Chin. Ann. Math.
27B(1), 2006, 73–94
DOI: 10.1007/s11401-005-0333-3

Chinese Annals of Mathematics, Series B

Riemann-Finsler Geometry with Applications to Information Geometry

Zhongmin SHEN*

(*Dedicated to the memory of Shiing-Shen Chern*)

Abstract Information geometry is a new branch in mathematics, originated from the applications of differential geometry to statistics. In this paper we briefly introduce Riemann-Finsler geometry, by which we establish Information Geometry on a much broader base, so that the potential applications of Information Geometry will be beyond statistics.

Keywords Riemann-Finsler geometry, Information geometry
2000 MR Subject Classification 53

1 Introduction

Information geometry has emerged from investigating the geometrical structure of a family of probability distributions, and has been applied successfully to various areas including statistical inference, control system theory and multi-terminal information theory (see [**?**, **?**]). The purpose of this paper is to give a brief introduction to Information Geometry from a more general point of view using Riemann-Finsler geometry and spray geometry.

Consider a set $\mathcal{F}$ of objects such as 2D/3D images, or probability distributions, etc. To measure the difference from one object to another in $\mathcal{F}$, one defines a function, $\mathcal{D}$, called a *divergence*, on the product space $\mathcal{F} \times \mathcal{F}$ with the following properties

$$\mathcal{D}(p, q) \geq 0, \quad \text{equality holds if and only if } p = q.$$

The number $\mathcal{D}(p, q)$ measures the "divergence" of p from q. The pair $(\mathcal{F}, \mathcal{D})$ is called a *divergence space*. To allow a great generality, the divergence $\mathcal{D}$ is not required to satisfy the reversibility condition: $\mathcal{D}(p, q) = \mathcal{D}(q, p)$.

For a divergence space $(\mathcal{F}, \mathcal{D})$, the set $\mathcal{F}$ is usually not finite-dimensional in any sense. In practice, one considers a family of objects in $\mathcal{F}$, parametrized in a domain of R^n. Such a family is called a model of $(\mathcal{F}, \mathcal{D})$. More precisely, a *model* of a divergence space $(\mathcal{F}, \mathcal{D})$ is an n-dimensional C^∞ manifold M as an embedded subset of $\mathcal{F}$ with the induced divergence $D = \mathcal{D}|_M$. Thus, a model (M, D) itself is also a divergence space.

Below are several examples.

Example 1.1 Let $(\mathcal{M}, d)$ be a metric space. Then $\mathcal{D} := \frac{1}{2}d^2$ is a divergence. This divergence is reversible, i.e., $\mathcal{D}(p, q) = \mathcal{D}(q, p)$.

Manuscript received August 12, 2005.
*Department of Mathematical Sciences, Indiana University Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202-3216, USA. E-mail: zshen@math.iupui.edu

Example 1.2 Let $\Omega \subset \mathrm{R}^n$ be an open subset and $\psi = \psi(x)$ be a C^∞ function on Ω with

$$\frac{\partial^2 \psi}{\partial x^i \partial x^j}(x) > 0.$$

Then

$$\psi(z) - \psi(x) - (z-x)^i \frac{\partial \psi}{\partial x^i}(x) \geq 0.$$

Define $D : \Omega \times \Omega \to [0, \infty)$ by

$$D(x,z) := \psi(z) - \psi(x) - (z-x)^i \frac{\partial \psi}{\partial x^i}(x). \tag{1}$$

D is a divergence on Ω.

More interesting examples are from other fields in natural science, such as mathematical psychology (see [5–7]).

Our goal is to use differential geometry to study *regular models* and the induced information structures. The regularity of divergence spaces and information structures will be defined in the following sections.

2 f-Divergences on Probability Distributions

An important class of divergence spaces comes from Probability Theory.

Let $\mathcal{X} = (\mathcal{X}, \mathcal{B}, \nu)$ be a measure space, where $\mathcal{X}$ is a set, $\mathcal{B}$ is a completely additive class consisting of $\mathcal{X}$ and its subsets, and ν is a σ-finite measure on $(\mathcal{X}, \mathcal{B})$. Let $\mathcal{P} = \mathcal{P}(\mathcal{X})$ be the space of probability distributions on $\mathcal{X}$.

$$\mathcal{P}(\mathcal{X}) := \Big\{ p : \mathcal{X} \to [0, \infty) \ \Big| \ \int_{\mathcal{X}} p(r) dr = 1 \Big\}.$$

The space $\mathcal{P}$ is convex in the sense that

$$\lambda p + (1-\lambda) q \in \mathcal{P}, \quad \text{if } p, q \in \mathcal{P}.$$

There is a special family of divergences on $\mathcal{P}$. Let $f : (0, \infty) \to \mathrm{R}$ be a convex function with

$$f(1) = 0, \quad f''(1) = 1. \tag{2}$$

Define $\mathcal{D}_f : \mathcal{P} \times \mathcal{P} \to \mathrm{R}$ by

$$\mathcal{D}_f(p,q) := \int_{\mathcal{X}} p(r) f\Big(\frac{q(r)}{p(r)}\Big) dr, \quad p = p(r), \ q = q(r) \in \mathcal{P}. \tag{3}$$

By Jensen's inequality, we have

$$\mathcal{D}_f(p,q) \geq f\Big(\int p(r) \frac{q(r)}{p(r)} dr \Big) = f(1) = 0,$$

where the equality holds if and only if $p = q$. Thus $\mathcal{D}_f$ is indeed a divergence on $\mathcal{P}$. We call $\mathcal{D}_f$ the *f-divergence* following I. Csiszàr. The f-divergence plays an important role in statistics.

There is a more special family of f-divergences on $\mathcal{P}$. For $\rho \in \mathrm{R}$, let

$$f_\rho(t) := \begin{cases} \frac{4}{1-\rho^2}\big(\frac{1+t}{2} - t^{(1+\rho)/2}\big) & \text{if } \rho \neq \pm 1, \\ t \ln t & \text{if } \rho = 1, \\ \ln(1/t) & \text{if } \rho = -1. \end{cases} \tag{4}$$

We have

$$f_\rho(1) = 0, \quad f'_\rho(1) = \frac{2}{\rho - 1}, \quad f''_\rho(1) = 1, \quad f'''_\rho(1) = \frac{\rho - 3}{2}.$$

For $\rho = 0$,

$$f_0(t) = 4\Big(\frac{1+t}{2} - \sqrt{t}\Big).$$

The divergence $\mathcal{D}_0$ on $\mathcal{P}$ is given by

$$\mathcal{D}_0(p,q) = 4\Big\{1 - \int \sqrt{p(r)q(r)}dr\Big\} = 2\int (\sqrt{p(r)} - \sqrt{q(r)}\,)^2 dr. \tag{5}$$

We see that $d_0(p,q) := \sqrt{2\mathcal{D}_0(p,q)}$ is a distance function. d_0 is called the *Hellinger distance* and $\mathcal{D}_0 = \frac{1}{2}d_0^2$ the *Hellinger divergence*.

For $\rho = -1$,

$$f_{-1}(t) = \ln(1/t).$$

The divergence $\mathcal{D}_{-1}$ on $\mathcal{P}$ is given by

$$\mathcal{D}_{-1}(p,q) = \int p(r) \ln \frac{p(r)}{q(r)} dr.$$

$\mathcal{D}_{-1}$ is called the *Kullback-Leibler divergence*.

3 Regular Divergences

Before we discuss regular divergences, let us first introduce Finsler metrics and H-functions.

Definition 3.1 *A Finsler metric on a manifold M is a scalar function $L = L(x,y)$ on TM with the following properties:*

(L1) $L(x,y) \geq 0$*, and the equality holds if and only if* $y = 0$;

(L2) $L(x, \lambda y) = \lambda^2 L(x,y)$, $\lambda > 0$;

(L3) $L(x,y)$ *is* C^∞ *on* $TM \setminus \{0\}$*, and for any* $y \in T_xM \setminus \{0\}$,

$$g_{ij}(x,y) := \frac{1}{2} L_{y^i y^j}(x,y) > 0. \tag{6}$$

For a Finsler metric L on a manifold M, the function $F_x := \sqrt{L}\,|_{T_xM}$ can be viewed as a norm on T_xM. Indeed, it satisfies the triangle inequality

$$F_x(u+v) \leq F_x(u) + F_x(v), \quad u, v \in T_xM.$$

But the reversibility $(F_x(-u) = F_x(u))$ is not assumed.

Let $g = g_{ij}(x)dx^i \otimes dx^j$ be a Riemannian metric as a tensor in the traditional notation. Then we get a scalar function L on TM:

$$L = g_{ij}(x)y^i y^j, \quad y = y^i \frac{\partial}{\partial x^i}\Big|_x.$$

By the above definition, L is a Finsler metric. Namely, Riemannian metrics are special Finsler metrics. Usually, we denote a Riemannian metric by the letter $g = g_{ij}(x)y^i y^j$. Riemannian metrics are the most important metrics and have been studied throughly in the last century.

Let (M, L) be a Finsler manifold. For a curve C parametrized by $c = c(t)$, $0 \leq t \leq 1$, the length of C is defined by

$$\mathcal{L}(C) = \int_0^1 \sqrt{L(c(t), c'(t))}\, dt.$$

Using the length structure, we can define a function $d = d(p, q)$ on $M \times M$ by

$$d(p, q) = \inf L(C),$$

where the infimum is taken over all curves from p to q. The distance function d satisfies

(a) $d(p, q) \geq 0$, and the equality holds if and only if $p = q$;

(b) $d(p, q) \leq d(p, r) + d(r, q)$.

d is called the *distance function* of L.

Definition 3.2 *An H-function on a manifold M is a scalar function $H = H(x, y)$ on TM with the following properties:*

(H1) *$H(x, \lambda y) = \lambda^3 H(x, y)$, $\lambda > 0$.*

(H2) *$H(x, y)$ is C^∞ on $TM \setminus \{0\}$.*

H-functions are positively homogeneous functions of degree three. There are lots of H-functions. If $L = L(x, y)$ is a Finsler metric on a manifold M, then the following function

$$H := L(x, y)^{3/2}$$

is an H-function on M. If $L = L(x, y)$ is a Finsler metric on an open subset $\Omega \subset \mathrm{R}^n$, then

$$H := \frac{1}{2} L_{x^k}(x, y) y^k$$

is an H-function on Ω.

Let $d = d(p, q)$ be the distance function of a Finsler metric L on M. Let

$$D(p, q) := \frac{1}{2} d(p, q)^2, \quad p, q \in M.$$

D is a divergence on M. In general, the divergence D is not C^∞ along the diagonal $\Delta = \{(p, p) \in M \times M\}$ unless L is Riemannian. Nevertheless we have the following

Lemma 3.3 *If D is the divergence of a Finsler metric L on a manifold M, then at any point p, there is a local coordinate system (U, ϕ) in M such that*

$$2D(\phi^{-1}(x), \phi^{-1}(x + y)) = L(x, y) + \frac{1}{2} L_{x^k}(x, y) y^k + o(|y|^3). \tag{7}$$

Now we are ready to define regular divergences.

Definition 3.4 *Let M be a manifold. A divergence function D on M is said to be regular if in any local coordinate system (U, ϕ) at any point in M (restricted to a smaller domain if necessary),*

$$2D(\phi^{-1}(x), \phi^{-1}(x + y)) = L(x, y) + P(x, y) + o(|y|^3), \tag{8}$$

where $L = L(x,y)$ is a Finsler metric on U and $P = P(x,y)$ is a C^∞ function on $TU \setminus \{0\}$ with

$$P(x, \lambda y) = \lambda^3 P(x,y), \quad \lambda > 0.$$

The Finsler metrics L in (**??**) form a global Finsler metric on M, while the functions P in (**??**) do not form a global scalar function on TM. However, one can use P to define an H-function on M.

Lemma 3.5 *Let D be a regular divergence on M. Let L and P be the local functions defined by (**??**) in a local coordinate system (U, ϕ). Then*

$$H := P(x,y) - \frac{1}{2} L_{x^k}(x,y) y^k \tag{9}$$

is a well-defined H-function on M.

Proof Let $\overline{L} = \overline{L}(\bar{x}, \bar{y})$ and $\overline{P} = \overline{P}(\bar{x}, \bar{y})$ be the local functions defined by (**??**) in another local coordinate system $(\overline{U}, \bar{\phi})$. Let $\bar{x} = \bar{\phi} \circ \phi^{-1}$.

$$\bar{x}(x+y) = \bar{x} + \bar{y} + \frac{1}{2} \frac{\partial^2 \bar{x}}{\partial x^i \partial x^j}(x) y^i y^j + o(|y|^2),$$

where

$$\bar{y} = \frac{\partial \bar{x}}{\partial x^i} y^i.$$

By comparing the expansions (**??**) in both coordinate systems, we get

$$L(x,y) = \overline{L}(\bar{x}, \bar{y}), \tag{10}$$

$$P(x,y) = \overline{P}(\bar{x}, \bar{y}) + \frac{1}{2} \overline{L}_{\bar{y}^k}(\bar{x}, \bar{y}) \frac{\partial^2 \bar{x}}{\partial x^i \partial x^j}(x) y^i y^j. \tag{11}$$

Differentiating (**??**) yields

$$\frac{1}{2} L_{x^k}(x,y) y^k = \frac{1}{2} \overline{L}_{\bar{x}^k}(\bar{x}, \bar{y}) \bar{y}^k + \frac{1}{2} \overline{L}_{\bar{y}^k}(\bar{x}, \bar{y}) \frac{\partial^2 \bar{x}}{\partial x^i \partial x^j}(x) y^i y^j.$$

Subtracting it from (**??**), we obtain

$$P(x,y) - \frac{1}{2} L_{x^k}(x,y) y^k = \overline{P}(\bar{x}, \bar{y}) - \frac{1}{2} \overline{L}_{\bar{x}^k}(\bar{x}, \bar{y}) \bar{y}^k.$$

Therefore the above function H is well-defined on M.

Now for a regular divergence D we have the following local expansion

$$2D(\phi^{-1}(x), \phi^{-1}(x+y)) = L(x,y) + \frac{1}{2} L_{x^k}(x,y) y^k + H(x,y) + o(|y|^3). \tag{12}$$

By Lemma 3.3, we have the following

Proposition 3.6 *If D is the divergence of a Finsler metric L on a manifold M, then it is regular with $H = 0$.*

Example 3.7 Let Ω be an open subset in a Minkowski space $(\mathrm{R}^n, \|\cdot\|)$ and $\psi(y) = a_{ijk}y^iy^jy^k$. Let

$$D(x,x') := \frac{1}{2}\|x'-x\|^2 + \frac{1}{2}\psi(x'-x), \quad x,x' \in \Omega.$$

Using the natural coordinate system $\varphi(x) = x$, we have

$$2D(x,x+y) = \|y\|^2 + \psi(y).$$

Thus D is a regular divergence with

$$L(x,y) = \|y\|^2, \quad H(x,y) = \psi(y).$$

4 Sprays of Finsler Metrics

Every Finsler metric L on a manifold M induces a vector field on TM,

$$\mathcal{G} := y^i \frac{\partial}{\partial x^i} - 2\mathcal{G}^i(x,y)\frac{\partial}{\partial y^i},$$

where

$$\mathcal{G}^i(x,y) := \frac{1}{4}g^{il}(x,y)\{L_{x^k y^l}(x,y)y^k - L_{x^l}(x,y)\}, \tag{13}$$

where $(g^{ij}(x,y)) := (g_{ij}(x,y))^{-1}$. From (**??**), one can see that

$$\mathcal{G}^i(x,\lambda y) = \lambda^2 \mathcal{G}^i(x,y), \quad \lambda > 0.$$

$\mathcal{G}$ is a well-defined C^∞ vector field on $TM \setminus \{0\}$. We call $\mathcal{G}$ the *spray* of L.

It is possible that two distinct Finsler metrics having the same spray. For example, if L is an arbitrary Finsler metric on a manifold, then the metric $\widetilde{L} := kL$ has the same spray as L for any positive constant k.

If $L = g_{ij}(x)y^iy^j$ is a Riemannian metric, then

$$\mathcal{G}^i(x,y) = \frac{1}{2}\gamma^i_{jk}(x)y^jy^k, \quad \gamma^i_{jk}(x) = \gamma^i_{kj}(x),$$

where

$$\gamma^i_{jk}(x) = \frac{1}{2}g^{il}(x)\Big\{\frac{\partial g_{jl}}{\partial x^k}(x) + \frac{\partial g_{kl}}{\partial x^j}(x) - \frac{\partial g_{jk}}{\partial x^l}(x)\Big\}. \tag{14}$$

The local functions $\gamma^i_{jk}(x)$ are called the *Christoffel symbols*. Note that $\mathcal{G}^i$ are quadratic in y.

A Finsler metric L is called a *Berwald metric* if its spray coefficients $\mathcal{G}^i = \frac{1}{2}\gamma^i_{jk}(x)y^jy^k$ are quadratic in y. There are many non-Riemannian Berwald metrics. An important fact is that every Berwald metric has the same spray as a Riemannian metric. This is due to Z. I. Szabo.

If $c = c(t)$ is an integral curve of $\mathcal{G}$ in $TM \setminus \{0\}$, then the local coordinates $(x(t), y(t))$ of $c(t)$ satisfy

$$\dot{x}^i \frac{\partial}{\partial x^i}\Big|_{c(t)} + \dot{y}^i(t)\frac{\partial}{\partial y^i}\Big|_{c(t)} = y^i(t)\frac{\partial}{\partial x^i}\Big|_{c(t)} - 2\mathcal{G}^i(x(t),y(t))\frac{\partial}{\partial y^i}\Big|_{c(t)}. \tag{15}$$

We obtain that $y^i(t) = \dot{x}^i(t)$ and

$$\ddot{x}^i(t) + 2\mathcal{G}^i(x(t),\dot{x}(t)) = 0. \tag{16}$$

Let $\sigma(t) := \pi(c(t))$ be the projection of $c = c(t)$ by $\pi : TM \to M$. The local coordinates of $\sigma(t)$ are $x(t) = (x^i(t))$, which satisfy (**??**). Conversely, if a curve $\sigma = \sigma(t)$ satisfies (**??**), then the canonical lift $c(t) = \dot{\sigma}(t)$ in TM is an integral curve of $\mathcal{G}$ such that $\sigma(t) = \pi(c(t))$.

Definition 4.1 *A curve σ in a Finsler manifold (M, L) is called a geodesic if its canonical lift $c := \dot{\sigma}$ in $TM \setminus \{0\}$ is an integral curve of the induced spray $\mathcal{G}$ by L.*

5 Sprays

The notion of sprays induced by a Finsler metric can be generalized.

Definition 5.1 *Let M be a manifold. A spray G on M is a vector field on the tangent bundle TM such that in any standard local coordinate system (x^i, y^i) in TM, it can be expressed in the following form*

$$G = y^i \frac{\partial}{\partial x^i} - 2G^i(x, y) \frac{\partial}{\partial y^i},$$

where $G^i(x, y)$ are C^∞ functions of (x^i, y^i) with $y \neq 0$ and

$$G^i(x, \lambda y) = \lambda^2 G^i(x, y), \quad \lambda > 0.$$

The notion of geodesics can also be extended to sprays. A curve $\sigma(t)$ is called a *geodesic* of $G := y^i \frac{\partial}{\partial x^i} - 2G^i(x, y) \frac{\partial}{\partial y^i}$ on a manifold M if it satisfies the following system of equations:

$$\ddot{x}^i(t) + 2G^i(x(t), \dot{x}(t)) = 0,$$

where $x(t) = (x^i(t))$ denotes the coordinates of $\sigma(t)$. Geodesics are also called *paths*. The collection of all paths of a spray is called a *path structure*.

A spray $G = y^i \frac{\partial}{\partial x^i} - 2G^i(x, y) \frac{\partial}{\partial y^i}$ is said to be *affine*, if in any local coordinate system,

$$G^i(x, y) = \frac{1}{2} \Gamma^i_{jk}(x) y^j y^k, \quad \Gamma^i_{jk}(x) = \Gamma^i_{kj}(x). \tag{17}$$

By definition, a Finsler metric is a Berwald metric if and only if its spray is affine.

Every affine spray G with coefficients $G^i(x, y) = \frac{1}{2} \Gamma^i_{jk}(x) y^j y^k$, $\Gamma^i_{jk}(x) = \Gamma^i_{kj}(x)$, defines a *connection* ∇ on TM,

$$\nabla_y X := \{dX^i(y) + X^j \Gamma^i_{jk}(x) y^k\} \frac{\partial}{\partial x^i}\Big|_x, \tag{18}$$

where $X = X^i \frac{\partial}{\partial x^i} \in C^\infty(TM)$ and $y = y^i \frac{\partial}{\partial x^i}\big|_x \in T_x M$. ∇ is *linear* in the following sense:

$$\nabla_{\lambda y + \mu v} X = \lambda \nabla_y X + \mu \nabla_v X,$$
$$\nabla_y (X + Y) = \nabla_y X + \nabla_y Y,$$
$$\nabla_y (fX) = df_x(y) X + f(x) \nabla_y X,$$

where $y, v \in T_x M$, $f \in C^\infty(M)$ and $X, Y \in C^\infty(TM)$. It is *torsion-free* in the following sense:

$$\nabla_X Y - \nabla_Y X = [X, Y],$$

where $X, Y \in C^\infty(TM)$. Torsion-free linear connections are also called *affine connections*.

Every affine spray defines an affine connection by (**??**). Conversely, every affine connection ∇ on TM defines a spray by (**??**). Thus affine connections one-to-one correspond to affine sprays.

$$\{\text{affine connections}\} \longleftrightarrow \{\text{affine sprays}\}.$$

Definition 5.2 *A spray G on a manifold is said to be flat if at every point, there is a standard local coordinate system (x^i, y^i) in TM such that $G = y^i \frac{\partial}{\partial x^i}$, i.e., $G^i = 0$. In this case, (x^i, y^i) is called an adapted coordinate system.*

Flat sprays are very special affine sprays. If G is flat, then in an adapted coordinate system, the geodesics of G are linear, i.e., the coordinates $(x^i(t))$ of every geodesic $\sigma(t)$ are in the following linear form

$$x^i(t) = a^i t + b^i.$$

6 Information Structures

By definition, any regular divergence D on a manifold M induces a Finsler metric L and an H-function. They can be obtained by the following formulas

$$L(x,y) = \lim_{\epsilon \to 0^+} \frac{2D(c(0), c(\epsilon))}{\epsilon^2}, \tag{19}$$

where $c(t)$ is an arbitrary C^1 curve in M with $c(0) = x$ and $c'(0) = y$;

$$H(x,y) = \lim_{\epsilon \to 0^+} \frac{2D(\sigma(0), \sigma(\epsilon)) - L(x,y)\epsilon^2}{\epsilon^3}, \tag{20}$$

where $\sigma = \sigma(t)$ is the geodesic with $\sigma(0) = x$ and $\dot{\sigma}(0) = y$.

Definition 6.1 *An information structure on a manifold M is a pair $\{L, H\}$, where $L = L(x,y)$ is a Finsler metric on M and $H = H(x,y)$ is a H-function.*

Every regular divergence induces an information structure. Conversely, every information structure is induced by a regular divergence as shown below.

Proposition 6.2 *Let (L, H) be an information structure on a manifold M. There is a regular divergence D on M such that the induced structure by D is $\{L, H\}$.*

Proof Let d denote the distance function of L on M. For $p, q \in M$, define

$$D(p,q) = \frac{1}{2} d(p,q)^2 + \inf_{c(0)=p, c(1)=q} \int_0^1 H(c(t), c'(t))dt,$$

where the infimum is taken over all minimizing geodesic c from p to q. Then it is easy to verify that D induces $\{L, H\}$.

7 The α-Sprays of an Information Structure

Let (L, H) be an information structure on a manifold M. Let $\mathcal{G} = y^i \frac{\partial}{\partial x^i} - 2\mathcal{G}^i \frac{\partial}{\partial y^i}$ be the spray of L. Using H, we can define a family of sprays $G_\alpha = y^i \frac{\partial}{\partial x^i} - 2G^i_\alpha(x,y) \frac{\partial}{\partial y^i}$ by

$$G^i_\alpha(x,y) := \mathcal{G}^i(x,y) + \frac{\alpha}{2} g^{ij}(x,y) H_{y^j}(x,y). \tag{21}$$

G_α is called the *α-spray* of (L, H). Our motivation to find a spray better than $\mathcal{G}$ so that the geodesics of the spray are simple. However, the rate of change of the divergence along any geodesic of the α-spray is not sensitive to α.

Lemma 7.1 *Let D be a regular divergence on a manifold M and (L, H) be the induced information structure and G_α be the α-spray of (L, H). Let $\sigma = \sigma(t)$ be a geodesic. Then for any geodesic σ of G_α,*

$$\frac{2D(\sigma(t_o), \sigma(t_o + \epsilon))}{d(\sigma(t_o), \sigma(t_o + \epsilon))^2} = 1 + \frac{H(x, y)}{3L(x, y)}\epsilon + o(\epsilon), \tag{22}$$

where $x = \sigma(t_o)$ and $y = \dot{\sigma}(t_o)$,

Proof Let $\phi = (x^i)$ be a local coordinate system in M. Let $x(t) := \phi(\sigma(t))$ and $\Delta x := x(t_o + \epsilon) - x(t_o)$. We have

$$\Delta x^i = \dot{x}^i(t_o)\epsilon + \frac{1}{2}\ddot{x}^i(t_o)\epsilon^2 + o(\epsilon^2) = y^i\epsilon - G^i_\alpha(x, y)\epsilon^2 + o(\epsilon^2).$$

By the above identity, we have

$$\begin{aligned}
L(x, \Delta x) &= L\epsilon^2 - L_{y^k}G^k_\alpha\epsilon^3 + o(\epsilon^3),\\
L_{x^k}(x, \Delta x)\Delta x^k &= L_{x^k}y^k\epsilon^3 + o(\epsilon^3),\\
H(x, \Delta x) &= H(x, y)\epsilon^3 + o(\epsilon^3).
\end{aligned}$$

It follows from (**??**) that

$$L_{y^k}\mathcal{G}^k = \frac{1}{2}L_{x^k}y^k. \tag{23}$$

Then by (**??**) we obtain

$$\begin{aligned}
2D(\sigma(t_o), \sigma(t_o + \epsilon)) &= 2D(\phi^{-1}(x), \phi^{-1}(x + \Delta x))\\
&= L(x, \Delta x) + \frac{1}{2}L_{x^k}(x, \Delta x)\Delta x^k + H(x, \Delta x) + o(\Delta x^3)\\
&= L\epsilon^2 - L_{y^k}G^k_\alpha\epsilon^3 + \frac{1}{2}L_{x^k}y^k\epsilon^3 + H\epsilon^3 + o(\epsilon^3)\\
&= L\epsilon^2 - L_{y^k}G^k_\alpha\epsilon^3 + L_{y^k}\mathcal{G}^k\epsilon^3 + H\epsilon^3 + o(\epsilon^3)\\
&= L\epsilon^2 + (1 - 3\alpha)H\epsilon^3 + o(\epsilon^3).
\end{aligned}$$

By a similar argument, we have

$$d(\sigma(t_o), \sigma(t_o + \epsilon))^2 = L\epsilon^2 - 3\alpha H\epsilon^3 + o(\epsilon^3).$$

Combining the above two expansions, we obtain (**??**).

Definition 7.2 *An information structure (L, H) on a manifold is said to be α-flat for some α if the α-spray G_α of (L, H) is flat. (L, H) is said to be flat if it is 1-flat.*

Let (L, H) be an information structure on M. Let

$$L^*(x, y) := L(x, -y), \quad H^*(x, y) := H(x, -y).$$

Then (L^*, H^*) is an information structure on M too. We call (L^*, H^*) the *dual information structure* of (L, H). The following lemma is trivial.

Lemma 7.3 *Let (L, H) be an information structure on a manifold M. Then*
(i) *(L, H) is α-flat if and only if $(L, \alpha H)$ is 1-flat.*
(ii) *(L, H) is α-flat if and only if the dual (L^*, H^*) is $(-\alpha)$-flat.*

Proof We only prove (ii). Let (L^*, H^*) be its dual structure of (L, H). Let G_α and G^*_α denote the α-sprays of (L, H) and (L^*, H^*), respectively. First we have

$$\mathcal{G}^{*i}(x, y) = \mathcal{G}^i(x, -y),$$
$$H^*_{y^j}(x, y) = -H_{y^j}(x, -y).$$

Thus

$$G^i_{-\alpha}(x, y) = G^i_\alpha(x, -y).$$

By this, it is easy to see that (L, H) is α-flat if and only if (L^*, H^*) is $(-\alpha)$-flat.

Lemma 7.4 *Let (L, H) be an information structure on a manifold M. For some $\alpha \neq 0$, (L, H) is α-flat if and only if at any point there is a local coordinate system (x^i) such that*

$$L_{x^k y^l} y^k = 2L_{x^l}, \tag{24}$$
$$\alpha H = -\frac{1}{6} L_{x^k} y^k. \tag{25}$$

Proof Suppose that (L, H) is α-flat. By assumption, there is a standard coordinate system (x^i, y^i) in which $G^i_\alpha(x, y) = 0$ hold. It follows from (**??**) and (**??**) that

$$H(x, y) = -\frac{1}{3\alpha} L_{y^k}(x, y)\mathcal{G}^k(x, y) = -\frac{1}{6\alpha} L_{x^k}(x, y) y^k.$$

Thus

$$\mathcal{G}^i(x, y) = -\frac{\alpha}{2} g^{il}(x, y) H_{y^l}(x, y) = \frac{1}{12} g^{il}(x, y)[L_{x^k}(x, y) y^k]_{y^l}.$$

Comparing it with (**??**), we obtain (**??**).

Conversely, if L satisfies (**??**), then the spray coefficients of L are given by

$$\mathcal{G}^i(x, y) = \frac{1}{4} g^{il}(x, y) L_{x^l}(x, y).$$

By (**??**) and (**??**), we have

$$\frac{\alpha}{2} g^{il}(x, y) H_{y^l}(x, y) = -\frac{1}{12} g^{il}(x, y)[L_{x^k}(x, y) y^k]_{y^l} = -\frac{1}{4} g^{il}(x, y) L_{x^l}(x, y).$$

Thus

$$G^i_\alpha(x, y) = \mathcal{G}^i(x, y) + \frac{\alpha}{2} g^{il}(x, y) H_{y^l}(x, y) = 0.$$

Thus the α-spray G_α is flat.

8 Dually Flat Finsler Metrics

In virtue of Lemma 7.4, we make the following

Definition 8.1 *A Finsler metric L on a manifold M is said to be locally dually flat if at any point, there is a local coordinate system (x^i) in which $L = L(x, y)$ satisfies (**??**), i.e.,*

$$L_{x^k y^l} y^k = 2L_{x^l}. \tag{26}$$

*Such a local system is called an adapted local system. L is said to be (globally) dually flat if there is an H-function H such that (L, H) is 1-flat, that is, at every point there is a local coordinate system (x^i) in which $L = L(x, y)$ satisfies (**??**) and the following equation*

$$L_{x^k} y^k = -6H. \tag{27}$$

If L is a locally dually flat Finsler metric on a manifold M, then at any point, there is a local coordinate system (x^i) in which the spray coefficients $\mathcal{G}^i$ of L satisfy

$$\mathcal{G}^i + \frac{1}{2} g^{ij} H_{y^j} = 0, \tag{28}$$

where $H := -\frac{1}{6} L_{x^k} y^k$.

Let us first consider locally dually flat Riemannian metrics.

Proposition 8.2 *A Riemannian metric $g = g_{ij}(x) y^i y^j$ on a manifold M is locally dually flat if and only if it can be locally expressed as*

$$g_{ij}(x) = \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x), \tag{29}$$

where $\psi = \psi(x)$ is a local scalar function on M.

Proof Assume that g is locally dually flat. There is a local coordinate system (x^i) in which $L := g$ satisfies (**??**).

$$\frac{\partial g_{il}}{\partial x^k}(x) + \frac{\partial g_{kl}}{\partial x^i}(x) = 2 \frac{\partial g_{ik}}{\partial x^l}(x). \tag{30}$$

Permutating i and l yields

$$\frac{\partial g_{il}}{\partial x^k}(x) + \frac{\partial g_{ik}}{\partial x^l}(x) = 2 \frac{\partial g_{kl}}{\partial x^i}(x). \tag{31}$$

Subtracting (**??**) from (**??**) yields

$$\frac{\partial g_{ik}}{\partial x^l}(x) = \frac{\partial g_{kl}}{\partial x^i}(x).$$

Thus there is a function $\psi(x)$ such that (**??**) holds. The converse is trivial.

Example 8.3 Let $\Omega \subset \mathrm{R}^n$ be a strongly convex domain defined by a Minkowski norm $\phi(y)$ on R^n,

$$\Omega := \{y \in \mathrm{R}^n \mid \phi(y) < 1\}.$$

Define $\Theta(x, y) > 0$, $y \neq 0$, by

$$\Theta(x, y) = \phi(y + \Theta(x, y)\, x), \quad y \in T_x\Omega = \mathrm{R}^n. \tag{32}$$

It is easy to verify that $\Theta(x,y)$ satisfies

$$\Theta_{x^k}(x,y)=\Theta(x,y)\Theta_{y^k}(x,y). \tag{33}$$

Let

$$L(x,y):=\Theta(x,y)^2.$$

Using (**??**), one obtains

$$\begin{aligned} L_{x^k} &= 2\Theta^2\Theta_{y^k}, \\ L_{x^k y^l}y^k &= [2\Theta^2\Theta_{y^k}]_{y^l}y^k = \frac{4}{3}[\Theta^3]_{y^l} = 4\Theta^2\Theta_{y^l}, \\ \frac{L_{x^k}y^k}{2L}L_{y^l} &= \frac{2\Theta^2}{2\Theta^2}\cdot 2\Theta\Theta_{y^l} = 2\Theta\Theta_{y^l}. \end{aligned}$$

Thus L satisfies (**??**). Namely, L is dually flat.

A Finsler metric L on an open domain $\mathcal{U}\subset \mathrm{R}^n$ is called a Funk metric, if $F:=\sqrt{L}$ satisfies

$$F_{x^k}=FF_{y^k}.$$

Every Funk metric is projectively flat, i.e., the geodesics are straight lines, or equivalently,

$$F_{x^k y^l}y^k = F_{x^l}. \tag{34}$$

A Finsler metric L is mutually dually flat and projectively flat if $F:=\sqrt{L}$ satisfies (**??**) and L satisfies (**??**). It can be shown that every mutually dually flat and projectively flat Finsler metric must be a Funk metric up to a scaling (see [**?**]).

9 Affine Divergences and Affine Information Structures

In general, a regular divergence $D: M\times M\to[0,\infty)$ is not C^∞ along the diagonal $\Delta:=\{(x,x),\ x\in M\}$.

Definition 9.1 *A regular divergence D on a manifold M is called an affine divergence if D is a C^∞ function on a neighborhood of the diagonal in $M\times M$.*

Lemma 9.2 *Let D be a regular affine divergence on a manifold M. Then the induced information structure (L,H) has the following properties:*

(i) $L=g_{ij}(x)y^iy^j$ *is Riemannian,*

(ii) $H=H_{ijk}(x)y^iy^jy^k$.

Proof Let

$$D(x,x'):=D(\phi^{-1}(x),\phi^{-1}(x')).$$

By assumption $D(x,x')$ is C^∞ in x,x'. Since $D(x,x)=0$, we have the following Taylor expansion

$$2D(x,x+y)=g_{ij}(x)y^iy^j+\frac{1}{3}h_{ijk}(x)y^iy^jy^k+o(|y|^3),$$

where

$$g_{ij}(x):=\frac{\partial^2 D}{\partial x'^i\partial x'^j}(x,x')\Big|_{x'=x},\quad h_{ijk}(x)=\frac{\partial^3 D}{\partial x'^i\partial x'^j\partial x'^k}(x,x')\Big|_{x'=x}.$$

Let

$$H_{ijk}(x) := \frac{1}{3}h_{ijk}(x) - \frac{1}{6}\Big\{\frac{\partial g_{ij}}{\partial x^k}(x) + \frac{\partial g_{ik}}{\partial x^j}(x) + \frac{\partial g_{jk}}{\partial x^i}(x)\Big\}.$$

Then

$$2D(x, x+y) = g_{ij}(x)y^iy^j + \frac{1}{2}\frac{\partial g_{ij}}{\partial x^k}(x)y^iy^jy^k + H_{ijk}(x)y^iy^jy^k + o(|y|^3).$$

Thus $L = g_{ij}(x)y^iy^j$ and $H = H_{ijk}(x)y^iy^jy^k$ are the induced metric and H-function.

Remark 9.3 For an affine divergence,

$$\frac{\partial^2 D}{\partial x^i \partial x^j}(x, x')\Big|_{x'=x} = \frac{\partial^2 D}{\partial x'^i \partial x'^j}(x, x')\Big|_{x'=x}.$$

Definition 9.4 *An information structure $\{L, H\}$ on a manifold M is said to be affine if*

(i) $L = g_{ij}(x)y^iy^j$ *is Riemannian, and*

(ii) $H = H_{ijk}(x)y^iy^jy^k$ *is a homogeneous polynomial.*

If $\{L, H\}$ is an affine information structure, then $(L^*, H^*) = (L, -H)$.

Lemma 9.5 *For an affine divergence D on a manifold M and its dual D^*, the induced information structure $\{L, H\}$ by D is dual to the induced information structure $\{L^*, H^*\}$ by D^*.*

Proof It suffices to prove that the induced information structure of D^* is $\{L, -H\}$.

10 α-Flat Affine Information Structures

We are particularly interested in α-flat information structures. If an information structure is α-flat, then the associated α-spray is flat.

In this section we are going to study flat affine information structures, and show that an affine information structure (L, H) is α-flat if and only if its dual (L^*, H^*) is α-flat.

Lemma 10.1 *Let (L, H) be an affine information structure on a manifold M and $\alpha \neq 0$. (L, H) is α-flat if and only if there is a local coordinate system (x^i) and a local function $\psi = \psi(x)$ such that*

$$L(x, y) = \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x)y^iy^j, \tag{35}$$

$$H(x, y) = -\frac{1}{6\alpha}\frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x)y^iy^jy^k. \tag{36}$$

Proof Assume that (L, H) is α-flat. By Lemma 7.4, there is a local coordinate system (x^i) such that

$$L_{x^k y^l}y^k = 2L_{x^l}.$$

Plugging $g_{ij}y^iy^j$ for L into the above equation, one can find a function $\psi(x)$ such that

$$g_{ij}(x) = \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x). \tag{37}$$

It follows from (**??**) that

$$H_{ijk}(x) = -\frac{1}{6\alpha}\frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x). \tag{38}$$

Conversely, if $L = g_{ij}(x)y^iy^j$ and $H = H_{ijk}(x)y^iy^jy^k$ are given by (**??**) and (**??**) respectively, then L satisfies (**??**) and H satisfies (**??**). Thus (L, H) is α-flat.

Lemma 10.2 *Let (L, H) be an affine information structure on a manifold M and $\alpha \neq 0$. Assume that in a local coordinate system (x^i), (L, H) is given by (**??**) and (**??**) respectively. Let $x^*_i := \frac{\partial\psi}{\partial x^i}(x)$ and*

$$\psi^*(x^*) := -\psi(x) + \sum_{i=1}^n x^*_i x^i. \tag{39}$$

Then in the new coordinate system (x^{*i}), the dual information structure $(L^*, H^*) = (L, -H)$ is given by

$$L^*(x^*, y^*) = \frac{\partial^2\psi^*}{\partial x^*_i \partial x^*_j}(x^*)y^*_i y^*_j, \tag{40}$$

$$H^*(x^*, y^*) = -\frac{1}{6\alpha}\frac{\partial^3\psi^*}{\partial x^*_i \partial x^*_j \partial x^*_k}(x^*)y^*_i y^*_j y^*_k. \tag{41}$$

Thus (L^*, H^*) is α-flat.

Proof First by (**??**), we have

$$\mathcal{G}^i = \frac{1}{4}g^{ik}(x)\frac{\partial^3\psi}{\partial x^i\partial x^j\partial x^k}(x)y^iy^j.$$

By definition,

$$g^*_{ij}(x) = g_{ij}(x), \quad H^*_{ijk}(x) = -H_{ijk}(x).$$

The α-spray G^*_α of (L^*, H^*) is given by

$$G^{*i}_\alpha(x, y) = \mathcal{G}^i(x, y) - \frac{\alpha}{2}g^{ik}H_{y^k}(x, y) = \frac{1}{2}g^{ik}(x)\frac{\partial^3\psi}{\partial x^i\partial x^j\partial x^k}(x)y^iy^j,$$

where $(g^{ij}(x)) := (g_{ij}(x))^{-1}$. That is, the Christoffel symbols $(\Gamma_\alpha)^{*i}_{jk}$ of G^*_α are given by

$$(\Gamma_\alpha)^{*i}_{jk}(x) = g^{il}(x)\frac{\partial^3\psi}{\partial x^j\partial x^k\partial x^l}(x).$$

Our goal is to find another local coordinate system (x^*_i) in which G^* is trivial. Consider the following map

$$x^*_i := \frac{\partial\psi}{\partial x^i}(x).$$

Since the Jacobian of $x^* = x^*(x)$ is just $(g_{ij}(x))$, this map is a local diffeomorphism which can serve as a coordinate transformation. Define ψ^* in (x^*_i) by (**??**). By a direct computation, we obtain

$$\frac{\partial\psi^*}{\partial x^*_k}(x^*) = x^k.$$

Since (L^*, H^*) is affine, we can express L^* and H^* in the new coordinate system (X^{*i}) by $L^* = g^{*kl}(x^*)y^*_k y^*_l$ and $H^* = H^{*ijk}(x^*)y^*_i y^*_j y^*_k$. It is easy to show that

$$g^{*kl}(x^*) = \frac{\partial^2\psi^*}{\partial x^*_k\partial x^*_l}(x^*),$$

and

$$\frac{\partial^2 x_i^*}{\partial x^j \partial x^k}(x) - \frac{\partial x_i^*}{\partial x^l}(x)(\Gamma_\alpha)^{*l}_{jk}(x) = 0.$$

Thus, in the local coordinate system (x_i^*), the spray coefficients of G_α^* vanish. This implies that

$$H^{*ijk}(x^*) = -\frac{1}{6\alpha}\frac{\partial^3 \psi^*}{\partial x_i^* \partial x_j^* \partial x_k^*}(x^*).$$

By the above lemmas, we get the following

Theorem 10.3 *Let $\alpha \neq 0$. An affine information structure (L, H) is α-flat if and only if its dual (L^*, H^*) is α-flat.*

11 Dualistic Affine Connections

We know that affine connections one-to-one correspond to affine sprays. An affine connection on a Riemannian manifold (M, g) is said to be *dualistic* if the dual linear connection ∇^* with respect to g is also affine. In this section we are going to characterize dualistic affine connections.

Let $L = g_{ij}y^iy^j$ be a Riemannian metric on a manifold M and $g = g_{ij}dx^i \otimes dx^j$ the associated inner product on tangent spaces. For a linear connection ∇ on M, define ∇^*:

$$g(\nabla^*_Z X, Y) + g(X, \nabla_Z Y) = Z[g(X, Y)], \tag{42}$$

where $X, Y, Z \in C^\infty(TM)$. It is easy to see that ∇^* is a linear connection too. ∇^* is called the *dual connection* of ∇ with respect to g. The concept of duality between two linear connections on a Riemannian manifold is introduced by S. -I. Amari and H. Nagaoka [**?**].

An important phenomenon is that if a linear connection ∇ is affine, the dual linear connection ∇^* (with respect to g) is not necessarily affine (i.e., it might not be torsion-free).

Theorem 11.1 *Let g be a Riemannian metric on a manifold M. Every polynomial H-function on (M, g) determines a dualistic affine connection. Conversely, every dualistic affine connection ∇ determines a polynomial H-function. The correspondence is canonical,*

$$\Gamma^i_{jk}(x) = \gamma^i_{jk}(x) + 3g^{il}H_{jkl}(x), \tag{43}$$

where Γ^i_{jk} denote the Christoffel symbols of ∇ and γ^i_{jk} denote the Christoffel symbols of g.

Proof Let H be a polynomial H-function on a Riemannian manifold (M, g). Let ∇ and $\overline{\nabla}$ be the affine connections corresponding to the associated 1-sprays G_1 and $\overline{G}_1$ of (g, H) and $(g, -H)$, respectively. Note that $(g, -H)$ is dual to (g, H). We claim that ∇ and $\overline{\nabla}$ satisfy

$$g(\overline{\nabla}_Z X, Y) + g(X, \nabla_Z Y) = Z[g(X, Y)], \tag{44}$$

namely, $\overline{\nabla}$ is dual to ∇ with respect to g.

Let $g = g_{ij}(x)y^iy^j$ and $H = H_{ijk}(x)y^iy^jy^k$. Let $\Gamma^i_{jk}(x)$ and $\overline{\Gamma}^i_{jk}(x)$ denote the Christoffel symbols of G_1 and $\overline{G}_1$ respectively. Let $\Gamma_{jk,i}(x) := g_{il}(x)\Gamma^l_{jk}(x)$, $\overline{\Gamma}_{jk,i}(x) := g_{il}(x)\overline{\Gamma}^l_{jk}(x)$, and etc. From (**??**), we have

$$\Gamma_{jk,i}(x) = \gamma_{jk,i}(x) + 3H_{ijk}(x), \tag{45}$$

$$\overline{\Gamma}_{ik,j}(x) = \gamma_{ik,j}(x) - 3H_{ijk}(x). \tag{46}$$

Adding (**??**) and (**??**) yields

$$\overline{\Gamma}_{ik,j}(x) + \Gamma_{jk,i}(x) = \gamma_{ik,j}(x) + \gamma_{jk,i}(x) = \frac{\partial g_{ij}}{\partial x^k}(x). \tag{47}$$

(**??**) can be written as (**??**). That is $\overline{\nabla} = \nabla^*$ is the dual linear connection of ∇ on (M, g). By definition, ∇ is dualistic.

Let ∇ be an affine connection on (M, g). Define H_{ijk} by (**??**). Clearly,

$$H_{ijk} = H_{ikj}.$$

Let ∇^* be the dual linear connection. Let Γ^{*i}_{jk} denote the Christoffel symbols of ∇^* and $\Gamma^*_{jk,l} = g_{il}\Gamma^{*i}_{jk}$. Then

$$\Gamma^*_{ik,j}(x) + \Gamma_{jk,i}(x) = \frac{\partial g_{ij}}{\partial x^k}(x) = \gamma_{ik,j}(x) + \gamma_{jk,i}(x). \tag{48}$$

It follows from (**??**) and (**??**) that

$$\Gamma^*_{ik,j}(x) = \gamma_{ik,j}(x) - 3H_{ijk}(x). \tag{49}$$

Suppose ∇^* is affine, i.e, $\Gamma^{*i}_{jk} = \Gamma^{*i}_{kj}$. Then

$$H_{ijk} = H_{kji}.$$

Thus H_{ijk} is symmetric in i, j, k. We obtain a polynomial H-function $H = H_{ijk}(x)y^iy^jy^k$. By (**??**), we see that H satisfies (**??**).

Since on a Riemannian manifold (M, g), dualistic affine connections one-to-one correspond to polynomial H-functions, we immediately obtain the following

Theorem 11.2 (See [**?**]) *Let ∇ and ∇^* be dual affine connections on a Riemannian manifold (M, g). Then ∇ is flat if and only if ∇^* is flat.*

Proof Let H be the polynomial H-function corresponding to ∇. Then $H^* := -H$ is the polynomial H-function corresponding to ∇^*. Note that the spray of (g, H) (resp. (g, H^*)) is the spray defined by ∇ (resp. ∇^*). Thus ∇ is flat if and only if (g, H) is 1-flat; (g, H) is 1-flat if and only if (g, H^*) is 1-flat by Theorem 10.3; (g, H^*) is 1-flat if and only if ∇^* is flat.

12 Statistical Models

Let $\mathcal{P}$ be a space of probability distributions on a measure space $\mathcal{X}$ and $\mathcal{D}$ a divergence on $\mathcal{P}$. A *statistical model* in $(\mathcal{P}, \mathcal{D})$ is a pair (M, D), where M is a finite C^∞ manifold embedded in $\mathcal{P}$ and D is the restriction of $\mathcal{D}$ on M. If f is a function satisfying (**??**), then it defines the f-divergence $\mathcal{D}_f$ on $\mathcal{P}$ by (**??**).

In this section, we are going to prove that for any manifold $M \subset \mathcal{P}$, the induced divergence $D_f = \mathcal{D}_f|_M$ is affine, namely, the induced metric $L = g_{ij}(s)y^iy^j$ is Riemannian and the induced H-function $H = H_{ijk}(x)y^iy^jy^k$ is a polynomial.

Theorem 12.1 *Let $f = f(t)$ be a function with $f(1) = 0$ and $f''(1) = 1$. For any regular statistical model (M, D_f) of $(\mathcal{P}, \mathcal{D}_f)$, the induced information structure on M is given* *by* $(L_f, H_f) = (L, \rho N)$, *where* $\rho := 3 + 2f'''(1)$, *and*

$$L = \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 p\, dr, \tag{50}$$

$$N = \frac{1}{6} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p\, dr. \tag{51}$$

The α*-spray* $G_{\alpha,\rho}$ *of* D_f *is given by* $G^i_{\alpha,\rho} = \overline{G}^i + (\rho\alpha + 1)A^i$, *where*

$$\overline{G}^i = \frac{g^{il}(x)}{2} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] \frac{\partial}{\partial x^l} p\, dr, \tag{52}$$

$$A^i = \frac{g^{il}(x)}{4} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 \frac{\partial}{\partial x^l} p\, dr. \tag{53}$$

Proof The natural embedding $M \to \mathcal{P}$ is given by $x \to p = p(r;x)$. Let $D(x,z) := D_f(p(r;x), p(r;z))$, i.e.,

$$D(x,z) := \int_{\mathcal{X}} p(r;x) f\Big(\frac{p(r;z)}{p(r;x)}\Big) dr.$$

We have

$$2D(x, x+y) = \frac{\partial^2 D}{\partial z^i \partial z^j}\Big|_{z=x} y^i y^j + \frac{1}{3} \frac{\partial^3 D}{\partial z^i \partial z^j \partial z^k}\Big|_{z=x} y^i y^j y^k + o(|y|^3).$$

By a direct computation, we obtain

$$\begin{aligned}
D|_{z=x} &= 0, \\
\frac{\partial D}{\partial z^i}\Big|_{z=x} y^i &= 0, \\
\frac{\partial^2 D}{\partial z^i \partial z^j}\Big|_{z=x} y^i y^j &= \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 p\, dr, \\
\frac{\partial^3 D}{\partial z^i \partial z^j \partial z^k}\Big|_{z=x} y^i y^j y^k &= \frac{\rho}{2} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p\, dr + \frac{3}{2} \Big\{ - \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p \\
&\quad + 2 \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} p \Big] \Big[y^k \frac{\partial}{\partial x^k} \ln p \Big] \Big\} dr.
\end{aligned}$$

Let

$$L := \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 p\, dr.$$

Then

$$\begin{aligned}
L_{x^k} y^k &= \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p\, dr + 2 \int_{\mathcal{X}} \Big[y^k \frac{\partial}{\partial x^k} \ln p \Big] \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr \\
&= - \int_{\mathcal{X}} \Big\{ \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p + 2 \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} p \Big] \Big[y^k \frac{\partial}{\partial x^k} \ln p \Big] \Big\} dr.
\end{aligned}$$

Let

$$N := \frac{1}{6} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^3 p\, dr.$$

We obtain

$$2D(x, x+y) = L(x,y) + \frac{1}{2} L_{x^k}(x,y) y^k + \rho N(x,y) + o(|y|^3).$$

Thus D_f is regular and the induced information structure $(L_f, H_f) = (L, \rho N)$ is affine.

Let $\mathcal{G} = y^i \frac{\partial}{\partial x^i} - 2\mathcal{G}^i \frac{\partial}{\partial y^i}$ denote the induced spray of L and $G_{\alpha,f} = y^i \frac{\partial}{\partial x^i} - 2G^i_{\alpha,\rho} \frac{\partial}{\partial y^i}$ be the α-spray of D_f. Without much difficulty, we obtain

$$\begin{aligned} G^i_{\alpha,\rho} &= \mathcal{G}^i(x,y) + \frac{\rho\alpha}{2} g^{il}(x) N_{y^l}(x,y) \\ &= (\rho\alpha + 1) \frac{g^{il}(x)}{4} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 \frac{\partial}{\partial x^l} p(r;x)\, dr \\ &\quad + \frac{g^{il}(x)}{2} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p(r;x) \Big] \frac{\partial}{\partial x^l} p\, dr. \end{aligned}$$

This gives a formula for $G_{\alpha,\rho}$.

Now let us express L and N in a different form. Observe that

$$\begin{aligned} L &= \int_{\mathcal{X}} y^j \frac{\partial}{\partial x^j} \Big\{ \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big] p \Big\} dr - \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr \\ &= \int_{\mathcal{X}} y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} p\, dr - \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr \\ &= y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \int_{\mathcal{X}} p dr - \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr \\ &= - \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr. \end{aligned}$$

This gives

$$L = - \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr. \tag{54}$$

By a similar argument, we obtain

$$\begin{aligned} 6N &= y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big]^2 p\, dr - 2 \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] \Big[y^k \frac{\partial}{\partial x^k} p \Big] dr \\ &= y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i \frac{\partial}{\partial x^i} \ln p \Big] \Big[y^j \frac{\partial}{\partial x^j} p \Big] dr - 2 y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr \\ &\quad + 2 \int_{\mathcal{X}} \Big[y^i y^j y^k \frac{\partial^3}{\partial x^i \partial x^j \partial x^k} \ln p \Big] p\, dr \\ &= -3 y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr + 2 \int_{\mathcal{X}} \Big[y^i y^j y^k \frac{\partial^3}{\partial x^i \partial x^j \partial x^k} \ln p \Big] p\, dr. \end{aligned}$$

This gives

$$N = \frac{1}{3} \int_{\mathcal{X}} \Big[y^i y^j y^k \frac{\partial^3}{\partial x^i \partial x^j \partial x^k} \ln p \Big] p\, dr - \frac{1}{2} y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2}{\partial x^i \partial x^j} \ln p \Big] p\, dr. \tag{55}$$

13 Exponential Family of Distributions

In this section, we will consider the exponential family of probability distributions, on which the α-spray of D_f with $\rho\alpha = -1$ is flat.

Definition 13.1 *A manifold M in $\mathcal{P}$ is called an exponential manifold if it is covered by injections*

$$\varpi : \Omega \subset \mathrm{R}^n \to M,$$

such that $p := \varpi(x) \in \mathcal{P}$ is in the following form

$$p(r;x) = \exp[x^i f_i(r) + k(r) - \psi(x)], \quad r \in \mathcal{X}. \tag{56}$$

Observe that the integral

$$\int_{\mathcal{X}} \frac{\partial p}{\partial x^i} dr = 0.$$

This implies that

$$\frac{\partial \psi}{\partial x^i}(x) = \int_{\mathcal{X}} p(r;x) f_i(r) dr.$$

The Kullback-Leibler divergence D_{KL} on M is the f-divergence with $f(t) = \ln(1/t)$. We have

$$\begin{aligned} D_{KL}(p(r;x), p(r;x')) &= \int p(r;x)[\psi(x') - \psi(x) - (x'-x)^i f_i(r)] dr \\ &= \psi(x') - \psi(x) - (x'-x)^i \frac{\partial \psi}{\partial x^i}(x). \end{aligned}$$

The pull-back of D_{KL} onto Ω is given by

$$D_{KL}(x,x') = \psi(x') - \psi(x) - (x'-x)^i \frac{\partial \psi}{\partial x^i}(x).$$

Proposition 13.2 *Let M be the exponential family of distributions in the form (**??**). The induced information structure of D_f is given by $(L_f, H_f) = (L, \rho N)$, $\rho = 3 + 2f'''(1)$, and*

$$L = \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x) y^i y^j, \quad N = \frac{1}{6} \frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x) y^i y^j y^k.$$

Proof Note that

$$\ln p(r;x) = x^i f_i(r) + k(r) - \psi(x).$$

It follows from (**??**) that

$$L(x,y) = \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x) \Big] p(r;x) dr = y^i y^j \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x).$$

Then the spray coefficients of L are given by

$$\mathcal{G}^i = \frac{1}{4} g^{ik} \frac{\partial^2 \psi}{\partial x^i \partial x^j \partial x^k}(x) y^i y^j.$$

It follows from (**??**) that

$$\begin{aligned} N(x,y) = &-\frac{1}{3} \int_{\mathcal{X}} \Big[y^i y^j y^k \frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x) \Big] p(r;x) dr + \frac{1}{2} y^k \frac{\partial}{\partial x^k} \int_{\mathcal{X}} \Big[y^i y^j \frac{\partial^2 \psi}{\partial x^i \partial x^j}(x) \Big] p(r;x) dr \\ &= \frac{1}{6} y^i y^j y^k \frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x). \end{aligned}$$

By Lemma 10.1, we obtain the following

Corollary 13.3 *Let M be the exponential family of distributions in the form (**??**). Let (L_f, H_f) be the information structure induced by the f-divergence. When $\rho\alpha = -1$, (L_f, H_f) is α-flat, namely, the α-spray of (L_f, H_f) is flat.*

Proof The α-spray is given by

$$G^i_{\alpha,\rho} = \mathcal{G}^i + \frac{\rho\alpha}{2} g^{ik} N_{y^k} = \frac{\rho\alpha+1}{4} g^{ik} \frac{\partial^3 \psi}{\partial x^i \partial x^j \partial x^k}(x) y^i y^j.$$

If $\rho\alpha = -1$, then the induced information structure (L_f, H_f) is α-flat.

Example 13.4 Consider the family M of Gaussian probability distributions with mean μ and variance σ^2:

$$p(r; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{(r-\mu)^2}{\sigma^2}\Big].$$

We can reparametrize M by

$$p(r; x) = \exp[x^1 f_1(r) + x^2 f_2(r) - \psi(x)],$$

where

$$x^1 = \frac{\mu}{\sigma^2}, \quad x^2 = \frac{1}{2\sigma^2}$$

and

$$f_1(r) = r, \quad f_2(r) = -r^2, \quad \psi(x) = \frac{\mu^2}{\sigma^2} + \ln(\sqrt{2\pi}\sigma) = \frac{(x^1)^2}{4x^2} + \ln\sqrt{\frac{\pi}{x^2}}.$$

Thus M is an exponential manifold in $\mathcal{P}$. The induced Riemannian metric $L = g_{ij}(x) y^i y^j$ of an f-divergence on M is given by

$$g_{11} = \frac{\partial^2\psi}{\partial x^1 \partial x^1}, \quad g_{12} = \frac{\partial^2\psi}{\partial x^1 \partial x^2}, \quad g_{22} = \frac{\partial^2\psi}{\partial x^2 \partial x^2}.$$

The Gauss curvature of L is a negative constant $K = -\frac{1}{2}$.

Example 13.5 Let M be the family of gamma distributions with event space $\Omega = \mathrm{R}^+$ and parameters $\tau, \nu \in \mathrm{R}^+$ which are defined by

$$p(r; \tau, \nu) = \Big(\frac{\nu}{\tau}\Big)^{\nu} \frac{r^{\nu-1}}{\Gamma(\nu)} \exp\Big[-\frac{r\nu}{\tau}\Big], \tag{57}$$

where Γ is the gamma function defined by

$$\Gamma(\nu) = \int_0^\infty s^{\nu-1} e^{-s} ds.$$

Note that $\tau = \langle r \rangle$ is the mean and $\tau^2/\nu = \mathrm{Var}(r)$ is the variance. Thus the coefficient of variation $\sqrt{\mathrm{Var}(r)}/\tau = 1/\sqrt{\nu}$ is independent of the mean.

Let $\mu := \nu/\tau$. Then gamma distributions can be expressed by

$$p(r; \mu, \nu) = \exp[-\mu r + \nu \ln r - \ln r - \psi(\mu, \nu)], \tag{58}$$

where

$$\psi(\mu,\nu) := \ln \Gamma(\nu) - \nu \ln \mu.$$

Thus M is an exponential manifold in $\mathcal{P}$. See [**?**] for related discussion.

Let L be the induced Riemannian metric by any f-divergence. In the coordinate system (τ,ν),

$$g_{11} = \frac{\nu}{\tau^2}, \quad g_{12} = 0 = g_{21}, \quad g_{22} = \Psi'(\nu) - \frac{1}{\nu},$$

where $\Psi(\nu) := \Gamma'(\nu)/\Gamma(\nu)$ is the logarithmic derivative of the gamma function. Since $\Psi(\nu)$ satisfies

$$\frac{1}{2\nu^2} \leq \Psi'(\nu) - \frac{1}{\nu} \leq \frac{1}{\nu^2}.$$

We have

$$L_1 := \frac{\nu}{\tau^2}\, u^2 + \frac{1}{2\nu^2}\, v^2 < L < \frac{\nu}{\tau^2}\, u^2 + \frac{1}{2\nu^2}\, v^2 := L_2.$$

The Gauss curvature K_i of L_i and the Gauss curvature K of L are given

$$K_1 = -\frac{1}{2} < K = \frac{\Psi'(\nu) + \Psi''(\nu)\nu}{4\nu^2(\Psi'(\nu) - 1/\nu)^2} < -\frac{1}{4} = K_2.$$

The reader is referred to [**?**] for the geometry of Gamma distributions and its applications.

14 Duality of f-Divergences

Let $(\mathcal{P},\mathcal{D})$ be a divergence space $(\mathcal{P},\mathcal{D})$. By definition, the dual divergence D^* is defined by

$$\mathcal{D}^*(p,q) := \mathcal{D}(q,p), \quad p,q \in \mathcal{P}.$$

Given a convex function $f : (0,\infty) \to \mathrm{R}$ with $f(1) = 0$ and $f''(1) = 1$. Let

$$f^*(t) := t f\Big(\frac{1}{t}\Big), \quad t > 0.$$

Then $f^*(t)$ satisfies that $f^*(1) = 0$ and $f^{*''}(1) = f''(1) = 1$. Let $\rho := 3 + 2f'''(1)$ and $\rho^* := 3 + 2f^{*'''}(1)$. We have

$$\rho + \rho^* = 0.$$

Note that

$$(D_f)^*(p,q) := D_f(q,p) = D_{f^*}(p,q).$$

Thus D_{f^*} is dual to D_f. By the above argument, $(D_f)^* = D_{f^*}$ induces an information structure

$$(L_{f^*}, H_{f^*}) = (L, \rho^* N) = (L, -\rho N).$$

That is, $L_{f^*}(x,y) = L_f(x,-y)$ and $H_{f^*}(x,y) = H_f(x,-y)$. The information structure of $(D_f)^*$ is dual to that of D_f. In this sense, D_f is said to be *dualistic*.

According to Lemmas 10.1 and 10.2, we have the following

Proposition 14.1 *The information structure (L_f, H_f) is α-flat if and only if the dual structure $(L_{f^*}, H_{f^*}) = (L_f(x,-y), H_f(x,-y))$ is α-flat.*

Let f_ρ be the function defined in (**??**). Let $D_\rho := D_{f_\rho}$. It is easy to see that

$$(f_\rho)^*(t) = f_{-\rho}(t).$$

Thus

$$(D_\rho)^*(p,q) = D_\rho(q,p) = D_{-\rho}(p,q).$$

For $\rho \neq \pm 1$,

$$D_\rho(p,q) = \frac{4}{1-\rho^2}\Big\{1 - \int p(r)^{(1-\rho)/2} q(r)^{(1+\rho)/2} dr\Big\}; \tag{59}$$

for $\rho = \pm 1$,

$$D_{-1}(p,q) = D_{+1}(q,p) = \int p(r) \ln \frac{p(r)}{q(r)} dr. \tag{60}$$

References

[1] Amari, S.-I. and Nagaoka, H., Methods of Information Geometry, Oxford University Press and Amer. Math. Soc., 2000.

[2] Bao, D., Chern, S. S. and Shen, Z., An Introduction to Riemann-Finsler Geometry, Springer, 2000.

[3] Amari, S.-I., Differential Geometrical Methods in Statistics, Springer Lecture Notes in Statistics, **20**, Springer, 2002.

[4] Dodson, C. T. J. and Matsuzoe, H., An affine embedding of the gamma manifold, *Appl. Sci.* (*electronic*), **5**(1), 2003, 7–12.

[5] Dzhafarov, E. D. and Colonius, H., Fechnerian metrics in unidimensional and multidimensional stimulus, *Psychological Bulletin and Review*, **6**, 1999, 239–268.

[6] Dzhafarov, E. D. and Colonius, H., Fechnerian scaling, probability-distance hypothesis, and Thurstonian link, Technical Report #**45**, Purdue Mathematical Psychology Program.

[7] Dzhafarov, E. D. and Colonius, H., Fechnerian Metrics, Looking Back: The End of the 20th Century Psychophysics, P. R. Kileen & W. R. Uttal (eds.), Arizona University Press, Tempe, AZ, 111–116.

[8] Hwang, T.-Y. and Hu, C.-Y., On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation, *Annals Inst. Statist. Math.*, **51**, 1999, 749–753.

[9] Lauritzen, S. L., Statistical manifolds, Differential Geometry in Statistical Inferences, IMS Lecture Notes Monograph Series, **10**, Hayward California, 1987, 96–163.

[10] Murray, M. K. and Rice, J. W., Differential Geometry and Statistics, Chapman & Hall, London, 1995.

[11] Nagaoka, H. and Amari, S., Differential geometry of smooth families of probability distributions, Univ. Tokyo, Tokyo, Japan, METR, 82–7, 1982.

[12] Shen, Z., Differential Geometry of Spray and Finsler Spaces, Kluwer Academic Publishers, Dordrecht, 2001.

[13] Shen, Z. and Yildirim, G. C., A characterization of Funk metrics, preprint, 2005.