) \\ &= \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_2}{\lambda_1} - 1. \end{aligned}$$

A remarkable property is that invariant standard f-divergences yield the Fisher information matrix and the α-connections. Indeed, the invariant standard f-divergences is related infinitesimally to the Fisher metric as follows:

$$\begin{aligned} I_f[p(x;\theta) : p(x;\theta + \mathrm{d}\theta)] &= \int p(x;\theta) f\left(\frac{p(x;\theta + \mathrm{d}\theta)}{p(x;\theta)}\right) \mathrm{d}\mu(x) & (116) \\ &\overset{\Sigma}{=} \frac{1}{2}{}_F g_{ij}(\theta) \mathrm{d}\theta^i \mathrm{d}\theta^j & (117) \end{aligned}$$

A *statistical parameter divergence* D on a parametric family of distributions $\mathcal{P}$ yields an equivalent *parameter divergence* ${}_{\mathcal{P}}D$:

$${}_{\mathcal{P}}D(\theta : \theta') := D[p(x;\theta) : p(x;\theta')]. \quad (118)$$

Thus we can build the information manifold induced by this parameter divergence ${}_{\mathcal{P}}D(\cdot : \cdot)$. For ${}_{\mathcal{P}}D(\cdot : \cdot) = I_f[\cdot : \cdot]$, the induced ± 1-divergence connections $\overset{I_f}{\mathcal{P}}\nabla := {}_{\mathcal{P}}{I_f}\nabla$ and $\overset{(I_f)^*}{\mathcal{P}}\nabla := {}_{\mathcal{P}}{I_f^*}\nabla$ are precisely the *expected* $\pm\alpha$*-connections* (derived from the exponential/mixture connections) with:

$$\alpha = 2f'''(1) + 3. \quad (119)$$

Thus the invariant connections which coincide with the connections induced by the invariant statistical divergences are the expected α-connections. Note that the curvature of an expected α-connection depends both on α and on the considered statistical model [64].

3.11 Fisher-Rao expected Riemannian manifolds: $(\mathcal{P}, {}_{\mathcal{P}}g)$

Historically, a first manifold modeling of a regular parametric family of distributions $\mathcal{P} = \{p_\theta(x)\}_\theta$ was to consider the *Fisher Information Matrix* (FIM) as the Riemannian metric tensor g (see [53, 111]), with:

$$ {}_{\mathcal{P}}I(\theta) := E_{p_\theta}\left[\partial_i l \partial_j l\right], \tag{120}$$

where $\partial_i l =: \frac{\partial}{\partial \theta_i} \log p(x;\theta)$. Under some regularity conditions, we can rewrite the FIM:

$$ {}_{\mathcal{P}}I(\theta) := -E_{p_\theta}\left[\partial_i \partial_j l\right]. \tag{121}$$

The Riemannian geodesic metric distance D_ρ is commonly called the *Fisher-Rao distance*:

$$D_\rho(p_{\theta_1}, p_{\theta_2}) = \int_0^1 \sqrt{\dot\gamma(t)^\top g_{\gamma(t)} \dot\gamma(t)} \mathrm{d}t, \tag{122}$$

where γ denotes the geodesic passing through $\gamma(0) = \theta_1$ and $\gamma(1) = \theta_2$. The Fisher-Rao distance can also be defined as the shortest path length: $D_\rho(p_{\theta_1}, p_{\theta_2}) = \inf_\gamma \int_0^1 \sqrt{\dot\gamma(t)^\top g_{\gamma(t)} \dot\gamma(t)} \mathrm{d}t$.

Definition 6 (Fisher-Rao distance)**.** *The Fisher-Rao distance is the geodesic metric distance of the Fisher-Riemannian manifold* $(\mathcal{P}, {}_{\mathcal{P}}g)$.

Let us give some examples of Fisher-Riemannian manifolds:

- The Fisher-Riemannian manifold of the family of categorical distributions (also called finite discrete distributions in [8]) amount to the spherical geometry [58] (spherical manifold).
- The Fisher-Riemannian manifold of the family of bivariate location-scale families amount to hyperbolic geometry (hyperbolic manifold).
- The Fisher-Riemannian manifold of the family of location families amount to Euclidean geometry (Euclidean manifold).

The first fundamental form of the Riemannian geometry is $\mathrm{d}s^2 = \langle \mathrm{d}x, \mathrm{d}x\rangle \overset{\Sigma}{=} g_{ij} \mathrm{d}x^i \mathrm{d}x^j$ where $\mathrm{d}s$ denotes the line element.

This Riemannian geometric structure applied to a family of parametric probability distributions was first proposed by Harold Hotelling [53] (in a handwritten note of 1929, reprinted typeset in [123]) and independently later by C. R. Rao [111] (1945, reprinted in [110]). In a similar vein, Jeffreys [55] proposed to use the volume element of a manifold as an invariant prior: The eponym Jeffreys prior in 1946.

Notice that for a parametric family of probability distributions $\mathcal{P}$, the Riemannian structure $(\mathcal{P}, {}_{\mathcal{P}}g)$ coincides with the self-dual conjugate connection manifold $(\mathcal{P}, {}_{\mathcal{P}}g, {}^{I_f}_{\mathcal{P}}\nabla, {}^{I_f}_{\mathcal{P}}\nabla^*)$ induced by a *symmetric* f-divergence like the squared Hellinger divergence.

The exponential map $\exp_p$ at point $p \in M$ provides a way to map back a vector $v \in T_p$ to a point $\exp_p(v) \in M$ (when well-defined). The exponential map can be used to parameterize a geodesic γ with $\gamma(0) = p$ and unit tangent vector $\dot\gamma(0) = v$: $t \mapsto \exp_p(tv)$. For geodesically complete manifolds, the exponential map is defined everywhere.

3.12 The monotone α-embeddings and the metric gauge freedom

Another common mathematically equivalent expression of the FIM [25] is given by:

$$I_{ij}(\theta) := 4 \int \partial_i \sqrt{p(x;\theta)} \partial_j \sqrt{p(x;\theta)} \mathrm{d}\mu(x). \tag{123}$$

This form of the FIM is well-suited to prove that the FIM is always a positive semi-definite matrix [25] ($I(\theta) \succeq 0$). It turns out that we can define a family of *equivalent representations* of the FIM using the *α-embedding* [139] of the parametric family.

First, we define the α-*representation* of densities $l^\alpha(x;\theta) := k_\alpha(p(x;\theta))$ with:

$$k_\alpha(u) := \begin{cases} \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \log u, & \text{if } \alpha = 1. \end{cases} \tag{124}$$

The function $l^\alpha(x;\theta)$ is called the α-*likelihood* function. Then the α-representation of the FIM, the α-FIM for short, is expressed as:

$$I_{ij}^\alpha(\theta) := \int \partial_i l^\alpha(x;\theta) \partial_j l^{-\alpha}(x;\theta) \mathrm{d}\mu(x). \tag{125}$$

We can rewrite compactly the α-FIM, as $I_{ij}^\alpha(\theta) = \int \partial_i l^\alpha \partial_j l^{-\alpha} \mathrm{d}\mu(x)$. Expanding the α-FIM, we get:

$$I_{ij}^\alpha(\theta) = \begin{cases} \frac{1}{1-\alpha^2} \int \partial_i p(x;\theta)^{\frac{1-\alpha}{2}} \partial_j p(x;\theta)^{\frac{1+\alpha}{2}} \mathrm{d}\mu(x) & \text{for } \alpha \neq \pm 1 \\ \int \partial_i \log p(x;\theta) \partial_j p(x;\theta) \mathrm{d}\mu(x) & \text{for } \alpha \in \{-1, 1\} \end{cases} \tag{126}$$

The 1-representation of the density is called the *logarithmic representation* (or e-representation), the -1-representation the *mixture representation* (or m-representation), and its 0-representation is called the *square root representation*. The set of α-*scores* vectors $B_\alpha := \{\partial_i l^\alpha\}_i$ are interpreted as the tangent basis vectors of the α-base B_α. Thus the FIM is α-independent.

Furthermore, the α-representation of the FIM can be rewritten under mild conditions [25] as:

$$I_{ij}^\alpha(\theta) = -\frac{2}{1+\alpha} \int p(x;\theta)^{\frac{1+\alpha}{2}} \partial_i \partial_j l^\alpha(x;\theta) \mathrm{d}\mu(x). \tag{127}$$

Since we have:

$$\partial_i \partial_j l^\alpha(x;\theta) = p^{\frac{1-\alpha}{2}} \left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right), \tag{128}$$

it follows that:

$$I_{ij}^\alpha(\theta) = -\frac{2}{1+\alpha} \left(-I_{ij}(\theta) + \frac{1-\alpha}{2} I_{ij} \right) = I_{ij}(\theta). \tag{129}$$

Notice that when $\alpha = 1$, we recover the equivalent expression of the FIM (under mild conditions):

$$I_{ij}^1(\theta) = -E[\nabla^2 \log p(x;\theta)]. \tag{130}$$

In particular, when the family is an exponential family [84] with cumulant function $F(\theta)$ (satisfying the mild conditions), we have:

$$I(\theta) = \nabla^2 F(\theta). \tag{131}$$

Zhang [139, 69] further discussed the representation/reference biduality which was confounded in the α-geometry.

Gauge freedom of the Riemannian metric tensor has been investigated under the framework of (ρ, τ)-monotone embeddings [139, 98, 69] in information geometry: Let ρ and τ be two strictly increasing functions, and f a strictly convex function such that $f'(\rho(u)) = \tau(u)$ (with f^* denoting its convex conjugate). Observe that the set of strictly increasing real-valued univariate functions has a group structure for the group operation chosen as the functional composition $\circ$. Let us write $p_\theta(x) = p(x;\theta)$.

The (ρ, τ)-metric tensor ${}^{\rho,\tau}g(\theta) = [\, {}^{\rho,\tau}g_{ij}(\theta)]_{ij}$ can be derived from the (ρ, τ)-divergence:

$$D_{\rho,\tau}(p:q) = \int \left(f(\rho(p(x))) + f^*(\tau(q(x))) - \rho(p(x))\tau(q(x)) \right) \mathrm{d}\nu(x) \tag{132}$$

We have:

$$
\begin{aligned}
{}^{\rho,\tau}g_{ij}(\theta) &= \int (\partial_i \rho(p_\theta(x)))\,(\partial_j \tau(p_\theta(x)))\,\mathrm{d}\nu(x), && (133)\\
&= \int \rho'(p_\theta(x))\tau'(p_\theta(x))\,(\partial_i p_\theta(x))\,(\partial_j p_\theta(x))\,\mathrm{d}\nu(x), && (134)\\
&= \int f''(\rho(p_\theta(x)))\,(\partial_i \rho(p_\theta(x)))\,(\partial_j \rho(p_\theta(x)))\,\mathrm{d}\nu(x), && (135)\\
&= \int (f^*)''(\tau(p_\theta(x)))\,(\partial_i \tau(p_\theta(x)))\,(\partial_j \tau(p_\theta(x)))\,\mathrm{d}\nu(x). && (136)
\end{aligned}
$$

3.13 Dually flat spaces and canonical Bregman divergences

We have described how to build a dually flat space from any strictly convex and smooth generator F: A Hessian structure is built from $F(\theta)$ with Riemannian Hessian metric $\nabla^2 F(\theta)$, and the convex conjugate $F^*(\eta)$ (obtained by the Legendre-Fenchel duality) yields the dual Hessian structure with Riemannian Hessian metric $\nabla^2 F^*(\eta)$. The dual connections ∇ and ∇^* are coupled with the metric. The connections are defined by their respective Christoffel symbols $\Gamma(\theta) = 0$ and $\Gamma^*(\eta) = 0$, showing that they are flat connections.

Conversely, it can be proved [8] that given two dually flat connections ∇ and ∇^*, we can reconstruct two dual canonical strictly convex potential functions $F(\theta)$ and $F^*(\eta)$ such that $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$. The canonical divergence A_{F,F^*} yields the dual Bregman divergences B_F and B_{F^*}.

The only symmetric Bregman divergences are squared Mahalanobis distances M_Q^2 [20] with the Mahalanobis distance defined by:

$$M_Q(\theta, \theta') = \sqrt{(\theta' - \theta)^\top Q (\theta' - \theta)}. \quad (137)$$

Let $Q = LL^\top$ be the Cholesky decomposition of a positive-definite matrix $Q \succ 0$. It is well-known that the Mahalanobis distance M_Q amounts to the Euclidean distance on affinely transformed points:

$$
\begin{aligned}
M_Q^2(\theta, \theta') &= \Delta\theta^\top Q \Delta\theta, && (138)\\
&= \Delta\theta^\top L L^\top \Delta\theta, && (139)\\
&= M_I^2(L^\top\theta, L^\top\theta') = \|L^\top\theta - L^\top\theta'\|^2, && (140)
\end{aligned}
$$

where $\Delta\theta = \theta' - \theta$.

The squared Mahalanobis distance M_Q^2 does not satisfy the triangle inequality, but the Mahalanobis distance M_Q is a metric distance. We can convert a Mahalanobis distance M_{Q_1} into another Mahalanobis distance M_{Q_2}, and vice versa, as follows:

Proof. Let us write matrix $Q = L^\top L \succ 0$ using the Cholesky decomposition. Then we have

$$M_Q(\theta_1, \theta_2) = M_I(L^\top\theta_1, L^\top\theta_2) \Leftrightarrow M_I(\theta_1, \theta_2) = M_Q((L^\top)^{-1}\theta_1, ((L^\top)^{-1}\theta_2). \quad (141)$$

Then we have for two symmetric positive-definite matrices $Q_1 = L_1^\top L_1 \succ 0$ and $Q_2 = L_2^\top L_2 \succ 0$:

$$M_{Q_1}(\theta_1, \theta_2) = M_I(L_1^\top\theta_1, L_1^\top\theta_2) = M_{Q_2}((L_2^\top)^{-1}L_1^\top\theta_1, (L_2^\top)^{-1}L_1^\top\theta_2). \quad (142)$$

It follows that we have:

$$M_{Q_1}(\theta_1, \theta_2) = M_{Q_2}((L_2^\top)^{-1}L_1^\top\theta_1, (L_2^\top)^{-1}L_1^\top\theta_2). \quad (143)$$

□

We have $M_Q^2(\theta_1, \theta_2) = B_F(\theta_1, \theta_2)$ (Bregman divergence) with $F(\theta) = \frac{1}{2}\theta^\top Q\theta$ for a positive-definite matrix $Q \succ 0$. The convex conjugate $F^*(\eta) = \frac{1}{2}\eta^\top Q^{-1}\eta$ (with $Q^{-1} \succ 0$). We have $\eta = Q^{-1}\theta$ and $\eta = Q\theta$. We have the following identity between the *dual Mahalanobis divergences* M_Q^2 and $M_{Q^{-1}}^2$:

$$M_Q^2(\theta_1, \theta_2) = M_{Q^{-1}}^2(\eta_1, \eta_2). \quad (144)$$

When the Bregman generator is based on an integral, i.e., the log-normalizer $F(\theta) = \log\left(\int \exp(\langle t(x), \theta\rangle \mathrm{d}\mu(x)\right)$ for exponential families $\mathcal{E}$, or the negative Shannon entropy $F(\theta) = \int m_\theta(x) \log m(\eta) \mathrm{d}\mu(x)$ for mixture families $\mathcal{M}$, the associated Bregman divergences $B_{F,\mathcal{E}}$ or $B_{F,\mathcal{M}}$ can be relaxed and interpreted as a statistical distance. We explain how to obtain the reconstruction below:

- Consider an exponential family $\mathcal{E}$ of order D with densities defined according to a dominating measure μ:

$$\mathcal{E} = \{p_\theta(x) = \exp(\theta^\top t(x) - F(\theta)) \ : \ \theta \in \Theta\}, \tag{145}$$

where the natural parameter θ and the sufficient statistic vector $t(x)$ belong to $\mathbb{R}^D$. We have the integral-based Bregman generator:

$$F(\theta) = F_{\mathcal{E}}(p_\theta) = \log\left(\int \exp(\theta^\top t(x)) \mathrm{d}\mu(x)\right), \tag{146}$$

and the dual convex conjugate

$$F^*(\eta) = -h(p_\theta) = \int p(x) \log p(x) \mathrm{d}\mu(x), \tag{147}$$

where $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$ denotes Shannon's entropy.

Let $\lambda(i)$ denotes the i-th coordinates of vector λ, and let us calculate the *inner product* $\theta_1^\top \eta_2 = \sum_i \theta_1(i)\eta_2(i)$ of the Legendre-Fenchel divergence. We have $\eta_2(i) = E_{p_{\theta_2}}[t_i(x)]$. Using the linear property of the expectation $E[\cdot]$, we find that $\sum_i \theta_1(i)\eta_2(i) = E_{p_{\theta_2}}\left[\sum_i \theta_1(i) t_i(x)\right]$. Moreover, we have $\sum_i \theta_1(i) t_i(x) = (\log p_{\theta_1}(x)) + F(\theta_1)$. Thus we have:

$$\theta_1^\top \eta_2 = E_{p_{\theta_2}}\left[\log p_{\theta_1} + F(\theta_1)\right] = F(\theta_1) + E_{p_{\theta_2}}\left[\log p_{\theta_1}\right]. \tag{148}$$

It follows that we get

$$\begin{aligned} B_{F,\mathcal{E}}[p_{\theta_1} : p_{\theta_2}] &= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2, & (149)\\ &= F(\theta_1) - h(p_{\theta_2}) - E_{p_{\theta_2}}[\log p_{\theta_1}] - F(\theta_1), & (150)\\ &= E_{p_{\theta_2}}\left[\log \frac{p_{\theta_2}}{p_{\theta_1}}\right] =: D_{\mathrm{KL}^*}[p_{\theta_1} : p_{\theta_2}], & (151) \end{aligned}$$

By relaxing the exponential family densities p_{θ_1} and p_{θ_2} to be arbitrary densities p_1 and p_2, we obtain the *reverse KL divergence* between p_1 and p_2 from the dually flat structure induced by the integral-based log-normalizer of an exponential family:

$$\begin{aligned} D_{\mathrm{KL}^*}[p_1 : p_2] &= E_{p_2}\left[\log \frac{p_2}{p_1}\right] = \int p_2(x) \log \frac{p_2(x)}{p_1(x)} \mathrm{d}\mu(x), & (152)\\ &= D_{\mathrm{KL}}[p_2 : p_1]. & (153) \end{aligned}$$

Thus we have recovered the reverse Kullback-Leibler divergence D_{KL^*} from $B_{F,\mathcal{E}}$.

The dual divergence $D^*[p_1 : p_2] := D[p_2 : p_1]$ is obtained by swapping the distribution parameter orders. We have:

$$D^*_{\mathrm{KL}^*}[p_1 : p_2] := D_{\mathrm{KL}^*}[p_2 : p_1] = E_{p_1}\left[\log \frac{p_1}{p_2}\right] =: D_{\mathrm{KL}}[p_1 : p_2], \tag{154}$$

and $D_{\mathrm{KL}^*}[p_1 : p_2] = D^*_{\mathrm{KL}^*}[p_2 : p_1] = D_{\mathrm{KL}}[p_2 : p_1]$.

To summarize, the canonical Legendre-Fenchel divergence associated with the log-normalizer of an exponential family amounts to the statistical reverse Kullback-Leibler divergence between p_{θ_1} and p_{θ_1}

(or the KL divergence between the swapped corresponding densities): $D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1) = A_{F,F^*}(\theta_2 : \eta_1)$. Notice that it is easy to check that $D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1)$ [14, 16]. Here, we took the opposite direction by constructing D_{KL} from B_F.

We may consider an auxiliary carrier term $k(x)$ so that the densities write $p_\theta(x) = \exp(\theta^\top t(x) - F(\theta) + k(x))$. Then the dual convex conjugate writes [88] as $F^*(\eta) = -h(p_\theta) + E_{p_\theta}[k(x)]$.

Notice that since the Bregman generator is defined up to an affine term, we may consider the equivalent generator $F(\theta) = -\log p_\theta(\omega)$ instead of the integral-based generator. This approach yields ways to build formula bypassing the explicit use of the log-normalizer for calculating various statistical distances [94].

- In this second example, we consider a mixture family

$$\mathcal{M} = \left\{ m_\theta = \sum_{i=1}^{D} \theta_i p_i(x) + (1 - \sum_{i=1}^{D} \theta_i) p_0(x) \right\}, \tag{155}$$

where $p_0, \ldots, p_D$ are $D+1$ linearly independent probability densities. The integral-based Bregman generator F is chosen as Shannon negentropy:

$$F(\theta) = F_{\mathcal{M}}(m_\theta) = -h(m_\theta) = \int m_\theta(x) \log m_\theta(x) \mathrm{d}\mu(x). \tag{156}$$

We have

$$\eta_i = [\nabla F(\theta)]_i = \int (p_i(x) - p_0(x)) \log m_\theta(x) \mathrm{d}\mu(x), \tag{157}$$

and the dual convex potential function is

$$F^*(\eta) = -\int p_0(x) \log m_\theta(x) \mathrm{d}\mu(x) = h^\times(p_0 : m_\theta), \tag{158}$$

i.e., the cross-entropy between the density p_0 and the mixture m_θ. Let us calculate the inner product $\theta_1^\top \eta_2$ of the Legendre-Fenchel divergence as follows:

$$\begin{aligned} \sum_i \theta_1(i) \int (p_i(x) - p_0(x)) \log m_{\theta_2}(x) \mathrm{d}\mu(x) &= \int \sum_i \theta_1(i) p_i(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x) \\ &\quad - \sum_i \theta_1(i) p_0(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x). \end{aligned} \tag{159}$$

That is

$$\theta_1^\top \eta_2 = \int \sum_i \theta_1(i) p_i \log m_{\theta_2} \mathrm{d}\mu - \sum_i \theta_1(i) p_0 \log m_{\theta_2} \mathrm{d}\mu. \tag{160}$$

Thus it follows that we have the following statistical distance:

$$\begin{aligned}
B_{F,\mathcal{M}}[m_{\theta_1} : m_{\theta_2}] &:= F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2, && (161)\\
&= -h(m_{\theta_1}) - \int p_0(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x) - \int \sum_i \theta_1(i) p_i(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x) \\
&\quad + \sum_i \theta_1(i) p_0(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (162)\\
&= -h(m_{\theta_1}) - \int ((1 - \sum_i \theta_1(i)) p_0(x) + \sum_i \theta_1(i) p_i(x)) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (163)\\
&= -h(m_{\theta_1}) - \int m_{\theta_1}(x) \log m_{\theta_2}(x) \mathrm{d}\mu(x), && (164)\\
&= \int m_{\theta_1}(x) \log \frac{m_{\theta_1}(x)}{m_{\theta_2}(x)} \mathrm{d}\mu(x), && (165)\\
&= D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}]. && (166)
\end{aligned}$$

Thus we have $D_{\mathrm{KL}}[m_{\theta_1} : m_{\theta_2}] = B_F(\theta_1 : \theta_2)$. By relaxing the mixture densities m_{θ_1} and m_{θ_2} to arbitrary densities m_1 and m_2, we find that the dually flat geometry induced by the negentropy of densities of a mixture family induces a statistical distance which corresponds to the (forward) KL divergence. That is, we have recovered the statistical distance D_{KL} from $B_{F,\mathcal{M}}$. Note that in general the entropy of a mixture is not available in closed-form (because of the log sum term), except when the component distributions have pairwise disjoint supports. This latter case includes the case of Dirac distributions whose mixtures represent the categorical distributions.

Dually flat spaces can be built from any strictly convex C^3 generator F. Vinberg and Koszul [120] showed how to obtain such a convex generator for homogeneous cones. A cone $\mathcal{C}$ in a vector space V yields a dual cone of positive linear functionals in the dual vector space V^*:

$$\mathcal{C}^* := \{\omega \in V^* : \forall v \in \mathcal{C}, \omega(v) \geq 0\}. \quad (167)$$

The characteristic function of the cone is defined by

$$\chi_{\mathcal{C}}(\theta) := \int_{\mathcal{C}^*} \exp(-\omega(\theta)) \mathrm{d}\omega \geq 0, \quad (168)$$

and the function $\log \chi_{\mathcal{C}}(\theta)$ defines a Bregman generator which induces a Hessian structure and a dually flat space.

Figure 10 displays the main types of information manifolds encountered in information geometry with their relationships.

4 Some applications of information geometry

Information geometry [8] found broad applications in information sciences. For example, we can mention:

- Statistics: Asymptotic inference, Expectation-Maximization (EM and the novel information-geometric em), time series (AutoRegressive Moving Average model, ARMA) models,
- Machine learning: Restricted Boltzmann machines (RBMs), neuromanifolds and natural gradient [124],
- Signal processing: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF),
- Mathematical programming: Barrier function of interior point methods,
- Game theory: Score functions.

Next, we shall describe a few applications, starting with the celebrated natural gradient descent.

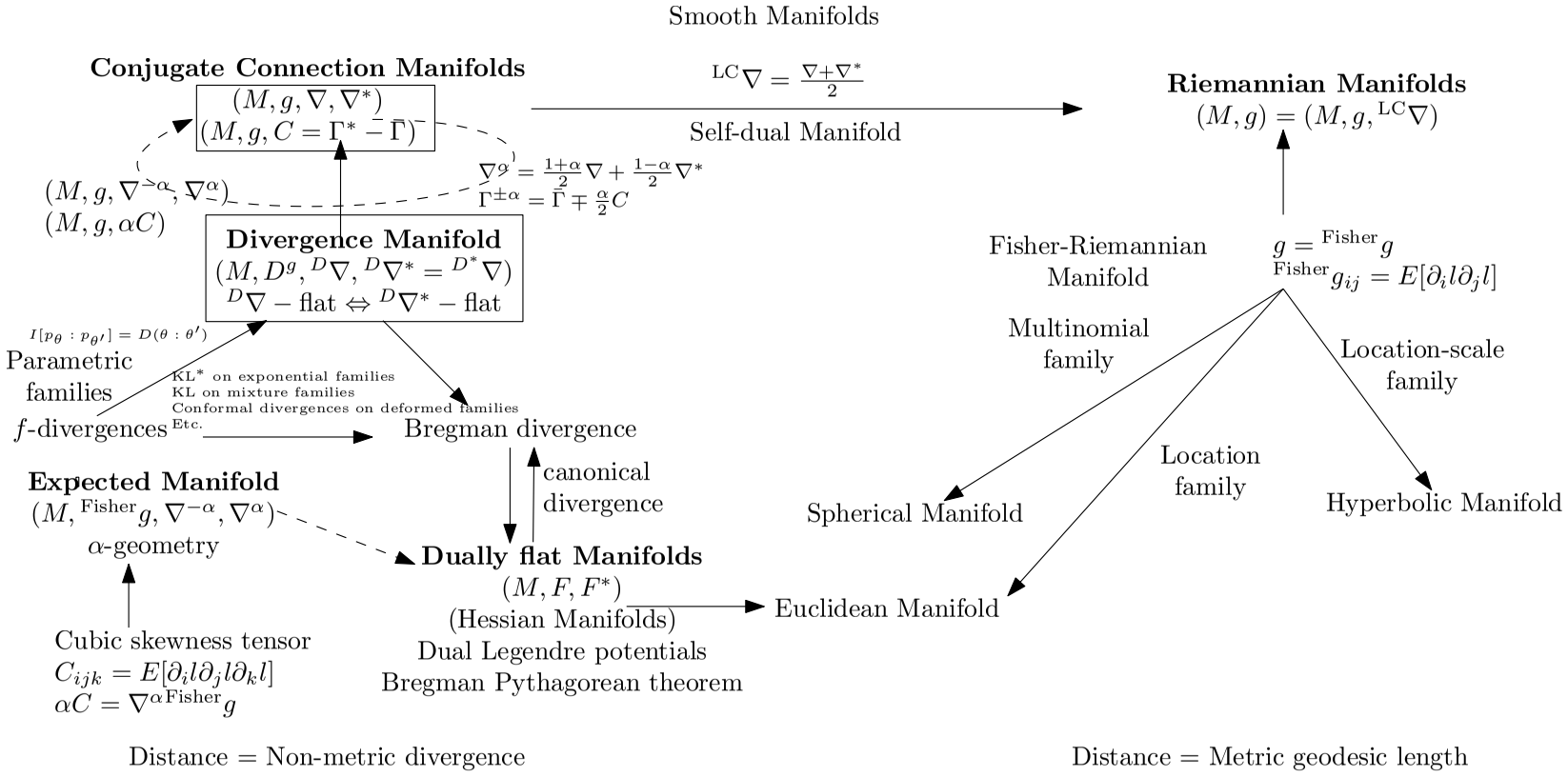


Figure 10: Overview of the main types of information manifolds with their relationships in information geometry.

4.1 Natural gradient in Riemannian space

The *Natural Gradient* [6] (NG) is an extension of the ordinary (Cartesian) gradient of Euclidean geometry to the gradient in a Riemannian space analyzed in an arbitrary coordinate system. We explain the natural gradient

4.1.1 The vanilla gradient descent method

Given a real-valued function $L_\theta(\theta)$ parameterized by a a D-dimensional vector θ on parameter space $\theta \in \Theta \subset \mathbb{R}^D$, we wish to minimize L_θ, i.e., solve $\min_{\theta\in\Theta} L_\theta(\theta)$. The *gradient descent* (GD) method, also called the *steepest descent* method, is a first-order local optimization procedure which starts by initializing the parameter to an arbitrary value (say, $\theta_0 \in \Theta$), and then iteratively updates at stage t the current location of θ_t to θ_{t+1} as follows:

$$\text{GD}: \quad \theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L_\theta(\theta_t). \tag{169}$$

The scalar $\alpha_t > 0$ is called the *step size* or *learning rate* in machine learning. The ordinary gradient (OG) $\nabla_\theta F_\theta(\theta)$ (vector of partial derivatives) represents the *steepest vector* at θ of the function graph $\mathcal{L}_\theta = \{(\theta, L_\theta(\theta)) \ : \ \theta \in \Theta\}$. The GD method was pioneered by Cauchy [28] (1847) and its convergence proof to a *stationary point* was first reported in Curry [35] (1944).

If we *reparameterize* the function L_θ using a one-to-one and onto differentiable mapping $\eta = \eta(\theta)$ (with reciprocal inverse mapping $\theta = \theta(\eta)$), the GD update rule transforms as:

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\eta L_\eta(\eta_t), \tag{170}$$

where

$$L_\eta(\eta) := L_\theta(\theta(\eta)). \tag{171}$$

Thus in general, the two gradient descent location sequences $\{\theta_t\}_t$ and $\{\eta_t\}_t$ (initialized at $\theta_0 = \theta(\eta_0)$ and $\eta_0 = \eta(\theta_0)$) are *different* (because usually $\eta(\theta) \neq \theta$), and the two GDs may potentially reach different stationary points. In other words, the GD local optimization *depends on the choice of the parameterization* of the function L (i.e., L_θ or L_η). For example, minimizing with the gradient descent a temperature function $L_\theta(\theta)$ with respect to Celsius degrees θ may yield a different result than minimizing the same temperature function $L_\eta(\eta) = L_\theta(\theta(\eta))$ expressed with respect to Fahrenheit degrees η. That is, the GD optimization is *extrinsic* since it depends on the choice of the parameterization of the function, and does not take into account the underlying geometry of the parameter space Θ.

The natural gradient precisely addresses this problem and solves it by choosing *intrinsically* the steepest direction with respect to a Riemannian metric tensor field on the parameter manifold. We shall explain the natural gradient descent method and highlight its connections with the Riemannian gradient descent, the mirror descent and even the ordinary gradient descent when the parameter space is dually flat.

4.1.2 Natural gradient and its connection with the Riemannian gradient

Let (M, g) be a D-dimensional Riemannian space [38] equipped with a metric tensor g, and $L \in C^\infty(M)$ a smooth function to minimize on the manifold M. The *Riemannian gradient* [21] uses the Riemannian *exponential map* $\exp_p : T_p \to M$ to update the sequence of points p_t's on the manifold as follows:

$$\text{RG}: \quad p_{t+1} = \exp_{p_t}(-\alpha_t \nabla_M L(p_t)), \tag{172}$$

where the Riemannian gradient ∇_M is defined according to a *directional derivative* ∇_v by:

$$\nabla_M L(p) := \nabla_v \left(L\left(\exp_p(v)\right)\right)\big|_{v=0}, \tag{173}$$

with

$$\nabla_v L(p) := \lim_{h\to 0} \frac{L(p+hv) - L(p)}{h}. \tag{174}$$

However, the Riemannian exponential mapping $\exp_p(\cdot)$ is often computationally intractable since it requires to solve a system of second-order differential equations [38, 1]. Thus instead of using $\exp_p$, we shall rather use a computable *Euclidean retraction* $R : T_p \to \mathbb{R}^D$ of the exponential map expressed in a local θ-coordinate system as:

$$\text{RetG}: \quad \theta_{t+1} = R_{\theta_t}\left(-\alpha_t \nabla_\theta L_\theta(\theta_t)\right). \tag{175}$$

Using the retraction [1] $R_p(v) = p + v$ which corresponds to a first-order Taylor approximation of the exponential map, we recover the *natural gradient descent* [6]:

$$\text{NG} : \theta_{t+1} = \theta_t - \alpha_t g_\theta^{-1}(\theta_t)\nabla_\theta L_\theta(\theta_t). \tag{176}$$

The *natural gradient* [6] (NG)

$$^{\text{NG}}\nabla L_\theta(\theta) := g_\theta^{-1}(\theta)\nabla_\theta L_\theta(\theta) \tag{177}$$

encodes the *Riemannian steepest descent* vector, and the natural gradient descent method yields the following update rule

$$\text{NG} : \theta_{t+1} = \theta_t - \alpha_t \, {}^{\text{NG}}\nabla L_\theta(\theta_t). \tag{178}$$

Notice that the natural gradient is a *contravariant vector* while the ordinary gradient is a *covariant vector*. Recall that a covariant vector $[v_i]$ is transformed into a contravariant vector $[v^i]$ by $v^i = \sum_j g^{ij} v_i$, that is by using the dual Riemannian metric $g^*_\eta(\eta) = g_\theta(\theta)^{-1}$. The natural gradient is *invariant* under an invertible smooth change of parameterization. However, the natural gradient *descent* does *not* guarantee that the locations θ_t's always stay on the manifold: Indeed, it may happen that for some t, $\theta_t \notin \Theta$ when $\Theta \neq \mathbb{R}^D$.

Property 5 ([21])**.** *The natural gradient descent approximates the intrinsic Riemannian gradient descent using a contravariant gradient vector induced by the Riemannian metric tensor g. The natural gradient is invariant to coordinate transformations.*

Next, we shall explain how the natural gradient descent is related to the *mirror descent* and the *ordinary gradient* when the Riemannian space Θ is dually flat.

4.1.3 Natural gradient in dually flat spaces: Connections to Bregman mirror descent and ordinary gradient

Recall that a dually flat space (M, g, ∇, ∇^*) is a manifold M equipped with a pair (∇, ∇^*) of dual torsion-free flat connections which are coupled to the Riemannian metric tensor g [8, 77] in the sense that $\frac{\nabla+\nabla^*}{2} = {}^{LC}\nabla$, where ${}^{LC}\nabla$ denotes the unique metric torsion-free Levi-Civita connection.

On a dually flat space, there exists a pair of dual global *Hessian structures* [120] with dual canonical Bregman divergences [23, 8]. The dual Riemannian metrics can be expressed as the Hessians of dual convex potential functions F and F^*. Examples of Hessian manifolds are the *manifolds of exponential families* or the *manifolds of mixture families* [86]. On a dually flat space induced by a strictly convex and C^3 function F (Bregman generator), we have two dual global coordinate system: $\theta(\eta) = \nabla F^*(\eta)$ and $\eta(\theta) = \nabla F(\theta)$, where F^* denotes the Legendre-Fenchel convex conjugate function [70, 71]. The Hessian metric expressed in the primal θ-coordinate system is $g_\theta(\theta) = \nabla^2 F(\theta)$, and the dual Hessian metric expressed in the dual coordinate system is $g^*_\eta(\eta) = \nabla^2 F^*(\eta)$. Crouzeix's identity [32] shows that $g_\theta(\theta) g_\eta(\eta) = I$, where I denotes the $D \times D$ matrix identity.

The ordinary gradient descent method can be extended using a *proximity function* $\Phi(\cdot, \cdot)$ as follows:

$$\text{PGD}: \quad \theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \langle \theta, \nabla L_\theta(\theta_t) \rangle + \frac{1}{\alpha_t} \Phi(\theta, \theta_t) \right\}. \tag{179}$$

When $\Phi(\theta, \theta_t) = \frac{1}{2}\|\theta - \theta_t\|^2$, the PGD update rule becomes the ordinary GD update rule.

Consider a Bregman divergence [23] B_F for the proximity function Φ: $\Phi(p,q) = B_F(p:q)$. Then the PGD yields the following *mirror descent* (MD):

$$\text{MD}: \quad \theta_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \langle \theta, \nabla L(\theta_t)\rangle + \frac{1}{\alpha_t} B_F(\theta:\theta_t) \right\}. \tag{180}$$

This mirror descent can be interpreted as a natural gradient descent as follows:

Property 6 ([112])**.** *Bregman mirror descent on the Hessian manifold $(M, g = \nabla^2 F(\theta))$ is equivalent to natural gradient descent on the dual Hessian manifold $(M, g^* = \nabla^2 F(\eta))$, where F is a Bregman generator, $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$.*

Indeed, the mirror descent rule yields the following natural gradient update rule:

$$\begin{aligned} \text{NG}^*: \eta_{t+1} &= \eta_t - \alpha_t (g_\eta^*)^{-1}(\eta_t) \nabla_\eta L_\theta(\theta(\eta_t)), & (181)\\ &= \eta_t - \alpha_t (g_\eta^*)^{-1}(\eta_t) \nabla_\eta L_\eta(\eta_t), & (182) \end{aligned}$$

where $g_\eta^*(\eta) = \nabla^2 F^*(\eta) = (\nabla_\theta^2 F(\theta))^{-1}$ and $\theta(\eta) = \nabla F^*(\theta)$.

The method is called mirror descent [24] because it performs that gradient step in the *dual space* (ie., mirror space) $H = \{\eta = \nabla F(\theta) \ : \ \theta \in \Theta\}$, and thus solves the inconsistency contravariant/covariant type problem of subtracting a covariant vector from a contravariant vector of the ordinary GD (Eq. 169).

Let us prove now the following property of the natural gradient in a dually flat space or Bregman manifold [77]:

Property 7 ([137])**.** *In a dually flat space induced by potential convex function F, the natural gradient amounts to the ordinary gradient on the dually parameterized function: ${}^{\text{NG}}\nabla L_\theta(\theta) = \nabla_\eta L_\eta(\eta)$ where $\eta = \nabla_\theta F(\theta)$ and $L_\eta(\eta) = L_\theta(\theta(\eta))$.*

Proof. Let (M, g, ∇, ∇^*) be a dually flat space. We have $g_\theta(\theta) = \nabla^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$ since $\eta = \nabla_\theta F(\theta)$. The function to minimize can be written either as $L_\theta(\theta) = L_\theta(\theta(\eta))$ or as $L_\eta(\eta) = L_\eta(\eta(\theta))$. Recall the chain rule in the calculus of differentiation:

$$\nabla_\theta L_\theta(\theta) = \nabla_\theta (L_\eta(\eta(\theta))) = (\nabla_\theta \eta)(\nabla_\eta L_\eta(\eta)). \tag{183}$$

Thus we have:

$$\begin{aligned} {}^{\text{NG}}\nabla L_\theta(\theta) &:= g_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta), & (184)\\ &= (\nabla_\theta \eta)^{-1} (\nabla_\theta \eta) \nabla_\eta L_\eta(\eta), & (185)\\ &= \nabla_\eta L_\eta(\eta). & (186) \end{aligned}$$

□

It follows that the natural gradient descent on a loss function $L_\theta(\theta)$ amounts to an ordinary gradient descent on the *dually parameterized* loss function $L_\eta(\eta) := L_\theta(\theta(\eta))$. In short, ${}^{\text{NG}}\nabla_\theta L_\theta = \nabla_\eta L_\eta$.

4.1.4 An application of the natural gradient: Natural Evolution Strategies (NESs)

A nice family of applications of the natural gradient are the Natural Evolution Strategies (NESs) for black-box minimization [19]: Let $f(x)$ for $x \in \mathbb{X} \subset \mathbb{R}^d$ be a real-valued function to minimize. Berny [18] proposed to *relax* the optimization problem $\min_{x\in\mathbb{X}} f(x)$ by considering a parametric search distribution p_λ, and minimize instead:

$$\min_{\lambda\in\Lambda} E_{p_\lambda}[f(x)], \tag{187}$$

where $\lambda \in \Lambda \subset \mathbb{R}^D$ denotes the parameter space of the search distributions. Let $J(\lambda) = E_{p_\lambda}[f(x)]$. Minimizing $J(\lambda)$ instead of $f(x)$ is particularly useful when $\mathbb{X}$ is a discrete space: Indeed, the *combinatorial*

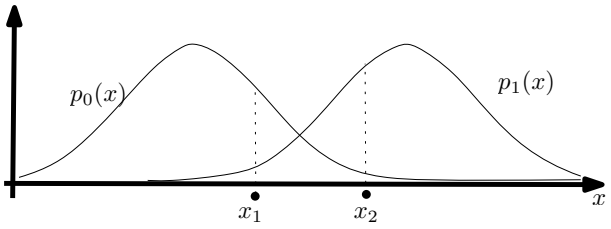


Figure 11: Statistical Bayesian hypothesis testing: The best Maximum A Posteriori (MAP) rule chooses to classify an observation from the class that yields the maximum likelihood.

optimization [18] $\min_{x\in\mathbb{X}} f(x)$ is replaced by a continuous optimization $\min_{\lambda\in\Lambda} J(\lambda)$ when Λ is a continuous parameter, and the ordinary or natural GD methods can be used. The gradient $\nabla J(\lambda)$ is called the *search gradient*, and it can be approximated stochastically using the log-likelihood trick [135] as

$$\widetilde{\nabla}\ J(\lambda) := \frac{1}{n}\sum_{i=1}^{n} f(x_i)\nabla \log p_\lambda(x_i) \approx \nabla J(\lambda), \tag{188}$$

where $x_1,\ldots,x_n \sim p_\lambda$. Similarly, the Fisher information matrix (FIM) may be approximated by the following empirical FIM:

$$\tilde{I}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\lambda l_\lambda(x_i)(\nabla_\lambda l_\lambda(x_i))^\top \approx I(\lambda), \tag{189}$$

where $l_\lambda(x) := \log p_\lambda(x)$ denote the log-likelihood function. Notice that the approximated FIM may potentially be degenerated and may not respect the structure of the true FIM. For example, we have $\nabla_\mu l(x;\mu,\sigma^2) = \frac{x-\mu}{\sigma^2}$ and $\nabla_{\sigma^2} = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$. The non-diagonal of the approximate FIM $\tilde{I}(\lambda)$ are close to but usually non-zero although the expected FIM is diagonal $I(\mu,\sigma^2) = \mathrm{diag}\left(\frac{1}{\sigma^2}, \frac{1}{2\sigma^4}\right)$. Thus we may estimate the FIM until the non-diagonal elements have absolute values less than a prescribed $\epsilon > 0$. For multivariate normals, we have $\nabla_\mu l(x;\mu,\Sigma) = \Sigma^{-1}(x-\mu)$ and $\nabla_\Sigma l(x;\mu,\Sigma) = \frac{1}{2}(\nabla_\mu l(x;\mu,\Sigma)\nabla_\mu l(x;\mu,\Sigma)^\top - \Sigma^{-1})$.

4.2 Some illustrating applications of dually flat manifolds

In this part, we describe how to use the dually flat structures for handling an exponential family $\mathcal{E}$ (in a hypothesis testing problem detailed in §4.3) and the mixture family $\mathcal{M}$ (clustering statistical mixtures §4.4). Note that for a general divergence, neither $(\mathcal{E}, D)$ nor $(\mathcal{M}, D)$ is dually flat. However, when $D = \mathrm{KL}$, the Kullback-Leibler divergence, we get dually flat spaces that are computationally attractive since the primal/dual geodesics are straight lines in the corresponding global affine coordinate system.

4.3 Hypothesis testing in the dually flat exponential family manifold $(\mathcal{E}, \mathrm{KL}^*)$

Given two probability distributions $P_0 \sim p_0(x)$ and $P_1 \sim p_1(x)$, we ask to classify a set of iid. observations $X_{1:n} = \{x_1,\ldots,x_n\}$ as either sampled from P_0 or from P_1? This is a statistical decision problem [73]. For example, P_0 can represent the signal distribution and P_1 the noise distribution. Figure 11 displays the probability distributions and the unavoidable error that is made by any statistical decision rule (on observations x_1 and x_2).

Assume that both distributions $P_0 \sim P_{\theta_0}$ and $P_1 \sim P_{\theta_1}$ belong to the same *exponential family* $\mathcal{E} = \{P_\theta : \theta \in \Theta\}$, and consider the exponential family manifold with the dually flat structure $(\mathcal{E}, {}_{\mathcal{E}}g, {}_{\mathcal{E}}\nabla^e, {}_{\mathcal{E}}\nabla^m)$. That is, the manifold equipped with the Fisher information metric tensor field and the expected exponential connection and conjugate expected mixture connection. More generally, the expected α-geometry of an

exponential family $\mathcal{E}$ with cumulant function F is given by:

$$g_{ij}(\theta) = \partial_i \partial_j F(\theta), \tag{190}$$

$$\Gamma^{\alpha}_{ij,k} = \frac{1-\alpha}{2} \partial_i \partial_j \partial_k F(\theta). \tag{191}$$

When $\alpha = 1$, $\Gamma^{\alpha}_{ij,k} = 0$ and ∇^1 is flat, and so is ∇^{-1} by using the fundamental theorem of information geometry.

The ± 1-structure can also be derived from a *divergence manifold structure* by choosing the reverse Kullback-Leibler divergence KL^*:

$$(\mathcal{E}, {}_{\mathcal{E}}g, {}_{\mathcal{E}}\nabla^e, {}_{\mathcal{E}}\nabla^m) \equiv (\mathcal{E}, \mathrm{KL}^*). \tag{192}$$

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

$$\mathrm{KL}^*[P_{\theta'} : P_\theta] = \mathrm{KL}[P_\theta : P_{\theta'}] = B_F(\theta' : \theta). \tag{193}$$

The *best exponent error* α^* of the best *Maximum A Priori* (MAP) decision rule is found by minimizing the *Bhattacharyya distance* to get the *Chernoff information* [106]:

$$C[P_1, P_2] = -\log \min_{\alpha \in (0,1)} \int_{x \in \mathcal{X}} p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}\mu(x) \geq 0. \tag{194}$$

On the exponential family manifold $\mathcal{E}$, the *Bhattacharyya distance*:

$$B_\alpha[p_1 : p_2] = -\log \int_{x \in \mathcal{X}} p_1^{\alpha}(x) p_2^{1-\alpha}(x) \mathrm{d}\mu(x), \tag{195}$$

amounts to a *skew Jensen parameter divergence* [83] (also called Burbea-Rao divergence):

$$J_F^{\alpha}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha) F(\theta_2) - F(\theta_1 + (1-\alpha)\theta_2). \tag{196}$$

It can be shown that the Chernoff information (that minimizes α) is equivalent to a Bregman divergence: Namely, the Bregman divergence for exponential families at the optimal exponent value α^*.

Theorem 10 (Chernoff information [73])**.** *The Chernoff information between two distributions belonging to the same exponential family amount to a Bregman divergence:*

$$C[P_{\theta_1} : P_{\theta_2}] = B(\theta_1 : \theta_{12}^{\alpha^*}) = B(\theta_2 : \theta_{12}^{\alpha^*}), \tag{197}$$

where $\theta_{12}^{\alpha} = (1-\alpha)\theta_1 + \alpha\theta_2$*, and* α^* *denote the best exponent error.*

Let $\theta_{12}^* := \theta_{12}^{\alpha^*}$ denote the best exponent error. The geometry [73] of the best error exponent can be explained on the dually flat exponential family manifold as follows:

$$P^* = P_{\theta_{12}^*} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2), \tag{198}$$

where G_e denotes the exponential geodesic γ_{∇^e} and Bi_m the m-bisector:

$$\mathrm{Bi}_m(P_1, P_2) = \{P \ : \ F(\theta_1) - F(\theta_2) + \eta(P)^\top (\theta_2 - \theta_1) = 0\}. \tag{199}$$

Figure 12 illustrates how to retrieve the best error exponent from an exponential arc (θ-geodesic) intersecting the m-bisector.

Furthermore, instead of considering two distributions for this statistical binary decision problem, we may consider a set of n distributions of $P_1, \ldots, P_n \in \mathcal{E}$. The geometry of the error exponent in this multiple hypothesis testing setting has been investigated in [72]. On the dually flat exponential family manifold, it corresponds to check the exponential arcs between *natural neighbors* (sharing Voronoi subfaces) of a Bregman Voronoi diagram [20]. See Figure 13 for an illustration.

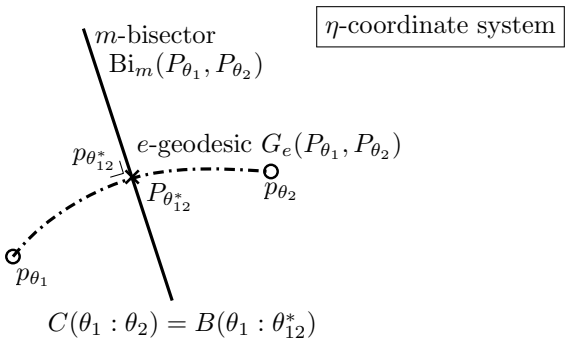


Figure 12: Exact geometric characterization (not necessarily i closed-form) of the best exponent error rate α^*.

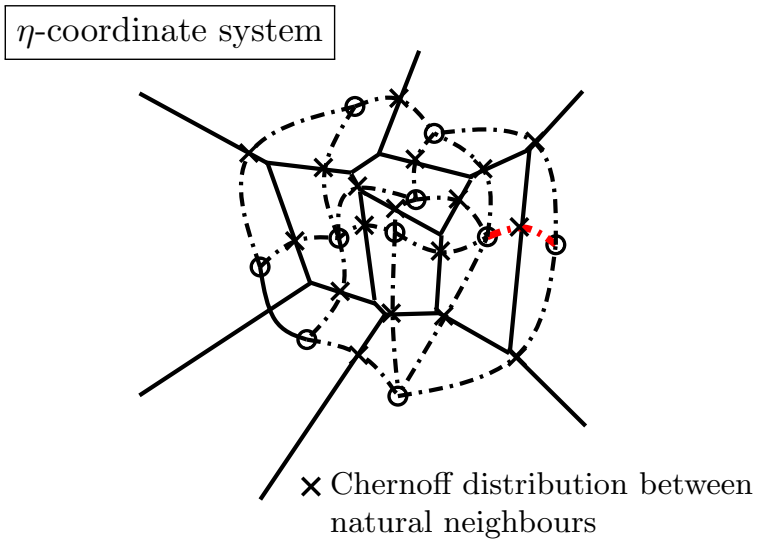


Figure 13: Geometric characterization of the best exponent error rate in the multiple hypothesis testing case.

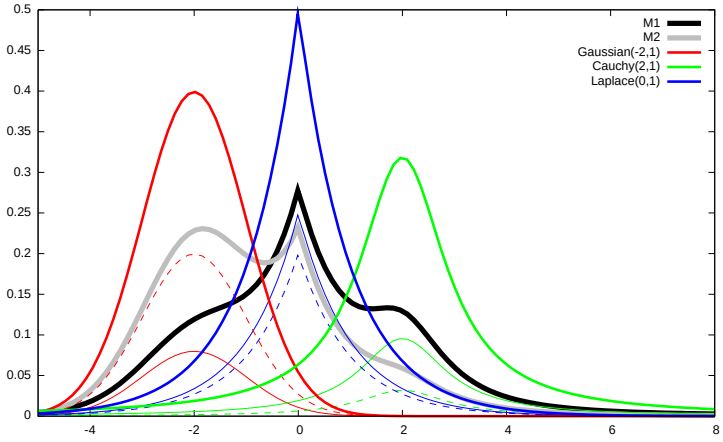


Figure 14: Example of a mixture family of order $D = 2$ (3 components: Laplacian, Gaussian and Cauchy prefixed distributions).

4.4 Clustering mixtures in the dually flat mixture family manifold $(\mathcal{M}, \mathrm{KL})$

Given a set of k prescribed statistical distributions $p_0(x), \ldots, p_{k-1}(x)$, all sharing the same support $\mathcal{X}$ (say, $\mathbb{R}$), a *mixture family* $\mathcal{M}$ of order $D = k - 1$ consists of all *strictly convex combinations* of these component distributions [93]:

$$\mathcal{M} := \left\{ m(x;\theta) = \sum_{i=1}^{k-1} \theta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \theta_i\right) p_0(x) \text{ such that } \theta_i > 0, \sum_{i=1}^{k-1} \theta_i < 1 \right\}. \tag{200}$$

Figure 14 displays two mixtures obtained as convex combinations of prescribed Laplacian, Gaussian and Cauchy component distributions ($D = 2$). When considering a set of prescribed Gaussian component distributions, we obtain a w-Gaussian Mixture Model, or w-GMM for short.

We consider the expected information manifold $(\mathcal{M}, {}^{\mathcal{M}}g, {}^{\mathcal{M}}\nabla^m, {}^{\mathcal{M}}\nabla^e)$ which is dually flat and equivalent to (M_Θ, KL). That is, the KL between two mixtures with prescribed components (w-mixtures, for short) is equivalent to a Bregman divergence for $F(\theta) = -h(m_\theta)$, where $h(p) = \int p(x) \log p(x) \mathrm{d}\mu(x)$ is the differential Shannon information (negative entropy) [93]:

$$\mathrm{KL}[m_{\theta_1} : m_{\theta_2}] = B_F(\theta_1 : \theta_2). \tag{201}$$

Consider a set $\{m_{\theta_1}, \ldots, m_{\theta_n}\}$ of n w-mixtures [93]. Because $F(\theta) = -h(m(x;\theta))$ is the *negative differential entropy* of a mixture (not available in closed form [95]), we approximate the untractable F by another close tractable generator $\tilde{F}$. We use Monte Carlo stochastic sampling to get Monte-Carlo convex $\tilde{F}_{\mathcal{S}}$ for an independent and identically distributed sample $\mathcal{S}$.

Thus we can build a *nested sequence* $(\mathcal{M}, \tilde{F}_{\mathcal{S}_1}), \ldots, (\mathcal{M}, \tilde{F}_{\mathcal{S}_m})$ of tractable dually flat manifolds for nested sample sets $\mathcal{S}_1 \subset \ldots \subset \mathcal{S}_m$ converging to the ideal mixture manifold $(\mathcal{M}, F)$: $\lim_{m\to\infty}(\mathcal{M}, \tilde{F}_{\mathcal{S}_m}) = (\mathcal{M}, F)$ (where convergence is defined with respect to the induced canonical Bregman divergence). A key advantage of this approach is that for a given sample $\mathcal{S}$, all computations carried inside the dually flat manifold $(\mathcal{M}, \tilde{F}_{\mathcal{S}})$ are *consistent*, see [93].

For example, we can apply Bregman k-means [87] on these Monte Carlo dually flat spaces [85] of w-GMMs (Gaussian Mixture Models) to cluster a set of w-GMMs. Figure 15 displays the result of such a clustering.

We have briefly described two applications using dually flat manifolds: (1) the *dually flat exponential manifold* induced by the statistical reverse Kullback-Leibler divergence on an exponential family (structure $(\mathcal{E}, \mathrm{KL}^*)$), and (2) the *dually flat mixture manifold* induced by the statistical Kullback-Leibler divergence

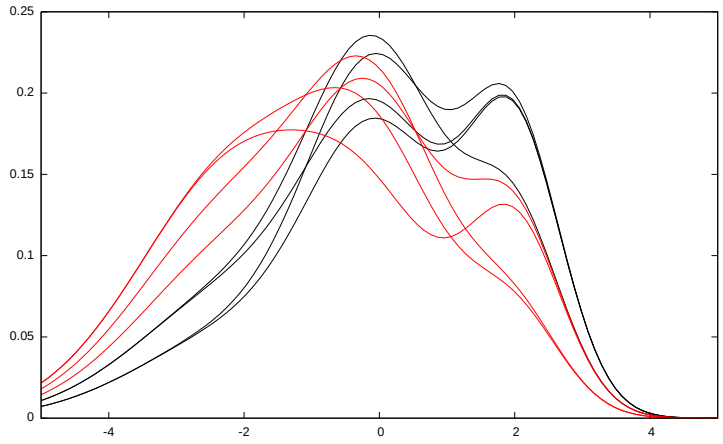


Figure 15: Example of w-GMM clustering into $k = 2$ clusters.

on a mixture family (structure $(\mathcal{M}, \mathrm{KL})$). There are many other dually flat structures that can be met in a statistical context: For example, two other dually flat structures for the D-dimensional probability simplex Δ_D are reported in Amari's textbook [8]: (1) the conformally deforming of the α-geometry (page 88, Eq. 4.95 of [8]), and (2) the χ-escort geometry (page 91, Eq. 4.114 of [8]).

5 Conclusion: Summary, historical background, and perspectives

5.1 Summary

We explained the dualistic nature of information manifolds (M, g, ∇, ∇^*) in information geometry. The dualistic structure is defined by a pair of conjugate connections coupled with the metric tensor that provides a dual parallel transport that preserves the metric. We showed how to extend this structure to a 1-parameter family of structures: From a pair of conjugate connections, the pipeline to build this 1-parameter family of structures can be informally summarized as:

$$(M, g, \nabla, \nabla^*) \Rightarrow (M, g, C) \Rightarrow (M, g, \alpha C) \Rightarrow (M, g, \nabla^{-\alpha}, \nabla^{\alpha}), \quad \forall \alpha \in \mathbb{R}. \tag{202}$$

We stated the fundamental theorem of information geometry on dual constant-curvature manifolds, including the special but important case of dually flat manifolds on which there exists two potential functions and global affine coordinate systems related by the Legendre-Fenchel transformation. Although, information geometry historically started with the Riemannian modeling $(\mathcal{P}, {}_{\mathcal{P}}g)$ of a parametric family of probability distributions $\mathcal{P}$ by letting the metric tensor be the Fisher information matrix, we have emphasized the dualistic view of information geometry which considers non-Riemannian manifolds that can be derived from any divergence, and not necessarily tied to a statistical context (e.g., information manifold can be used in mathematical programming [101]). Let us notice that for any symmetric divergence (e.g. any symmetrized f-divergence like the squared Hellinger divergence), the induced conjugate connections coincide with the Levi-Civita connection but the Fisher-Rao metric distance does not coincide with the squared Hellinger divergence.

On one hand, a Riemannian metric distance D_ρ is *never* a divergence because the rooted distance functions fail to be smooth at the extremities but a squared Riemmanian metric distance is always a divergence. On the other hand, taking the power δ of a divergence D (i.e., D^δ) for some $\delta > 0$ may yield a metric distance (e.g., the square root of the Jensen-Shannon divergence [44]), but this may not always be the case: The powered Jeffreys divergence J^δ is never a metric distance (see [127], page 889). Recently, the *Optimal Transport*

(OT) theory [130] gained interest in statistics and machine learning. But the optimal transport between two members of a same elliptically-contoured family has the same optimal transport formula distance (see [40] Eq. 16 and Eq. 17, although they have different Kullback-Leibler divergences). Another essential difference is that the Fisher-Rao manifold of location-scale families is hyperbolic but the Wasserstein manifold of location-scale families has positive curvature [40, 125].

Notice that we may convert back and forth a similarity $S(p,q) \in (0,1]$ to a dissimilarity $D(p,q) \in [0,\infty)$ as follows:

$$S(p,q) = \exp(-D(p,q)) \in (0,1] \quad (203)$$

$$D(p,q) = -\log S(p,q) \in [0,\infty) \quad (204)$$

When the dissimilarity satisfies the (additive) triangle inequality (i.e., $D(p,q)+D(q,r) \geq D(p,r)$ for any triple (p,q,r)) then the corresponding similarity satisfies the multiplicative triangle inequality: $S(p,q) \times S(q,r) \leq S(p,r)$. A metric transform on a metric distance D is a transformation T such that $T(D(p,q))$ is a metric. The transformation $T(u) = \frac{1}{1+u}$ is a metric transform which bounds potentially unbounded metric distances: That is, if D is an unbounded metric, then $T(D(p,q)) = \frac{D(p,q)}{1+D(p,q)}$ is a bounded metric distance. The transformation $S(u) = u^2$ is not a metric transform since the squared of the Euclidean metric distance is not a metric distance.

5.2 A brief historical review of information geometry

The field of Information Geometry (IG) was historically motivated by providing some differential-geometric structures to statistical models in order to reason geometrically about statistical problems with the endeavor goal of geometrizing mathematical statistics [29, 5, 39, 67, 54, 58, 9, 47]: Professor Harold Hotelling [53] first considered in the late 1920's the *Fisher Information Matrix* (FIM) I as a Riemannian metric tensor g (ie., the *Fisher Information metric*, FIm), and interpreted a parametric family of probability distributions M as a Riemannian manifold (M,g). Historically speaking, Hotelling attended the American Mathematical Society's Annual Meeting in Bethlehem (Pennsylvania, USA) on December 26–29, 1929, but left before his scheduled talk on December 27. His handwritten notes on the "Spaces of Statistical Parameters" was read by a colleague and are fully typeset in [123]. We warmly thank Professor Stigler for sending us the scanned handwritten notes and for discussing by emails some historical aspects of the birth of information geometry. In this pioneering work, Hotelling mentioned that location-scale probability families yield Riemannian manifolds of constant non-positive curvatures. This Riemannian modeling of parametric family of densities was further independently studied by Calyampudi Radhakrishna Rao (C.R. Rao) in his celebrated paper [111] (1945) that also includes the Cramér-Rao lower bound [71] and the Rao-Blackwellization technique used in statistics. Nowadays the induced Riemannian metric distance is often called the *Fisher-Rao distance* [122] or Rao distance [113]. Yet another use of Riemannian geometry in statistics was pioneered by Harold Jeffreys [55] that proposed to use as an invariant prior the normalized volume element of the Fisher-Riemannian manifold. In those seminal papers, there was no theoretical justification of using the Fisher information matrix as a metric tensor (besides the fact that it is a well-defined positive-definite matrix for regular identifiable models). Nowadays, this Riemmanian metric tensor is called the *information metric* for short. Modern information geometry considers a generalization of this approach using a non-Riemannian dualistic modeling (M,g,∇,∇^*) which coincides with the Riemannian manifold when $\nabla = \nabla^* = {}^{\mathrm{LC}}\nabla$, the Levi-Civita connection (the unique torsion-free affine connection compatible with the metric tensor). The Fisher-Rao geometry has also been explored in thermodynamics yielding the Ruppeiner geometry [134], and the geometry of thermodynamics is called nowadays called *geometrothermodynamics* [109].

In the 1960's, Nikolai Chentsov (also commonly written Čencov) studied the algebraic category of all statistical decision rules with its induced geometric structures: Namely, the α-geometries ("equivalent differential geometry") and the dually flat manifolds ("Nonsymmetric Pythagorean geometry" of the exponential families with respect to the Kullback-Leibler divergence). In the preface of the english translation of his 1972's russian monograph [29], the field of investigation is defined as "geometrical statistics." However in the original Russian monograph, Chentsov used the russian term *geometrostatistics*. According to Professor

Alexander Holevo, the geometrostatistics term was coined by Andrey Kolmogorov to define the field of differential geometry of statistical models. In the monograph of Chentsov [29], the Fisher information metric is shown to be the unique metric tensor (up to a scaling factor) yielding statistical invariance under Markov morphisms (see [26] for a simpler proof that generalizes to positive measures).

The dual nature of the information geometry was thoroughly investigated by Professor Shun-ichi Amari [4]. In the preface of his 1985's monograph [5], Professor Amari coined the term *information geometry* as follows: "The differential-geometrical method developed in statistics is also applicable to other fields of sciences such as information theory and systems theory... They together will open a new field, which I would like to call *information geometry*." Professor Amari mentioned in [5] that he considered the Gaussian Riemannian manifold as a hyperbolic manifold in 1959, and was strongly influenced by Efron's paper on statistical curvature [41] (1975) to study the family of α-connections in the 1980's [4, 68]. Professor Amari prepared his PhD under the supervision of Professor Kondo [31], an expert of differential geometry in touch with Professor Kawaguchi [59]. The role of differential geometry in statistics has been discussed in [17].

Note that the dual affine connections of information geometry have also been investigated independently in *affine differential geometry* [99] which considers invariance under volume-preserving affine transformations by defining a volume form instead of a metric form for Riemannian geometry. The notion of dual parallel transport compatible with the metric is due to Aleksandr Norden [100] and Rabindra Nath Sen [116, 117, 118] (See the Senian geometry in `http://insaindia.res.in/detail/N54-0728`).

We summarize the main fundamental structures of information manifolds below:

(M, g)	Riemannian manifold
$(\mathcal{P}, {}_{\mathcal{P}}g)$	Fisher-Riemannian (expected) Riemannian manifold
(M, g, ∇)	Riemannian manifold (M, g) with affine connection ∇
$(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}{}^{e}\nabla^{\alpha})$	Chentsov's manifold with affine exponential α-connection
(M, g, ∇, ∇^*)	Amari's dualistic information manifold
$(\mathcal{P}, {}_{\mathcal{P}}g, {}_{\mathcal{P}}\nabla^{-\alpha}, {}_{\mathcal{P}}\nabla^{\alpha})$	Amari's (expected) information α-manifold, α-geometry
(M, g, C)	Lauritzen's statistical manifold [62]
$(M, {}^{D}g, {}^{D}\nabla, {}^{D^*}\nabla)$	Eguchi's conjugate connection manifold induced by divergence D
$(M, {}^{F}g, {}^{F}C)$	Chentsov/Amari's dually flat manifold induced by convex potential F

We use the $\equiv$ symbol to denote the equivalence of geometric structures. For example, we have $(M, g) \equiv (M, g, {}^{\mathrm{LC}}\nabla, {}^{\mathrm{LC}}\nabla^* = {}^{\mathrm{LC}}\nabla)$.

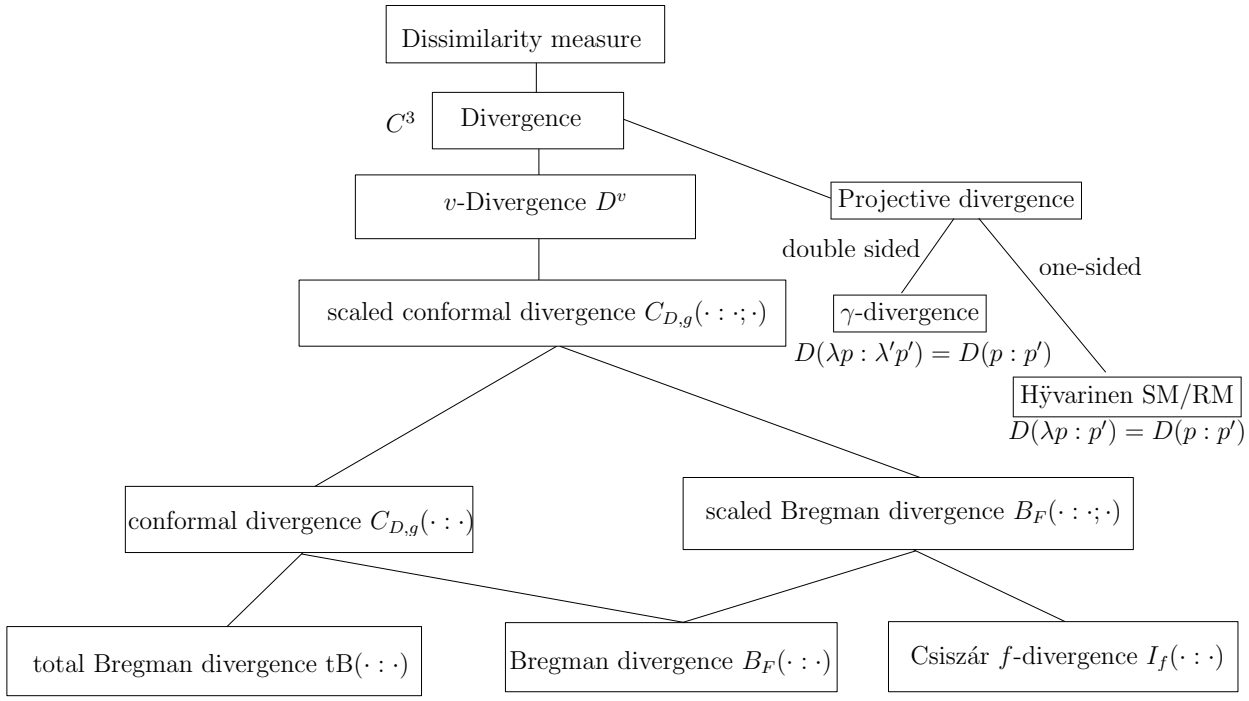
5.3 Perspectives

We recommend the two recent textbooks [25, 8] for an indepth covering of (parametric) information geometry, and the book [48] for a thorough description of some infinite-dimensional statistical models. (Japanese readers may refer to [7, 45].) We did not report the various coefficients of the metric tensors, Christoffel symbols and skewness tensors for the expected α-geometry of common parametric models like the multivariate Gaussian distributions, the Gamma/Beta distributions, etc. They can be found in [10, 25] and in various articles dealing with less common family of distributions [64, 65, 141, 105, 142, 104, 10]. Although we have focused on the finite parametric setting, information geometry is also considering non-parametric families of distributions [107], and quantum information geometry [51].

We have shown that we can always create an information manifold (M, D) from *any* divergence function D. It is therefore important to consider generic classes of divergences in applications, that are ideally axiomatized and shown to have exhaustive characteristics. Beyond the three main Bregman/Csiszár/Jensen classes (theses classes overlap [102]), we may also mention the class of conformal divergences [98, 91], the class of projective divergences [92, 97], etc. Figure 16 illustrates the relationships between the principal classes of distances.

There are many perspectives on information geometry as attested by the new Springer journal[13], and the biannual international conference "Geometric Sciences of Information" (GSI) [80, 81, 82] with its collective

[13] 'Information Geometry', `https://www.springer.com/mathematics/geometry/journal/41884`



$D^v(P:Q)=D(v(P):v(Q))$

$I_f(P:Q)=\int p(x) f\left(\left(\frac{q(x)}{p(x)}\right)\right) \mathrm{d}\nu(x)$

$B_F(P:Q)=F(P)-F(Q)-\langle P-Q,\nabla F(Q)\rangle$

$\mathrm{tB}_F(P:Q)=\frac{B_F(P:Q)}{\sqrt{1+\|\nabla F(Q)\|^2}}$

$C_{D,g}(P:Q)=g(Q)D(P:Q)$

$B_{F,g}(P:Q;W)=WB_F\left(\frac{P}{Q}:\frac{Q}{W}\right)$

Figure 16: Principled classes of distances/divergences

post-conference edited books [75, 74]. We also mention the edited book [13] on the Occasion of Shun-ichi Amari's 80th birthday.
Acknowledgments:FN would like to thank the organizers of the *Geometry In Machine Learning* workshop in 2018 (GiMLi, `http://gimli.cc/2018/`) for their kind keynote talk invitation, and specially Professor Søren Hauberg (Technical University of Denmark, DTU). This survey is based on the talk given at GiMLi. I am very thankful to Professor Stigler (University of Chicago, USA) and Professor Holevo (Steklov Mathematical Institute, Russia) for providing me feedback on some historical aspects of the field of information geometry. Finally, I would like to express my sincere thanks to Gaëtan Hadjeres (Sony Computer Science Laboratories Inc, Paris) for his careful proofreading and feedback.

A Monte Carlo estimations of f-divergences

Let (X, F, μ) be a probability space [60] with X denoting the sample space, F the σ-algebra, and μ a reference positive measure. The f-divergence [33, 90] between two probability measures P and Q both absolutely continuous with respect to a positive measure μ for a convex generator $f : (0, \infty) \to \mathbb{R}$ strictly convex at 1 and satisfying $f(1) = 0$ is

$$I_f(P : Q) = I_f(p : q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x), \tag{205}$$

where $P = p\mathrm{d}\mu$ and $Q = q\mathrm{d}\mu$ (i.e., p and q denote the Radon-Nikodym derivatives with respect to μ). We use the following conventions:

$$0 f\left(\frac{0}{0}\right) = 0, \quad f(0) = \lim_{u \to 0^+} f(u), \quad \forall a > 0, 0 f\left(\frac{a}{0}\right) = \lim_{u \to 0^+} u f\left(\frac{a}{u}\right) = a \lim_{u \to \infty} \frac{f(u)}{u}. \tag{206}$$

When $f(u) = -\log u$, we retrieve the Kullback-Leibler divergence (KLD):

$$D_{\mathrm{KL}}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x). \tag{207}$$

The KLD is usually difficult to calculate in closed-form, say, for example, between statistical mixture models [96]. A common technique is to estimate the KLD using Monte Carlo sampling using a proposal distribution r:

$$\widehat{\mathrm{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^{n} \frac{p(x_i)}{r(x_i)} \log \frac{p(x_i)}{q(x_i)}, \tag{208}$$

where $x_1, \ldots, x_n \sim_{\mathrm{iid}} r$. When r is chosen as p, the KLD can be estimated as:

$$\widehat{\mathrm{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i)}{q(x_i)}. \tag{209}$$

Monte Carlo estimators are consistent under mild conditions: $\lim_{n \to \infty} \widehat{\mathrm{KL}}_n(p : q) = \mathrm{KL}(p : q)$.

In practice, one problem when implementing Eq. 209, is that we may end up potentially with $\widehat{\mathrm{KL}}_n(p : q) < 0$. This may have disastrous consequences as algorithms implemented by programs consider non-negative divergences to execute a correct workflow. The potential negative value problem of Eq. 209 comes from the fact that $\sum_i p(x_i) \neq 1$ and $\sum_i q(x_i) \neq 1$. One way to circumvent this problem is to consider the *extended f-divergences*:

Definition 7 (Extended f-divergence)**.** *The extended f-divergence for a convex generator f, strictly convex at 1 and satisfying $f(1) = 0$ is defined by*

$$I_f^e(p : q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1) \left(\frac{q(x)}{p(x)} - 1\right)\right) \mathrm{d}\mu(x). \tag{210}$$

Indeed, for a strictly convex generator f, let us consider the *scalar Bregman divergence* [23]:

$$B_f(a:b) = f(a) - f(b) - (a-b)f'(b) \geq 0. \tag{211}$$

Setting $a = \frac{q(x)}{p(x)}$ and $b = 1$ in Eq. 211, and using the fact that $f(1) = 0$, we get

$$f\left(\frac{q(x)}{p(x)}\right) - \left(\frac{q(x)}{p(x)} - 1\right) f'(1) \geq 0. \tag{212}$$

Therefore we define the *extended f-divergences* as

$$I_f^e(p:q) = \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) \mathrm{d}\mu(x) \geq 0. \tag{213}$$

That is, the formula for the extended f-divergences is

$$I_f^e(p:q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right) \mathrm{d}\mu(x) \geq 0. \tag{214}$$

Then we estimate the extended f-divergence using importance sampling of the integral with respect to distribution r, using n variates $x_1, \ldots, x_n \sim_{\mathrm{iid}} p$ as:

$$\hat{I}_{f,n}(p:q) = \frac{1}{n}\sum_{i=1}^n f\left(\frac{q(x_i)}{p(x_i)}\right) - f'(1)\left(\frac{q(x_i)}{p(x_i)} - 1\right) \geq 0. \tag{215}$$

For example, for the KLD, we obtain the following Monte Carlo estimator:

$$\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n}\sum_{i=1}^n \left(\log\frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1\right) \geq 0, \tag{216}$$

since the extended KLD is

$$D_{\mathrm{KL}^e}(p:q) = \int \left(p(x)\log\frac{p(x)}{q(x)} + q(x) - p(x)\right) \mathrm{d}\mu(x). \tag{217}$$

Eq. 216 can be interpreted as a sum of scalar Itakura-Saito divergences since the Itakura-Saito divergence is scale-invariant: $\widehat{\mathrm{KL}}_n(p:q) = \frac{1}{n}\sum_{i=1}^n D_{\mathrm{IS}}(p(x_i):q(x_i))$ with the scalar Itakura-Saito divergence

$$D_{\mathrm{IS}}(a:b) = D_{\mathrm{IS}}\left(\frac{a}{b}:1\right) = \frac{a}{b} - \log\frac{a}{b} - 1 \geq 0, \tag{218}$$

a Bregman divergence obtained for the generator $f(u) = -\log u$.

Notice that the extended f-divergence is a f-divergence for the generator

$$f_e(u) = f(u) - f'(1)(u-1). \tag{219}$$

We check that the generator f_e satisfies both $f(1) = 0$ and $f'(1) = 0$, and we have $I_f^e(p:q) = I_{f_e}(p:q)$. Thus $D_{\mathrm{KL}^e}(p:q) = I_{f_{\mathrm{KL}}^e}(p:q)$ with $f_{\mathrm{KL}}^e(u) = -\log u + u - 1$.

Let us remark that we only need to have the scalar function strictly convex at 1 to ensure that $B_f\left(\frac{a}{b}:1\right) \geq 0$. Indeed, we may use the definition of Bregman divergences extended to strictly convex functions but not necessarily smooth functions [50, 126]:

$$B_f(x:y) = \max_{g(y)\in\partial f(y)} \{f(x) - f(y) - (x-y)g(y)\}, \tag{220}$$

where $\partial f(y)$ denotes the subderivative of f at y.

Furthermore, noticing that $I_{\lambda f}(p:q) = \lambda I_f(p:q)$ for $\lambda > 0$, we may enforce that $f''(1) = 1$, and obtain a *standard f-divergence* [8] which enjoys the property that $I_f(p_\theta(x) : p_{\theta+\mathrm{d}\theta}(x)) = \mathrm{d}\theta^\top I(\theta)\mathrm{d}\theta$, where $I(\theta)$ denotes the Fisher information matrix of the parameteric family $\{p_\theta\}_\theta$ of densities.

B The multivariate Gaussian family: An exponential family

We report the canonical decomposition of the multivariate Gaussian [136] family $\{N(\mu, \Sigma)$ such that $\mu \in \mathbb{R}^d, \Sigma \succ 0\}$ following [78]. The multivariate Gaussian family is also called the *MultiVariate Normal* family, or MVN family for short.

Let $\lambda := (\lambda_v, \lambda_M) = (\mu, \Sigma)$ denote the *composite* (vector,matrix) parameter of an MVN. The d-dimensional MVN density is given by

$$p_\lambda(x; \lambda) \quad := \quad \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2}(x-\lambda_v)^\top \lambda_M^{-1}(x-\lambda_v)\right), \tag{221}$$

where $|\cdot|$ denotes the matrix determinant. The natural parameters θ are also expressed using both a vector parameter θ_v and a matrix parameter θ_M in a compound object $\theta = (\theta_v, \theta_M)$. By defining the following *compound inner product* on a composite (vector,matrix) object

$$\langle \theta, \theta' \rangle := \theta_v^\top \theta'_v + \mathrm{tr}\left({\theta'_M}^\top \theta_M\right), \tag{222}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace, we rewrite the MVN density of Equation (221) in the canonical form of an exponential family [84]:

$$p_\theta(x; \theta) \quad := \quad \exp\left(\langle t(x), \theta\rangle - F_\theta(\theta)\right) = p_\lambda(x; \lambda(\theta)), \tag{223}$$

where

$$\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right) = \theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, -\frac{1}{2}\lambda_M^{-1}\right), \tag{224}$$

is the *compound natural parameter* and

$$t(x) = (x, -xx^\top) \tag{225}$$

is the *compound sufficient statistic*. The function F_θ is the strictly convex and continuously differentiable log-normalizer defined by:

$$F_\theta(\theta) = \frac{1}{2}\left(d\log\pi - \log|\theta_M| + \frac{1}{2}\theta_v^\top \theta_M^{-1}\theta_v\right), \tag{226}$$

The log-normalizer can be expressed using the ordinary parameters, $\lambda = (\mu, \Sigma)$, as:

$$F_\lambda(\lambda) \quad = \quad \frac{1}{2}\left(\lambda_v^\top \lambda_M^{-1}\lambda_v + \log|\lambda_M| + d\log 2\pi\right), \tag{227}$$

$$\quad = \quad \frac{1}{2}\left(\mu^\top \Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right). \tag{228}$$

The *moment/expectation parameters* [8] are

$$\eta = (\eta_v, \eta_M) = E[t(x)] = \nabla F(\theta). \tag{229}$$

We report the conversion formula between the three types of coordinate systems (namely the ordinary parameter λ, the natural parameter θ and the moment parameter η) as follows:

$$\begin{cases} \theta_v(\lambda) = \lambda_M^{-1}\lambda_v = \Sigma^{-1}\mu \\ \theta_M(\lambda) = \frac{1}{2}\lambda_M^{-1} = \frac{1}{2}\Sigma^{-1} \end{cases} \quad \Leftrightarrow \quad \begin{cases} \lambda_v(\theta) = \frac{1}{2}\theta_M^{-1}\theta_v = \mu \\ \lambda_M(\theta) = \frac{1}{2}\theta_M^{-1} = \Sigma \end{cases} \tag{230}$$

$$\begin{cases} \eta_v(\theta) = \frac{1}{2}\theta_M^{-1}\theta_v \\ \eta_M(\theta) = -\frac{1}{2}\theta_M^{-1} - \frac{1}{4}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^\top \end{cases} \quad \Leftrightarrow \quad \begin{cases} \theta_v(\eta) = -(\eta_M + \eta_v\eta_v^\top)^{-1}\eta_v \\ \theta_M(\eta) = -\frac{1}{2}(\eta_M + \eta_v\eta_v^\top)^{-1} \end{cases} \tag{231}$$

$$\begin{cases} \lambda_v(\eta) = \eta_v = \mu \\ \lambda_M(\eta) = -\eta_M - \eta_v\eta_v^\top = \Sigma \end{cases} \quad \Leftrightarrow \quad \begin{cases} \eta_v(\lambda) = \lambda_v = \mu \\ \eta_M(\lambda) = -\lambda_M - \lambda_v\lambda_v^\top = -\Sigma - \mu\mu^\top \end{cases} \tag{232}$$

The dual Legendre convex conjugate [8] is

$$F_\eta^*(\eta) = -\frac{1}{2}\left(\log(1+\eta_v^\top \eta_M^{-1}\eta_v) + \log|-\eta_M| + d(1+\log 2\pi)\right), \tag{233}$$

and $\theta = \nabla_\eta F_\eta^*(\eta)$.

We check the Fenchel-Young equality when $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$:

$$F_\theta(\theta) + F_\eta^*(\eta) - \langle \theta, \eta\rangle = 0. \tag{234}$$

The KullbackLeibler divergence between two d-dimensional Gaussians distributions $p_{(\mu_1,\Sigma_1)}$ and $p_{(\mu_2,\Sigma_2)}$ (with $\Delta_\mu = \mu_2 - \mu_1$) is

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) = \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \Delta_\mu^\top \Sigma_2^{-1}\Delta_\mu + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\right\} = \mathrm{KL}(p_{\lambda_1} : p_{\lambda_2}). \tag{235}$$

We check that $\mathrm{KL}(p_{(\mu,\Sigma)} : p_{(\mu,\Sigma)}) = 0$ since $\Delta_\mu = 0$ and $\mathrm{tr}(\Sigma^{-1}\Sigma) = \mathrm{tr}(I) = d$. Notice that when $\Sigma_1 = \Sigma_2 = \Sigma$, we have

$$\mathrm{KL}(p_{(\mu_1,\Sigma)} : p_{(\mu_2,\Sigma)}) = \frac{1}{2}\Delta_\mu^\top \Sigma^{-1}\Delta_\mu = \frac{1}{2}D^2_{\Sigma^{-1}}(\mu_1,\mu_2), \tag{236}$$

that is half the squared Mahalanobis distance for the precision matrix Σ^{-1} (a positive-definite matrix: $\Sigma^{-1} \succ 0$), where the Mahalanobis distance is defined for any positive matrix $Q \succ 0$ as follows:

$$D_Q(p_1 : p_2) = \sqrt{(p_1 - p_2)^\top Q (p_1 - p_2)}. \tag{237}$$

The KullbackLeibler divergence between two probability densities of the same exponential families amount to a Bregman divergence [8]:

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) = \mathrm{KL}(p_{\lambda_1} : p_{\lambda_2}) = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2), \tag{238}$$

where the Bregman divergence is defined by

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - \langle \theta - \theta', \nabla F(\theta')\rangle, \tag{239}$$

with $\eta' = \nabla F(\theta')$. Define the canonical divergence [8]

$$A_F(\theta_1 : \eta_2) = F(\theta_1) + F^*(\eta_2) - \langle \theta_1, \eta_2\rangle = A_{F^*}(\eta_2 : \theta_1), \tag{240}$$

since $F^{**} = F$. We have $B_F(\theta_1 : \theta_2) = A_F(\theta_1 : \eta_2)$.

Notations

Below is a list of notations we used in this document:

$[D]$	$[D] := \{1, \ldots, D\}$
$\langle \cdot, \cdot \rangle$	inner product
$M_Q(u,v) = \|u - v\|_Q$	Mahalanobis distance $M_Q(u,v) = \sqrt{\sum_{i,j}(u^i - v^i)(u^j - v^j)Q_{ij}}$, $Q \succ 0$
$D(\theta : \theta')$	parameter divergence
$D[p(x) : p'(x)]$	statistical divergence
D, D^*	Divergence and dual (reverse) divergence

Csiszár divergence I_f	$I_f(\theta:\theta') := \sum_{i=1}^D \theta_i f\left(\frac{\theta'_i}{\theta_i}\right)$ with $f(1)=0$
Bregman divergence B_F	$B_F(\theta:\theta') := F(\theta) - F(\theta') - (\theta-\theta')^\top \nabla F(\theta')$
Canonical divergence A_{F,F^*}	$A_{F,F^*}(\theta:\eta') = F(\theta) + F^*(\eta') - \theta^\top \eta'$
Bhattacharyya distance	$B_\alpha[p_1:p_2] = -\log \int_{x\in\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \mathrm{d}\mu(x)$
Jensen/Burbea-Rao divergence	$J_F^{(\alpha)}(\theta_1:\theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\theta_1 + (1-\alpha)\theta_2)$
Chernoff information	$C[P_1,P_2] = -\log \min_{\alpha\in(0,1)} \int_{x\in\mathcal{X}} p_1^\alpha(x) p_2^{1-\alpha}(x) \mathrm{d}\mu(x)$
F, F^*	Potential functions related by Legendre-Fenchel transformation
$D_\rho(p,q)$	Riemannian distance $D_\rho(p,q) := \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \mathrm{d}t$
B, B^*	basis, reciprocal basis
$B = \{e_1 = \partial_1, \ldots, e_D = \partial_D\}$	natural basis
$\{\mathrm{d}x_i\}_i$	covector basis (one-forms)
$(v)_B := (v^i)$	contravariant components of vector v
$(v)_{B^*} := (v_i)$	covariant components of vector v
$u \perp v$	vector u is perpendicular to vector v ($\langle u, v\rangle = 0$)
$\|v\| = \sqrt{\langle v, v\rangle}$	induced norm, length of a vector v
M, S	Manifold, submanifold
T_p	tangent plane at p
TM	Tangent bundle $TM = \cup_p T_p = \{(p,v), p\in M, v\in T_p\}$
$\mathfrak{F}(M)$	space of smooth functions on M
$\mathfrak{X}(M) = \Gamma(TM)$	space of smooth vector fields on M
vf	direction derivative of f with respect to vector v
$X, Y, Z \in \mathfrak{X}(M)$	Vector fields
$g \overset{\Sigma}{=} g_{ij} \mathrm{d}x_i \otimes \mathrm{d}x_j$	metric tensor (field)
$(\mathcal{U}, x)$	local coordinates x in a chat $\mathcal{U}$
$\partial_i :=: \frac{\partial}{\partial x_i}$	natural basis vector
$\partial^i :=: \frac{\partial}{\partial x^i}$	natural reciprocal basis vector
∇	affine connection
$\nabla_X Y$	covariant derivative
$\prod_c^\nabla$	parallel transport of vectors along a smooth curve c
$\prod_c^\nabla v$	Parallel transport of $v \in T_{c(0)}$ along a smooth curve c
γ, γ_∇	geodesic, geodesic with respect to connection ∇
$\Gamma_{ij,l}$	Christoffel symbols of the first kind (functions)
Γ_{ij}^k	Christoffel symbols of the second kind (functions)
R	Riemann-Christoffel curvature tensor
$[X,Y]$	Lie bracket $[X,Y](f) = X(Y(f)) - Y(X(f)), \forall f \in \mathfrak{F}(M)$
∇-projection	$P_S = \arg\min_{Q\in S} D(\theta(P):\theta(Q))$
∇^*-projection	$P_S^* = \arg\min_{Q\in S} D(\theta(Q):\theta(P))$
C	Amari-Chentsov totally symmetric cubic 3-covariant tensor
$\mathcal{P} = \{p_\theta(x)\}_{\theta_i n\Theta}$	parametric family of probability distributions
$\mathcal{E}, \mathcal{M}, \Delta_D$	exponential family, mixture family, probability simplex
${}_\mathcal{P}I(\theta)$ Fisher information matrix	
${}_\mathcal{P}I(\theta)$	Fisher Information Matrix (FIM) for a parametric family $\mathcal{P}$

${}_{\mathcal{P}}g$	Fisher information metric tensor field
exponential connection ${}^{e}_{\mathcal{P}}\nabla$	${}^{e}_{\mathcal{P}}\nabla := E_\theta\left[(\partial_i\partial_j l)(\partial_k l)\right]$
mixture connection ${}^{m}_{\mathcal{P}}\nabla$	${}^{m}_{\mathcal{P}}\nabla := E_\theta\left[(\partial_i\partial_j l + \partial_i l\partial_j l)(\partial_k l)\right]$
expected skewness tensor C_{ijk}	$C_{ijk} := E_\theta\left[\partial_i l\partial_j l\partial_k l\right]$
expected α-connections	${}_{\mathcal{P}}\Gamma^{\alpha}{}^{k}_{ij} := -\frac{1+\alpha}{2}C_{ijk} = E_\theta\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l\partial_j l\right)(\partial_k l)\right]$
$\equiv$	equivalence of geometric structures

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[3] Maks Aizikovich Akivis and Boris Abramovich Rosenfeld. *Élie Cartan (1869-1951)*, volume 123. American Mathematical Society, 2011.

[4] Shun-ichi Amari. Theory of information spaces: A differential geometrical foundation of statistics. *Post RAAG Reports*, 1980.

[5] Shun-ichi Amari. Differential-geometrical methods in statistics. *Lecture Notes on Statistics*, 28, 1985. second edition in 1990.

[6] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[7] Shun-ichi Amari. *New developments of information geometry*. Saiensu'sha, Tokyo, 2014. Jouhou kikagaku no shintenkai (in Japanese).

[8] Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[9] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2007.

[10] Khadiga A. Arwini and Christopher Terence John Dodson. *Information Geometry: Near Randomness and Near Independance*. Springer, 2008.

[11] John Ashburner and Karl J. Friston. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage*, 55(3):954–967, 2011.

[12] Nihat Ay and Shun-ichi Amari. A novel approach to canonical divergences within information geometry. *Entropy*, 17(12):8111–8129, 2015.

[13] Nihat Ay, Paolo Gibilisco, and Frantisek Matús. *Information Geometry and its Applications: On the Occasion of Shun-ichi Amari's 80th Birthday*, volume 252 of *Springer Proceedings in Mathematics & Statistics*. Springer, 2018. following the June 2016 event at Liblice, Czech Republic.

[14] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[15] John C. Baez and Derek K. Wise. Teleparallel gravity as a higher gauge theory. *Communications in Mathematical Physics*, 333(1):153–186, 2015.

[16] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[17] Ole E. Barndorff-Nielsen, David Roxbee Cox, and Nancy Reid. The role of differential geometry in statistical theory. *International Statistical Review*, pages 83–96, 1986.

[18] Arnaud Berny. Selection and reinforcement learning for combinatorial optimization. In *International Conference on Parallel Problem Solving from Nature*, pages 601–610. Springer, 2000.

[19] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies–a comprehensive introduction. *Natural computing*, 1(1):3–52, 2002.

[20] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281–307, 2010.

[21] Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[22] Jean-Pierre Bourguignon. Ricci curvature and measures. *Japanese Journal of Mathematics*, 4(1):27–45, 2009.

[23] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[24] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[25] Ovidiu Calin and Constantin Udriste. *Geometric Modeling in Probability and Statistics*. Mathematics and Statistics. Springer International Publishing, 2014.

[26] L. Lorne Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

[27] Elie Joseph Cartan. *On manifolds with an affine connection and the theory of general relativity*. Bibliopolis, 1986.

[28] AL Cauchy. Methode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences*, 25:536–538, 1847.

[29] Nikolai N. Chentsov. Statistical decision rules and optimal inference. *Monographs, American Mathematical Society, Providence, RI*, 1982.

[30] JM Corcuera and Federica Giummolè. A characterization of monotone and regular divergences. *Annals of the Institute of Statistical Mathematics*, 50(3):433–450, 1998.

[31] Grenville J Croll. The natural philosophy of Kazuo Kondo. *arXiv preprint arXiv:0712.0641*, 2007.

[32] Jean-Pierre Crouzeix. A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13(1):364–365, 1977.

[33] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[34] Imre Csiszár and Paul C Shields. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

[35] Haskell B Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.

[36] Anand Ganesh Dabak. *A geometry for detection theory*. PhD thesis, Rice University, 1993.

[37] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):380–393, 1997.

[38] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

[39] Christopher Terence John Dodson, editor. *Geometrization of statistical theory*. ULDM Publications, 1987. University of Lancaster, Department of Mathematics.

[40] D. C. Dowson and Basil V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[41] Bradley Efron et al. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975.

[42] Shinto Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, pages 793–803, 1983.

[43] Shinto Eguchi et al. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima mathematical journal*, 15(2):341–391, 1985.

[44] Bent Fuglede and Flemming Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium on Information Theory (ISIT)*, page 31. IEEE, 2004.

[45] Akio Fujiwara. *Foundations of Information Geometry*. Makino Shoten, Tokyo, 2015. Jouhou kikagaku no kisou (in Japanese).

[46] Hitoshi Furuhata. Hypersurfaces in statistical manifolds. *Differential Geometry and its Applications*, 27(3):420–429, 2009.

[47] Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn, editors. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009.

[48] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

[49] Erika Gomes-Gonçalves, Henryk Gzyl, and Frank Nielsen. Geometry and fixed-rate quantization in riemannian metric spaces induced by separable Bregman divergences. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information - 4th International Conference, GSI 2019, Toulouse, France, August 27-29, 2019, Proceedings*, volume 11712 of *Lecture Notes in Computer Science*, pages 351–358. Springer, 2019.

[50] G. J. Gordon. *Approximate solutions to Markov decision processes*. PhD thesis, Department of Computer Science, Carnegie Mellon University, 1999.

[51] Masahito Hayashi. *Quantum information*. Springer, 2006.

[52] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[53] Harold Hotelling. Spaces of statistical parameters. *Bulletin of the American Mathematical Society (AMS)*, 36:191, 1930.

[54] Shun ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. Iwanami Shoten, Japan, 1993. Jouhou kika no houhou (in Japanese).

[55] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, 186(1007):453–461, 1946.

[56] Jiantao Jiao, Thomas A Courtade, Albert No, Kartik Venkat, and Tsachy Weissman. Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626, 2014.

[57] Satoshi Kakihara, Atsumi Ohara, and Takashi Tsuchiya. Information geometry and interior-point algorithms in semidefinite programs and symmetric cone programs. *J. Optim. Theory Appl.*, 157(3):749–780, 2013.

[58] Robert E. Kass and Paul W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley-Interscience, 07 1997.

[59] Michiaki Kawaguchi. An introduction to the theory of higher order spaces I. the theory of Kawaguchi spaces. *RAAG Memoirs*, 3:718–734, 1960.

[60] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

[61] Takashi Kurose. On the divergences of 1-conformally flat statistical manifolds. *Tohoku Mathematical Journal, Second Series*, 46(3):427–433, 1994.

[62] Stefan L. Lauritzen. Statistical manifolds. *Differential geometry in statistical inference*, 10:163–216, 1987.

[63] Luigi Malagò and Giovanni Pistone. Information geometry of the Gaussian distribution in view of stochastic optimization. In *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*, pages 150–162, 2015.

[64] Ann F. S. Mitchell. Statistical manifolds of univariate elliptic distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 1–16, 1988.

[65] Ann F. S. Mitchell. The information matrix, skewness tensor and α-connections for the general multivariate elliptic distribution. *Annals of the Institute of Statistical Mathematics*, 41(2):289–304, 1989.

[66] Uwe Mühlich. *Fundamentals of tensor calculus for engineers with a primer on smooth manifolds*, volume 230. Springer, 2017.

[67] Michael Murray and John Rice. *Differential geometry and statistics*. Number 48 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1993.

[68] Hiroshi Nagaoka and Shun-ichi Amari. Differential geometry of smooth families of probability distributions. Technical report, University of Tokyo, 1982. METR 82-7.

[69] Jan Naudts and Jun Zhang. Rho–tau embedding and gauge freedom in information geometry. *Information Geometry*, Aug 2018.

[70] Frank Nielsen. Legendre transformation and information geometry, 2010.

[71] Frank Nielsen. Cramér-Rao lower bound and information geometry. In *Connected at Infinity II*, pages 18–37. Springer, 2013.

[72] Frank Nielsen. Hypothesis testing, information divergence and computational geometry. In *GSI*, pages 241–248, 2013.

[73] Frank Nielsen. An information-geometric characterization of Chernoff information. *IEEE SPL*, 20(3):269–272, 2013.

[74] Frank Nielsen. *Geometric Theory of Information*. Springer, 2014.

[75] Frank Nielsen. *Geometric Structures of Information*. Springer, 2018.

[76] Frank Nielsen. What is... an information projection? *Notices of the AMS*, 65(3)(10):321–324, 2018.

[77] Frank Nielsen. On geodesic triangles with right angles in a dually flat space. *arXiv preprint arXiv:1910.03935*, 2019.

[78] Frank Nielsen. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

[79] Frank Nielsen. On Voronoi diagrams on the information-geometric Cauchy manifolds. *Entropy*, 22(7):713, 2020.

[80] Frank Nielsen and Frédéric Barbaresco, editors. *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*. Springer, 2013.

[81] Frank Nielsen and Frédéric Barbaresco, editors. *Geometric Science of Information*, volume 9389 of *Lecture Notes in Computer Science*. Springer, 2015.

[82] Frank Nielsen and Frédéric Barbaresco, editors. *Geometric Science of Information*, volume 10589 of *Lecture Notes in Computer Science*. Springer, 2017.

[83] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8), 2011.

[84] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[85] Frank Nielsen and Gaëtan Hadjeres. Monte Carlo information geometry: The dually flat case. *CoRR*, abs/1803.07225, 2018.

[86] Frank Nielsen and Gaëtan Hadjeres. Monte Carlo information-geometric structures. In *Geometric Structures of Information*, pages 69–103. Springer, 2019.

[87] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6), 2009.

[88] Frank Nielsen and Richard Nock. Entropies and cross-entropies of exponential families. In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE, 2010.

[89] Frank Nielsen and Richard Nock. Hyperbolic Voronoi diagrams made easy. In *International Conference on Computational Science and Its Applications (ICCSA)*, pages 74–80. IEEE, 2010.

[90] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.

[91] Frank Nielsen and Richard Nock. Total Jensen divergences: Definition, properties and clustering. In *ICASSP*, pages 2016–2020, 2015.

[92] Frank Nielsen and Richard Nock. Patch matching with polynomial exponential families and projective divergences. In *SISAP*, pages 109–116, 2016.

[93] Frank Nielsen and Richard Nock. On the geometry of mixtures of prescribed distributions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2861–2865. IEEE, 2018.

[94] Frank Nielsen and Richard Nock. Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family. *arXiv preprint arXiv:2003.02469*, 2020.

[95] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

[96] Frank Nielsen and Ke Sun. Guaranteed bounds on the Kullback–Leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 23(11):1543–1546, 2016.

[97] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. On Hölder projective divergences. *Entropy*, 19(3):122, 2017.

[98] Richard Nock, Frank Nielsen, and Shun-ichi. Amari. On conformal divergences and their population minimizers. *IEEE TIT*, 62(1):527–538, 2016.

[99] Katsumi Nomizu, Nomizu Katsumi, and Takeshi Sasaki. *Affine differential geometry: Geometry of affine immersions*. Cambridge university press, 1994.

[100] Aleksandr Petrovich Norden. On pairs of conjugate parallel displacements in multidimensional spaces. *Doklady Akademii nauk SSSR*, 49(9):1345–1347, 1945. Kazan State University, Comptes rendus de l'Académie des sciences de l'URSS.

[101] Atsumi Ohara and Takashi Tsuchiya. An information geometric approach to polynomial-time interior-point algorithms: Complexity bound via curvature integral. *The Institute of Statistical Mathematics*, 1055, 2007. Research Memorandum.

[102] María del Carmen Pardo and Igor Vajda. About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE transactions on information theory*, 43(4):1288–1293, 1997.

[103] Charles Sanders Peirce. *Chance, love, and logic: Philosophical essays*. U of Nebraska Press, 1998.

[104] Linyu Peng, Huafei Sun, and Lin Jiu. The geometric structure of the Pareto distribution. *Boletın de la Asociación Matemática Venezolana*, 14(1-2):5–13, 2007.

[105] Tongzhu Li Linyu Peng and Huafei Sun. The geometric structure of the inverse gamma distribution. *Contributions to Algebra and Geometry*, 49(1):217–225, 2008.

[106] Gia-Thuy Pham, Rémy Boyer, and Frank Nielsen. Computational information geometry for binary classification of high-dimensional random tensors. *Entropy*, 20(3):203, 2018.

[107] Giovanni Pistone. Nonparametric information geometry. In *Geometric Science of Information*, pages 5–36. Springer, 2013.

[108] Yu Qiao and Nobuaki Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

[109] Hernando Quevedo. Geometrothermodynamics. *Journal of Mathematical Physics*, 48(1):013506, 2007.

[110] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[111] Radhakrishna C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[112] Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

[113] Ferran Reverter and Josep M. Oller. Computing the Rao distance for Gamma distributions. *Journal of computational and applied mathematics*, 157(1):155–167, 2003.

[114] Yoshiharu Sato, Kazuaki Sugawa, and Michiaki Kawaguchi. The geometrical structure of the parameter space of the two-dimensional normal distribution. *Reports on Mathematical Physics*, 16(1):111–119, 1979.

[115] Gerhard Schurz. Patterns of abduction. *Synthese*, 164(2):201–234, 2008.

[116] R. N. Sen. On parallelism in Riemannian space I. *Bull. Calcutta Math. Soc*, 36:102–107, 1944.

[117] R. N. Sen. On parallelism in Riemannian space II. *Bull. Calcutta Math. Soc*, 37:153–159, 1944.

[118] R. N. Sen. On parallelism in Riemannian space III. *Bull. Calcutta Math. Soc*, 38:161–167, 1946.

[119] Claude Elwood Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656, 1948.

[120] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

[121] Lene Theil Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian journal of statistics*, pages 211–223, 1984.

[122] Anuj Srivastava, Wei Wu, Sebastian Kurtek, Eric Klassen, and James Stephen Marron. Registration of Functional Data Using Fisher-Rao Metric. *ArXiv e-prints*, 03 2011.

[123] Stephen M. Stigler. The epic story of maximum likelihood. *Statistical Science*, pages 598–620, 2007.

[124] Ke Sun and Frank Nielsen. Relative Fisher information and natural gradient for learning large modular models. In *ICML*, pages 3289–3298, 2017.

[125] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[126] Matus Telgarsky and Sanjoy Dasgupta. Agglomerative Bregman clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1011–1018. Omnipress, 2012.

[127] Igor Vajda. On metric divergences of probability measures. *Kybernetika*, 45(6):885–900, 2009.

[128] Hông Vân Lê. Statistical manifolds are statistical models. *Journal of Geometry*, 84(1-2):83–93, 2006.

[129] Hông Vân Lê. The uniqueness of the Fisher metric as information metric. *Annals of the Institute of Statistical Mathematics*, 69(4):879–896, 2017.

[130] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[131] Abraham Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, pages 165–205, 1949.

[132] Abraham Wald. *Statistical decision functions*. Wiley, 1950.

[133] MI Wanas. Absolute parallelism geometry: Developments, applications and problems. *arXiv preprint gr-qc/0209050*, 2002.

[134] Shao-Wen Wei, Yu-Xiao Liu, and Robert B Mann. Ruppeiner geometry, phase transitions, and the microstructure of charged AdS black holes. *Physical Review D*, 100(12):124033, 2019.

[135] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

[136] Shintaro Yoshizawa and Kunio Tanabe. Dual differential geometry associated with kullback-leibler information on the gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.

[137] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pages 5852–5861, 2018.

[138] Jun Zhang. Divergence functions and geometric structures they induce on a manifold. In Frank Nielsen, editor, *Geometric Theory of Information*, pages 1–30. Springer, 2014.

[139] Jun Zhang. On monotone embedding in information geometry. *Entropy*, 17(7):4485–4499, 2015.

[140] Jun Zhang. Reference duality and representation duality in information geometry. *AIP Conference Proceedings*, 1641(1):130–146, 2015.

[141] Zhenning Zhang, Huafei Sun, and Fengwei Zhong. Information geometry of the power inverse Gaussian distribution. *Applied Sciences*, 9, 2007.

[142] Fengwei Zhong, Huafei Sun, and Zhenning Zhang. The geometry of the Dirichlet manifold. *Journal of the Korean Mathematical Society*, 45(3):859–870, 2008.