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

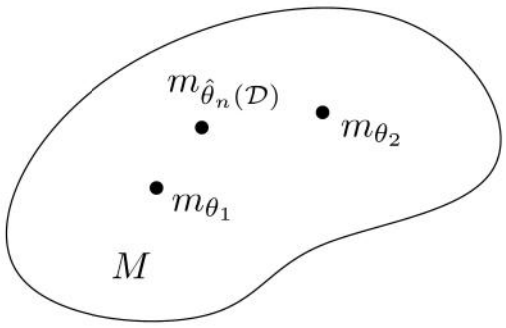

Figure 1: The parameter inference $\hat{\theta}$ of a model from data $\mathcal{D}$ can also be interpreted as a decision making problem: Decide which parameter of a parametric family of models $M = \{m_\theta\}_{\theta \in \Theta}$ suits the "best" the data. Information geometry provides a differential-geometric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

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 ∇.

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,

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 $g_p : T_p M \times T_p M \to \mathbb{R}$ is a 2-covariant tensor field:

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

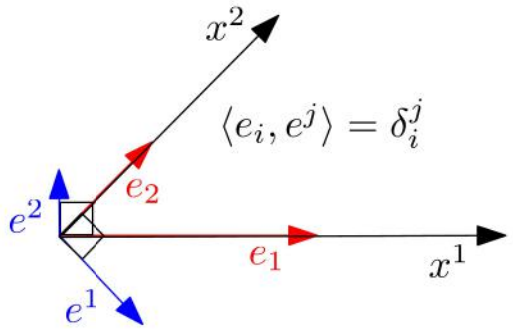

Figure 2: Primal basis (red) and reciprocal basis (blue) of an inner product $\langle \cdot, \cdot \rangle$ space. The primal/reciprocal basis are mutually orthogonal: e^1 is orthogonal to e_2, and e_1 is orthogonal to e^2.

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:

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

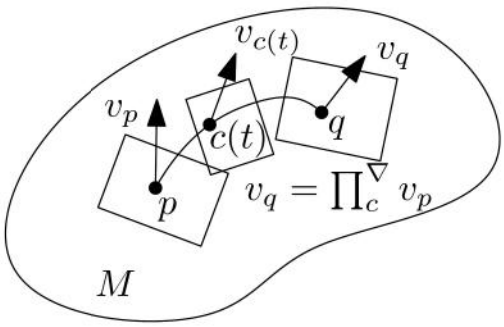

Figure 3: Illustration of the parallel transport of vectors on tangent planes along a smooth curve. For a smooth curve c, with $c(0) = p$ and $c(1) = q$, a vector $v_p \in T_p$ is parallel transported smoothly to a vector $v_q \in T_q$ such that for any $t \in [0, 1]$, we have $v_{c(t)} \in T_{c(t)}$.

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.

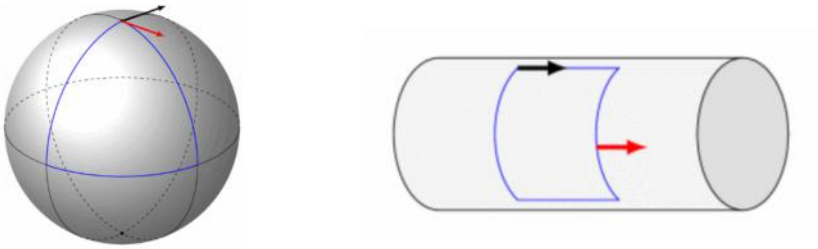

Figure 4: Parallel transport with respect to the metric connection: Curvature effect can be visualized as the angle defect along the parallel transport on smooth (infinitesimal) loops. For a sphere manifold, a vector parallel-transported along a loop does not coincide with itself, while it always conside 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.

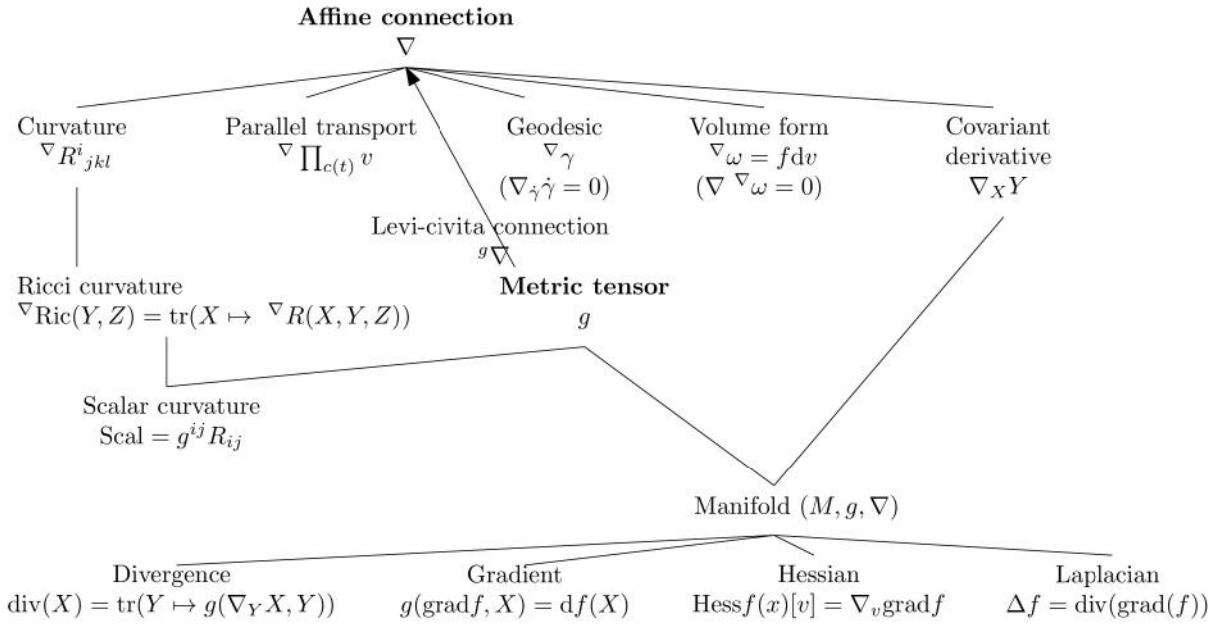

Figure 5: Differential-geometric concepts associated to an affine connection ∇ and a metric tensor g.

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 $E_{\cdot}[\cdot]$. 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*:

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

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):

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