Japan. J. Math. 16, 1–48 (2021)
DOI: 10.1007/s11537-020-1920-5

JAPANESE JOURNAL OF MATHEMATICS
SINCE 1924

Information geometry*

Shun-ichi Amari

Received: 10 September 2019 / Revised: 30 June 2020, 6 October 2020,
13 October 2020 / Accepted: 20 October 2020
Published online: 2 January 2021

Communicated by: Toshiyuki Kobayashi

Abstract. Information geometry has emerged from the study of the invariant structure in families of probability distributions. This invariance uniquely determines a second-order symmetric tensor g and third-order symmetric tensor T in a manifold of probability distributions. A pair of these tensors (g, T) defines a Riemannian metric and a pair of affine connections which together preserve the metric. Information geometry involves studying a Riemannian manifold having a pair of dual affine connections. Such a structure also arises from an asymmetric divergence function and affine differential geometry. A dually flat Riemannian manifold is particularly useful for various applications, because a generalized Pythagorean theorem and projection theorem hold. The Wasserstein distance gives another important geometry on probability distributions, which is non-invariant but responsible for the metric properties of a sample space. I attempt to construct information geometry of the entropy-regularized Wasserstein distance.

Keywords and phrases: canonical divergence, dual affine connection, information geometry, Pythagorean theorem, semiparametric statistics, Wasserstein geometry

Mathematics Subject Classification (2020): 53B12

* This article is based on the 23rd Takagi Lectures that the author delivered at Research Institute for Mathematical Sciences, Kyoto University on June 8, 2019.

S. Amari
RIKEN, Hirosawa, Wako-shi, Saitama, 351-0198 Japan
(e-mail: amari@brain.riken.jp)

1. Introduction

Statistics involves the study of a parameterized family of probability distributions, which is a statistical model that forms a manifold where the parameters play the role of local coordinates. We search for a natural geometric structure to be introduced in such a statistical manifold. [RAO45] introduced a Riemannian structure by using the Fisher information matrix. It was [CHEN72] who proposed the criterion of invariance such that the structure should be invariant under Markov morphisms. We reformulate generally that the geometry should be invariant when sufficient statistics are used instead of the original random sample.

The invariance criterion determines two quantities, a second-order positive-definite symmetric tensor g and third-order symmetric tensor T. The former is the Fisher information matrix, playing the role of a Riemannian metric. The g together with T gives two invariant affine connections, which are dually coupled in the sense that, although each is non-metric, they together preserve the Riemannian metric ([AMN00], [AM16], [AM85]).

Information geometry involves the study of the geometry of a manifold equipped with a Riemannian metric g and symmetric cubic tensor T, or equivalently a Riemannian manifold equipped with a pair of dual affine connections. Such a structure also emerges from a manifold in which a divergence function is defined. A divergence function is an asymmetric function $D[p : q]$ of two points p and q in the manifold such that it is non-negative, equal to 0 when and only when $p = q$ and, when q is infinitesimally close to p, the Taylor expansion of D gives a positive quadratic form, playing the role of a Riemannian metric. Hence, a divergence is a generalization of the square of Riemannian distance in an asymmetric manner. We show that a manifold equipped with a divergence introduces a Riemannian structure with a pair of dual affine connections ([EG83]).

Information geometry is closely connected to affine differential geometry ([NOS94]), which involves the study of the structure of an n-dimensional manifold immersed in an $(n+1)$-dimensional affine space together with a transversal vector field attached to it. A dual pair of affine connections may emerge from affine differential geometry.

A Riemannian manifold having a pair of flat affine connections is particularly interesting. Such a manifold is generally not Euclidean, because the Levi-Civita connection has non-zero curvature. It has a unique canonical divergence, for which a generalized Pythagorean theorem holds together with mutually orthogonal primal and dual geodesics. A projection theorem also holds, which gives a useful tool in many applications. Interestingly, this gives a geometrical meaning to the well-known Legendre

transformation. We give an application to statistical inference by using the semi-parametric statistical model.

The Wasserstein distance is an interesting topic of research concerning the distance of probability distributions ([VIL09]). It is not invariant but depends on the distance in the sample space. There has been extensive research on this topic with various applications [SAN15], [PEC18], where the distance in the sample space plays an important role, such as visual patterns. Since the original Wasserstein problem is computationally difficult to solve, [CUT13] used the entropy-regularized Wasserstein distance, showing its effectiveness in various applications. We give a divergence function derived from the entropy-regularized Wasserstein problem and study its relation to information geometry ([AKO18], [AKOC19]). A more fundamental approach is found in a recent paper ([LIZ19]).

We study mostly statistical models specified by finite numbers of parameters. However, Sect. 5 and Subsect. 7.7 include models of function spaces. We admit these parts are intuitive and not mathematically rigorously formulated.

2. Riemannian manifold with dually coupled affine connections

2.1. Invariant geometry of manifold of probability distributions

We begin with a finite-dimensional regular statistical model, which is a parameterized family of probability distributions $p(x, \boldsymbol{\xi})$ over a sample space Ω, $x \in \Omega$. Here, $\boldsymbol{\xi}$ is an n-dimensional vector in a parameter space $\boldsymbol{R}^n$, x is a random variable and $p(x, \boldsymbol{\xi})$ is a probability density of x with respect to measure $\mu(x)$ of Ω. (A random variable is conventionally denoted by capital X and its realization is by small case x. But we use only small case x for the both cases, hoping no confusion occurs.) We assume that $p(x, \boldsymbol{\xi})$ is differentiable with respect to $\boldsymbol{\xi}$. The set of such distributions

$$M = \{p(x, \boldsymbol{\xi})\} \tag{1}$$

forms an n-dimensional manifold (see [AM16]; more rigorously [AY17]), where $\boldsymbol{\xi}$ is a coordinate system in a local chart.

We show two simple examples:

1) Exponential family

$$p(x, \boldsymbol{\xi}) = \exp\{\xi^i x_i - \psi(\boldsymbol{\xi})\}, \tag{2}$$

where

$$x = (x_i), \quad \boldsymbol{\xi} = (\xi^i), \quad i = 1, \ldots, n \tag{3}$$

and the summation convention is used in the form $\xi^i x_i$, implying that $\xi^i x_i = \sum \xi^i x_i$. The function $\psi(\boldsymbol{\xi})$ is derived from the normalization condition

$$\int p(x, \boldsymbol{\xi})\, d\mu(x) = 1 \tag{4}$$

and given explicitly as

$$\psi(\boldsymbol{\xi}) = \log \int \exp\{\xi^i x_i\}\, d\mu(x). \tag{5}$$

This is the logarithm of the Laplace transform of $\mu(x)$. An exponential family is said to be regular and minimally represented, when $\psi(\boldsymbol{\xi})$ is differentiable and its Hessian,

$$g_{ij} = \frac{\partial^2}{\partial \xi^i \partial \xi^j} \psi(\boldsymbol{\xi}), \tag{6}$$

is positive-definite.

There are many exponential families depending on $\mu(x)$. One simple example is a family of Gaussian distributions of random variable z, which can be written as

$$p(z, m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(z-m)^2}{2\sigma^2} \right\} \tag{7}$$

in terms of mean and variance parameters (m, σ^2). When we introduce vector $\boldsymbol{x}$,

$$\boldsymbol{x} = (x_1, x_2), \quad x_1 = z, \quad x_2 = z^2, \tag{8}$$

and define

$$\boldsymbol{\xi} = (\xi^1, \xi^2), \quad \xi^1 = \frac{m}{\sigma^2}, \quad \xi^2 = -\frac{1}{2\sigma^2}, \tag{9}$$

it is rewritten in the standard form (2) of the exponential family since

$$\boldsymbol{\xi} \cdot \boldsymbol{x} = \xi^i x_i = -\frac{(z-m)^2}{2\sigma^2} + \frac{m^2}{2\sigma^2}. \tag{10}$$

The $\mu(z)$ is the Lebesgue measure on $\boldsymbol{R}^1$ and $\mu(\boldsymbol{x})$ is defined on $x_1^2 - x_2 = 0$. Multivariate Gaussian distributions form another exponential family. The class of exponential families covers many well-known families of probability distributions.

2) Discrete distributions

When Ω consists of $(n+1)$ points, a probability over Ω is represented by an $(n+1)$-dimensional vector

$$\boldsymbol{p} = (p_0, p_1, \ldots, p_n), \tag{11}$$

satisfying

$$\sum p_i = 1, \quad p_i > 0, \tag{12}$$

where p_i is the probability of $x = i$, $i = 0, 1, \ldots, n$. The family of probability distributions is called a probability simplex S_n. It is an exponential family since we have

$$p(x, \boldsymbol{\xi}) = \exp\{\xi^i x_i - \psi(\boldsymbol{\xi})\}, \tag{13}$$

where

$$\xi^i = \log \frac{p_i}{p_0}, \quad i = 1, \ldots, n, \tag{14}$$

$$x_i = \delta_i(x) = \begin{cases} 1, \text{ when } x = i, \\ 0, \text{ otherwise}, \end{cases} \tag{15}$$

$$\psi(\boldsymbol{\xi}) = -\log p_0 = \log\left(1 + \sum e^{\xi_i}\right), \tag{16}$$

with the trivial counting measure $\mu(\boldsymbol{x})$.

2.2. Invariance under sufficient statistic

We pose an invariance criterion for the purpose of introducing a differential-geometrical structure in a manifold of probability distributions. A statistic $s(x)$, a function of x, is said to be sufficient when the probability density is decomposed as

$$p(x, \boldsymbol{\xi}) = p(s, \boldsymbol{\xi}) p(x|s), \tag{17}$$

in which the conditional probability density $p(x|s)$ of x conditioned on s does not depend on $\boldsymbol{\xi}$. We used an abused notion of representing probabilities $p(s, \boldsymbol{\xi})$ and $p(x|s)$ by using the same letter p. Roughly speaking, only s part depends on $\boldsymbol{\xi}$, so s is sufficient for estimating parameter $\boldsymbol{\xi}$. When Ω is a real line $\boldsymbol{R}^1$, any invertible function of $s(x)$ is a sufficient statistic. We show a proposition posed by [AMN00] as the start of information geometry, which was originally due to [CHEN72] in the discrete case.

Invariance Criterion: Geometry is said to be invariant when the geometry of $M = \{p(x, \boldsymbol{\xi})\}$ is identical to that of $M' = \{p(s, \boldsymbol{\xi})\}$.

Proposition. *Manifold M of probability distributions has a unique invariant second-order symmetric tensor g and a third-order symmetric tensor T under the invariance criterion. They are given in the component form by*

$$g_{ij} = \mathrm{E}[\partial_i l(x, \boldsymbol{\xi}) \partial_j l(x, \boldsymbol{\xi})], \tag{18}$$

$$T_{ijk} = \mathrm{E}[\partial_i l(x, \boldsymbol{\xi}) \partial_j l(x, \boldsymbol{\xi}) \partial_k l(x, \boldsymbol{\xi})], \tag{19}$$

except for a common scale, where E *is the expectation with respect to* $p(x, \boldsymbol{\xi}), l$ *is log probability,*

$$l(x, \boldsymbol{\xi}) = \log p(x, \boldsymbol{\xi}) \tag{20}$$

and ∂_i *denotes differentiation*

$$\partial_i = \frac{\partial}{\partial \xi^i}. \tag{21}$$

It is easy to see that the tensors in (18) and (19) are invariant. The converse is not so easy. The proposition was originally proved by a Russian mathematician [CHEN72] in the discrete case S_n. There are many papers for justifying this proposition in the function space, see [AY17] and [BAU16]. A recent paper by [DOW18] proved it for exponential families and curved exponential families.

To show the implications of the invariance criterion, we give a simple example. Let $S_1 = \{p(x, \xi)\}$ be given by probability distributions

$$p(x, \xi) = \xi \delta_1(x) + (1 - \xi)\delta_0(x), \tag{22}$$

where $x = 0, 1, \delta_i(x)$ is the Kronecker delta, and $\xi = \mathrm{Prob}\{x = 1\} \in (0, 1)$. The manifold is an interval $S_1 = (0, 1)$. We next consider another family S_2 of probability distributions,

$$p(x, \boldsymbol{\xi}) = \xi^1 \delta_1(x) + \xi^2 \delta_2(x) + (1 - \xi^1 - \xi^2)\delta_0(x), \tag{23}$$

where $x = 0, 1, 2$, and $\boldsymbol{\xi} = (\xi^1, \xi^2)$, $\xi^1, \xi^2 > 0$, $\xi^1 + \xi^2 < 1$. When we introduce $\xi^0 = 1 - \xi^1 - \xi^2$, S_2 is represented by a triangle satisfying

$$\xi^0 + \xi^1 + \xi^2 = 1, \quad \xi^0, \xi^1, \xi^2 > 0 \tag{24}$$

in $\boldsymbol{R}^3$ and is called the probability simplex S_2. We consider the following probability model $\tilde{S}_1$ parameterized by ξ:

$$\mathrm{Prob}\{x = 0\} = 1 - \xi, \tag{25}$$

$$\mathrm{Prob}\{x = 1\} = r\xi, \tag{26}$$

$$\mathrm{Prob}\{x = 2\} = (1 - r)\xi, \tag{27}$$

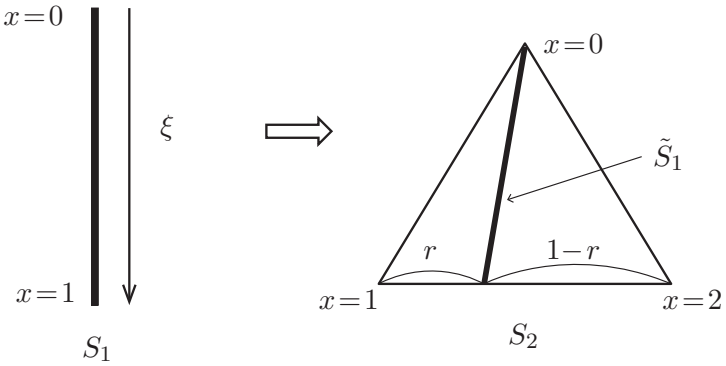

Fig. 1. Embedding of S_1 in S_2

where $0 < r < 1$ is a known constant. The model is one-dimensional with $\text{Prob}\{x = 0\} = 1 - \xi$ and when $x \neq 0$, $\text{Prob}\{x = 1 \mid x \neq 0\} = r$ and $\text{Prob}\{x = 2 \mid x \neq 0\} = 1 - r$. Hence, this model is specified by

$$\tilde{p}(x, \xi) = \xi r \delta_1(x) + \xi(1 - r)\delta_2(x) + (1 - \xi)\delta_0(x). \tag{28}$$

So it is a submanifold of S_2.

Let us put

$$s(x) = \delta_1(x) + \delta_2(x), \tag{29}$$

where $s(x) = 0$ when $x = 0$, and 1 when $x \neq 0$. By introducing

$$\tilde{p}(s, \xi) = \xi \delta_1(s) + (1 - \xi)\delta_0(s), \tag{30}$$

we have

$$\tilde{p}(x, \xi) = \tilde{p}(s, \xi)\tilde{p}(x \mid s), \tag{31}$$

where

$$\tilde{p}(x = 1 \mid s = 1) = r, \quad \tilde{p}(x = 2 \mid s = 1) = 1 - r, \quad \tilde{p}(x = 0 \mid s = 0) = 1, \tag{32}$$

and 0 otherwise. Since $\tilde{p}(x \mid s)$ does not depend on ξ, s is a sufficient statistic.

Note that $\tilde{S}_1$ is a submanifold of S_2 (see Fig. 1) specified by

$$\xi^1 = r\xi, \quad \xi^2 = (1 - r)\xi. \tag{33}$$

The invariance criterion requires that the geometry of S_1 is the same as $\tilde{S}_1$ embedded in S_2 for any r. We may consider an embedding of S_n in S_m $(n < m)$ in a similar manner. The center of S_n,

$$\boldsymbol{\xi}_{\text{center}} = \frac{1}{n + 1}(1, \ldots, 1) \tag{34}$$

is isotropical because a permutation of $(0, 1, \ldots, n)$ gives the same probability model. [CHEN72] proved this Proposition from the fact that geometry of S_n is the same as that of submanifolds $\tilde{S}_n \subset S_m$ inherited from S_m. See also a book in Japanese by [FUJ15].

2.3. Affine connections derived from (g, T)

We have two invariant tensors g and T in a manifold of probability distributions. Here, g plays the role of a Riemannian metric. It is the well-known Fisher information matrix, playing a fundamental role in statistics. It gives how much information is included in an observed sample x for estimating parameter $\boldsymbol{\xi}$. The other invariant tensor T has not been well studied in statistics. We call T a cubic tensor.

Riemannian geometry involves the study of a manifold $\{M, g\}$ equipped with g. We study a manifold $\{M, g, T\}$ equipped with g and T, not necessarily derived from probability distributions. Since it is motivated from the invariance criterion of statistics, it is called a statistical manifold by [LAU87].

In the case of a Riemannian manifold, we have a unique torsion-free metric affine connection, the Levi-Civita connection, that satisfies

$$\overset{0}{\nabla}_i g_{jk} = 0, \tag{35}$$

where $\overset{0}{\nabla}_i$ is the covariant derivative in the direction of $\partial/\partial\xi^i$, and $\overset{0}{\nabla}_i$ is metric preserving.

In the case of a statistical manifold, a pair of torsion-free affine connections, or equivalently covariant derivatives ∇ and ∇^*, are derived from g and T, which satisfy

$$\nabla_i g_{jk} = T_{ijk}, \tag{36}$$

$$\nabla^*_i g_{jk} = -T_{ijk}. \tag{37}$$

These covariant derivatives are not metric preserving but the pair (∇, ∇^*) is metric preserving in the dual sense, as described in the following subsection. The pair is called dually coupled affine connections. They are given in terms of the components of affine connections as

$$\Gamma^k_{ij} = \left\{ \begin{matrix} k \\ ij \end{matrix} \right\} - \frac{1}{2} T^k_{ij}, \tag{38}$$

$$\Gamma^{*k}_{ij} = \left\{ \begin{matrix} k \\ ij \end{matrix} \right\} + \frac{1}{2} T^k_{ij}, \tag{39}$$

where $\left\{ \begin{matrix} k \\ ij \end{matrix} \right\}$ is the Christoffel symbol showing the Levi-Civita connection,

$$\left\{ \begin{matrix} k \\ ij \end{matrix} \right\} = g^{km}[ij; m], \tag{40}$$

$$[ij; m] = \frac{1}{2}(\partial_i g_{jm} + \partial_j g_{im} - \partial_m g_{ij}), \tag{41}$$

where (g^{km}) is the inverse of (g_{mk}) and

$$T^k_{ij} = g^{km} T_{ijm}. \tag{42}$$

We may generalize these connections by using a real parameter α,

$$\overset{\alpha}{\Gamma}{}^k_{ij} = \left\{ \begin{matrix} k \\ ij \end{matrix} \right\} - \frac{\alpha}{2} T^k_{ij}, \tag{43}$$

$$\overset{-\alpha}{\Gamma}{}^k_{ij} = \left\{ \begin{matrix} k \\ ij \end{matrix} \right\} + \frac{\alpha}{2} T^k_{ij}, \tag{44}$$

forming a dually coupled pair of affine connections, which are called α- and $-\alpha$-connections, respectively. When $\alpha = 0$, it reduces to the Levi-Civita Riemannian connection. The $\pm\alpha$-connections define the α-geometry.

An affine connection gives covariant derivative $\nabla_Y X$ of vector field X in the direction of another vector field Y. It also gives a parallel transport of vector X in tangent space $T_{\boldsymbol{\xi}}$ at $\boldsymbol{\xi}$ to another tangent space $T_{\boldsymbol{\xi}'}$ at $\boldsymbol{\xi}'$ along a smooth path

$$c : \boldsymbol{\xi}(t), \tag{45}$$

where

$$\boldsymbol{\xi}(0) = \boldsymbol{\xi}, \tag{46}$$

$$\boldsymbol{\xi}(1) = \boldsymbol{\xi}', \tag{47}$$

connecting the two points $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$. Let $X(t)$ be a vector field along the curve c. When

$$\nabla_{\dot{\xi}(t)} X(t) = 0, \tag{48}$$

where

$$\dot{\boldsymbol{\xi}} = \frac{d}{dt} \boldsymbol{\xi}(t), \tag{49}$$

$X(t)$ is said to be a parallel field along c. The $X' = X(1)$ is the parallel transport of $X = X(0)$ at $\boldsymbol{\xi}$ to $\boldsymbol{\xi}'$ along c (Fig. 2). It is written as

$$X' = \prod\nolimits_{\boldsymbol{\xi},c}^{\boldsymbol{\xi}'} X. \tag{50}$$

A curve $\boldsymbol{\xi}(t)$ is a geodesic, when it satisfies

$$\nabla_{\dot{\boldsymbol{\xi}}(t)} \dot{\boldsymbol{\xi}}(t) = 0. \tag{51}$$

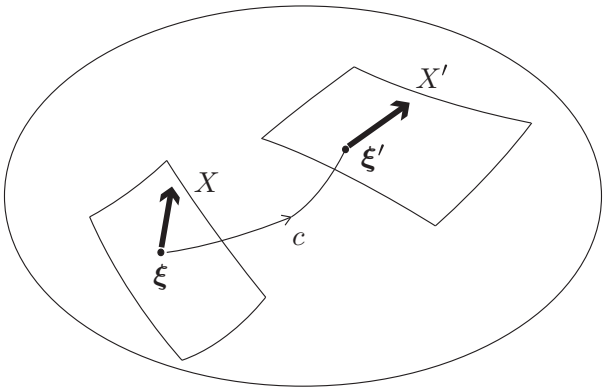

Fig. 2. Parallel transport of $\boldsymbol{X}$ at $\boldsymbol{\xi}$ to $\boldsymbol{X}'$ at $\boldsymbol{\xi}'$

2.4. Dual connections and metric preservation

The Levi-Civita connection $\overset{0}{\nabla}$ is metric preserving, and the parallel transport does not change the magnitude of a vector for any c. Equivalently, for any c,

$$\left\langle \prod_{\boldsymbol{\xi}}^{0}{}_{\boldsymbol{\xi}}^{\boldsymbol{\xi}'} X, \prod_{\boldsymbol{\xi}}^{0}{}_{\boldsymbol{\xi}}^{\boldsymbol{\xi}'} Y \right\rangle = \langle X, Y \rangle, \tag{52}$$

where $\langle\ ,\ \rangle$ is the inner product,

$$\langle X, Y \rangle = g_{ij} X^i Y^j. \tag{53}$$

This is rewritten in terms of the covariant derivative $\overset{0}{\nabla}$ for vector fields X, Y, Z as

$$Z\langle X, Y \rangle = \langle \overset{0}{\nabla}_Z\, X, Y \rangle + \langle X, \overset{0}{\nabla}_Z\, Y \rangle. \tag{54}$$

For

$$X = \boldsymbol{e}_i, \quad Y = \boldsymbol{e}_j, \quad Z = \boldsymbol{e}_k, \tag{55}$$

where

$$\boldsymbol{e}_i = \frac{\partial}{\partial \xi^i}, \quad i = 1, \ldots, n \tag{56}$$

are the natural basis vectors of the tangent space for coordinates $\boldsymbol{\xi}$, (54) is rewritten as

$$\partial_k g_{ij} = [ki\ ;\ j] + [kj\ ;\ i]. \tag{57}$$

The dual connections or covariant derivatives ∇ and ∇^* given in (38) and (39) are not generally metric preserving. Instead, the pair of dual affine connections ∇ and ∇^* preserves the metric in the following dual manner:

Theorem 1. *Let* $\prod$ *and* $\prod^*$ *be the parallel transports by two dually coupled connections. Then, we have*

$$\langle X, Y\rangle = \left\langle \prod X, \prod{}^* Y \right\rangle. \tag{58}$$

In terms of the covariant derivatives, this is written as

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle \tag{59}$$

and, in the component form,

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji}, \tag{60}$$

where

$$\Gamma_{kij} = g_{km}\Gamma^m_{ij}, \quad \Gamma^*_{kji} = g_{km}\Gamma^{*m}_{ji}. \tag{61}$$

Proof. We first prove (60). Since the Christoffel symbol is written as

$$[ij;k] = \left\{ \begin{matrix} l \\ ij \end{matrix} \right\} g_{lk} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) \tag{62}$$

in the covariant form, we easily have (60) by using (38), (39) and (57), since T_{ijk} is symmetric. Hence, (59) holds for the natural vector fields $X = \boldsymbol{e}_i$, $Y = \boldsymbol{e}_j$ and $Z = \boldsymbol{e}_k$. Therefore, it holds for any X, Y, Z. By considering $X(t)$ and $Y^*(t)$ which are parallel fields along curve c due to two covariant derivatives ∇ and ∇^*, we have

$$\frac{d}{dt}\langle X(t), Y^*(t)\rangle = 0, \tag{63}$$

implying the inner product is preserved by the two parallel shifts. □

The two dual connections (38), (39), or ∇, ∇^*, are obtained from g and T. On the contrary, when ∇ and ∇^* are dually coupled in the sense of metric preservation, we have a cubic tensor

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk} \tag{64}$$

which is symmetric, satisfying

$$\nabla_i g_{jk} = T_{ijk}, \quad \nabla^*_i g_{jk} = -T_{ijk}. \tag{65}$$

Recall that primal and dual geodesics $\boldsymbol{\xi}(t)$ and $\boldsymbol{\xi}^*(t)$ do not minimize the arc length, because ∇ and ∇^* are non-metric. Their average

$$\overset{0}{\nabla} = \frac{1}{2}(\nabla + \nabla^*) \tag{66}$$

is the Riemannian (Levi-Civita) connection and is metric. A straight line in a Euclidean space has the following properties:

1) It is a curve of the minimum length.
2) It does not change its direction.

Geodesics of the Riemannian (Levi-Civita) connection keep these properties. Geodesics in a statistical manifold have 2) but not 1). Instead, they have a duality.

It is easy to see that $\pm\alpha$-connections (43), (44) are dually coupled. When $\alpha = 0$, 0-connection is the Riemannian connection and is self-dual. Since a pair of dual connections are given from g and T, the α-geometry is given from g and αT.

The two connections in a statistical manifold give two Riemann–Christoffel curvatures. The two Riemann–Christoffel curvatures are mutually related, because of duality. The Riemann–Christoffel curvature tensor of ∇ is defined by a vector

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z \tag{67}$$

for vector fields X, Y, Z, where

$$[X,Y] = XY - YX. \tag{68}$$

Theorem 2. *The curvatures R and R^* of a statistical manifold satisfy*

$$\langle R(X,Y)Z, W\rangle = -\langle R^*(X,Y)W, Z\rangle. \tag{69}$$

Corollary. *$R = 0$ when and only when $R^* = 0$.*

We omit the proof of Theorem 2, since it is given by technical calculation from the definition of curvature. Instead, we give a simple proof for Corollary. Let $\prod$ and $\prod^*$ be dual parallel transport operators of a vector through a loop c which passes through $\boldsymbol{\xi}$. When $R = 0$, we have

$$A = \prod A \tag{70}$$

for any vector $A \in T_{\boldsymbol{\xi}}$ and vice versa. From the duality, we have

$$\langle A, B\rangle = \left\langle \prod A, \prod{}^* B \right\rangle = \left\langle A, \prod{}^* B \right\rangle \tag{71}$$

for any vector B. Hence

$$B = \prod{}^* B, \tag{72}$$

proving that $R^* = 0$.

In the case of probability distributions, g and T are determined from the log likelihood by the invariance principle. Let us consider the inverse problem. Let M be a manifold equipped with g and T. Is this a statistical manifold, in other words, is there a statistical model that gives g and T by (18) and (19)? This is a problem posed by [AM85] and affirmatively answered by [LE05].

Theorem 3. *Given an n-dimensional manifold M with g and T, there exists a probability simplex S_N with finite N, in which M is immersed isometrically and isocubically, that is, g and T are derived from those of S_N.*

The theorem justifies the use of the name 'statistical manifold'. For the proof, see [LE05] and [AY17].

3. Dual geometry induced from divergence

3.1. Divergence

Let $D[p(x,\boldsymbol{\xi}) : p(x,\boldsymbol{\xi}')]$ be a differentiable function of two points in a manifold $M = \{p(x,\boldsymbol{\xi})\}$ of probability distributions in a local chart. We denote it as $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$ in short for $D[p(x,\boldsymbol{\xi}) : p(x,\boldsymbol{\xi}')]$.

It is called a divergence, when the following three properties are satisfied:

1) $D[\boldsymbol{\xi} : \boldsymbol{\xi}'] \geq 0$,
2) $D[\boldsymbol{\xi} : \boldsymbol{\xi}'] = 0$, if and only if $\boldsymbol{\xi} = \boldsymbol{\xi}'$,
3) $D[\boldsymbol{\xi} : \boldsymbol{\xi} + d\boldsymbol{\xi}] = g_{ij}(\boldsymbol{\xi})\, d\xi^i\, d\xi^j + O(|d\boldsymbol{\xi}|^3)$,

for infinitesimally small $d\boldsymbol{\xi}$, where (g_{ij}) is a positive-definite matrix, $O(|d\boldsymbol{\xi}|^3)$ being higher-order terms of $d\boldsymbol{\xi}$.

We may add one more:

4) Let $U_\beta(\boldsymbol{\xi})$ be a subset of M called the β-neighborhood of $\boldsymbol{\xi}$, defined by

$$U_\beta(\boldsymbol{\xi}) = \{\boldsymbol{\xi}' \mid D[\boldsymbol{\xi} : \boldsymbol{\xi}'] < \beta\}. \tag{73}$$

Then, $U_\beta(\boldsymbol{\xi}) \subset U_{\beta'}(\boldsymbol{\xi})$, when $\beta < \beta'$.

A divergence $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$ is invariant, when it does not change if we use a sufficient statistic $s(x)$ instead of x. It is said to be additive when there exists a function $d(p,q)$ that

$$D[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \int d\{p(x;\boldsymbol{\xi}), q(x;\boldsymbol{\xi}')\}\, d\mu(x) \tag{74}$$

or in the discrete case

$$D[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \sum d\{p_i(\boldsymbol{\xi}), p_i(\boldsymbol{\xi}')\}. \tag{75}$$

A divergence $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$ induces a Riemannian metric g together with dually coupled affine connections ([EG83]). Let us introduce the following

notation $\partial_{i_1,\ldots,i_m\,;\,j_1,\ldots,j_k}$ for differentiation of $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$ with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, for example,

$$\partial_{i;j} D = \frac{\partial^2}{\partial \xi^i \partial \xi'^j} D[\boldsymbol{\xi} : \boldsymbol{\xi}'], \tag{76}$$

$$\partial_{i;jk} D = \frac{\partial^3}{\partial \xi^i \partial \xi'^j \partial \xi'^k} D[\boldsymbol{\xi} : \boldsymbol{\xi}']. \tag{77}$$

Theorem 4. *Given divergence* $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$,

$$g_{ij} = \partial_{ij} D[\boldsymbol{\xi} : \boldsymbol{\xi}']_{\boldsymbol{\xi}'=\boldsymbol{\xi}} = -\partial_{i;j} D[\boldsymbol{\xi} : \boldsymbol{\xi}']_{\boldsymbol{\xi}'=\boldsymbol{\xi}} \tag{78}$$

is positive-definite, playing the role of a Riemannian metric, and

$$\Gamma_{ijk} = -\partial_{ij;k} D[\boldsymbol{\xi} : \boldsymbol{\xi}']_{\boldsymbol{\xi}'=\boldsymbol{\xi}}, \tag{79}$$

$$\Gamma^*_{ijk} = -\partial_{k;ij} D[\boldsymbol{\xi} : \boldsymbol{\xi}']_{\boldsymbol{\xi}'=\boldsymbol{\xi}} \tag{80}$$

are the coefficients of dually coupled affine connections. The cubic tensor is given by

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}. \tag{81}$$

Proof. It is easy to see from the requirement for a divergence D that g in (78) is positive-definite. By differentiating the right side of (78) further with respect to $\boldsymbol{\xi}$, we have

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji}. \tag{82}$$

This shows that Γ and Γ^* are dually coupled. □

For a divergence D, we define its dual by

$$D^*[\boldsymbol{\xi} : \boldsymbol{\xi}'] = D[\boldsymbol{\xi}' : \boldsymbol{\xi}]. \tag{83}$$

Then, D^* gives the same dual geometry as that of D, except that ∇ and ∇^* are interchanged. A divergence induces a dual geometry. Conversely, there always exists a divergence for a Riemannian manifold having dually coupled affine connections. It is not difficult to construct a divergence from g and T ([MAT93]). However, this divergence is not unique.

Theorem 5. *Let* f *be a monotonically increasing differentiable function satisfying* $f(0) = 0$ *and* $f'(0) = 1$. *Then, given* $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$,

$$\tilde{D}[\boldsymbol{\xi} : \boldsymbol{\xi}'] = f(D[\boldsymbol{\xi} : \boldsymbol{\xi}']) \tag{84}$$

is a divergence inducing the same dual geometrical structure.

Proof. From

$$\partial_i f(D) = f'(D)\partial_i D, \tag{85}$$

we have

$$\partial_{i;j} f(D) = f''(D)\partial_i D\partial_{;j} D + f'(D)\partial_{i;j} D. \tag{86}$$

By evaluating the above at $\boldsymbol{\xi}' = \boldsymbol{\xi}$, we have

$$g_{ij} = \tilde{g}_{ij} \tag{87}$$

because of

$$\partial_i D[\boldsymbol{\xi} : \boldsymbol{\xi}']_{\boldsymbol{\xi}'=\boldsymbol{\xi}} = 0 \tag{88}$$

and $f'(0) = 1$. Similarly, we have

$$\Gamma_{ijk} = \tilde{\Gamma}_{ijk}. \tag{89}$$

□

A divergence $D[\boldsymbol{\xi} : \boldsymbol{\xi}']$ is not necessarily symmetric with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$. When it is symmetric, $T = 0$. In this case, the manifold is self-dual and $\nabla = \nabla^*$ is the Levi-Civita connection. If we use the half of the square of the Riemannian distance as a divergence, the derived geometry is the same as the original one.

We give a class of divergences for a manifold of probability distributions. A typical one is f-divergence ([CSI67], [MOR63]) using a convex function f satisfying

$$f(1) = 0, \quad f''(1) = 1. \tag{90}$$

The f-divergence between $p(x, \boldsymbol{\xi})$ and $p(x, \boldsymbol{\xi}')$ is defined by

$$D_f[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \int p(x, \boldsymbol{\xi}) f\Big\{\frac{p(x, \boldsymbol{\xi}')}{p(x, \boldsymbol{\xi})}\Big\}\, d\mu(x). \tag{91}$$

The class of f-divergences includes various well-known divergences.

The Kullback–Leibler divergence (KL-divergence) is an f-divergence with

$$f(u) = -\log u, \tag{92}$$

$$D_{KL}[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \int p(x, \boldsymbol{\xi}) \log \frac{p(x, \boldsymbol{\xi})}{p(x, \boldsymbol{\xi}')}\, d\mu(x). \tag{93}$$

For real α, the α-divergence is defined by the α-function

$$f_\alpha(u) = \frac{4}{1 - \alpha^2}(1 - u^{\frac{1+\alpha}{2}}), \tag{94}$$

giving

$$D_\alpha[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \frac{4}{1-\alpha^2}\Big(1 - \int p(x,\boldsymbol{\xi})^{\frac{1-\alpha}{2}} p(x,\boldsymbol{\xi}')^{\frac{1+\alpha}{2}}\, d\mu(x)\Big), \quad \alpha \neq \pm 1. \tag{95}$$

The square of the Hellinger distance is given by $\alpha = 0$,

$$f_{\frac{1}{2}}(u) = 4(1-\sqrt{u}), \tag{96}$$

$$D_0[\boldsymbol{\xi} : \boldsymbol{\xi}'] = 4\Big(1 - \int \sqrt{p(x)q(x)}\, d\mu(x)\Big). \tag{97}$$

This is a symmetric divergence. The KL-divergence and its dual are derived from the α-divergence by taking limit $\alpha \to \mp 1$. The f-divergence is invariant and additive. There are many non-invariant and non-additive divergences. The Wasserstein divergence we study later is such an example.

3.2. Transformation of divergence

Given divergence $D[\boldsymbol{\xi}, \boldsymbol{\xi}']$, we introduce another divergence

$$\tilde{D}[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \sigma(\boldsymbol{\xi}, \boldsymbol{\xi}') D[\boldsymbol{\xi} : \boldsymbol{\xi}'], \tag{98}$$

$$\sigma(\boldsymbol{\xi}, \boldsymbol{\xi}') = \exp\{\lambda(\boldsymbol{\xi}) + \tau(\boldsymbol{\xi}')\} \tag{99}$$

by using functions λ and τ. This is called a conformal transformation of divergence ([MATS10], [AOM12]). We study the change in the geometrical structure due to a conformal transformation of divergence.

Theorem 6. *The geometry of M is changed by a conformal transformation of divergence as*

$$\tilde{g}_{ij} = \sigma g_{jk}, \tag{100}$$

$$\tilde{T}_{ijk} = \sigma(T_{ijk} + \{\lambda, \tau, g\}_{ijk}), \tag{101}$$

$$\{\lambda, \tau, g\}_{ijk} = \partial_i(\tau - \lambda) g_{jk} + \partial_j(\tau - \lambda) g_{ik} + \partial_k(\tau - \lambda) g_{ij}, \tag{102}$$

where $\sigma = \sigma(\boldsymbol{\xi}, \boldsymbol{\xi})$.

Proof. By differentiation, we have

$$\partial_i \tilde{D} = \sigma(\partial_i \lambda D + \partial_i D), \tag{103}$$

$$\partial_{i;j} \tilde{D} = \sigma(\partial_i \lambda \partial_j \tau D + \partial_i \lambda \partial_{;j} D + \partial_j \tau \partial_i D + \partial_{i;j} D). \tag{104}$$

Evaluating the above at $\boldsymbol{\xi} = \boldsymbol{\xi}'$, we have (100). The $\tilde{T}$ is calculated similarly. □

4. Dually flat manifold

4.1. Geometry of exponential family

A statistical manifold is dually flat when

$$R = R^* = 0. \tag{105}$$

However, it is not generally Euclidean, because the Riemannian curvature due to g is generally not equal to 0. A dually flat manifold inherits nice properties from the Euclidean space as follows. Before stating them, we study the structure of an exponential family as a typical example of the dually flat manifold.

We rewrite (2) as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\theta^i x_i - \psi(\boldsymbol{\theta})\} \tag{106}$$

by using $\boldsymbol{\theta}$ denoting natural parameters instead of $\boldsymbol{\xi}$. The expectation of random variable $\boldsymbol{x} = (x_i)$ is given by

$$\boldsymbol{\eta} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}). \tag{107}$$

Hence $\boldsymbol{\eta}$ is called the expectation parameter. Differentiating (106), we have

$$\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{\theta}) = \Big\{ \boldsymbol{x} - \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \Big\} p(\boldsymbol{x}, \boldsymbol{\theta}). \tag{108}$$

From

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \int p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}) = 0, \tag{109}$$

we have

$$\boldsymbol{\eta} = \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta}). \tag{110}$$

Similarly, differentiating (108) again and integrating it, we have the variance of $\boldsymbol{x}$,

$$\mathrm{E}[(\boldsymbol{x} - \boldsymbol{\eta})(\boldsymbol{x} - \boldsymbol{\eta})^T] = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta}). \tag{111}$$

This shows that $\psi(\boldsymbol{\theta})$ is a convex function.

The expectation parameter $\boldsymbol{\eta}$ is the Legendre transform of $\boldsymbol{\theta}$ given by (110). Hence, it forms another local coordinate system. The Legendre dual of $\psi(\boldsymbol{\theta})$ is given by

$$\varphi(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}), \tag{112}$$

where $\boldsymbol{\theta}$ is regarded as a function of $\boldsymbol{\eta}$ implicitly given by (110). The $\varphi(\boldsymbol{\eta})$ is a convex function, and the inverse transform from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta} = \frac{\partial}{\partial \boldsymbol{\eta}} \varphi(\boldsymbol{\eta}). \tag{113}$$

We use the KL-divergence

$$D_{KL}[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{\theta}')} \, d\mu(\boldsymbol{x}) \tag{114}$$

to define the dual geometry. The divergence is calculated as

$$D_{KL}[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \psi(\boldsymbol{\theta}') + \varphi(\boldsymbol{\eta}) - \boldsymbol{\theta}' \cdot \boldsymbol{\eta}, \tag{115}$$

where $\boldsymbol{\eta}$ is the $\boldsymbol{\eta}$-coordinates of $\boldsymbol{\theta}$. The geometric quantities are calculated from (18), (19), (38) and (39) as

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}), \tag{116}$$
$$T_{ijk}(\boldsymbol{\theta}) = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}), \tag{117}$$
$$\Gamma_{ijk}(\boldsymbol{\theta}) = 0, \tag{118}$$
$$\Gamma^*_{ijk}(\boldsymbol{\theta}) = T_{ijk}. \tag{119}$$

Dually to the above, we calculate these quantities in the $\boldsymbol{\eta}$-coordinate system. We denote $\boldsymbol{\eta} = (\eta_i)$ by using the lower index and $\partial^i = \partial/\partial \eta_i$,

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}), \tag{120}$$
$$T^{ijk}(\boldsymbol{\eta}) = \partial^i \partial^j \partial^k \varphi(\boldsymbol{\eta}), \tag{121}$$
$$\Gamma^{ijk}(\boldsymbol{\eta}) = T^{ijk}, \tag{122}$$
$$\Gamma^{*ijk}(\boldsymbol{\eta}) = 0. \tag{123}$$

Note that

$$\partial_i = g_{ij} \partial^j, \tag{124}$$
$$\partial^j = g^{ji} \partial_i. \tag{125}$$

Therefore, g^{ij} and T^{ijk} are the contravariant components of g_{ij} and T_{ijk}, respectively.

From (118) and (123), we see that the manifold is dually flat, $R = R^* = 0$. Moreover, $\boldsymbol{\theta}$ denotes affine coordinates of ∇ connection and $\boldsymbol{\eta}$ denotes affine coordinates of ∇^* connection.

4.2. Fundamental theorem on dually flat manifold

We now study a general theory of a dually flat manifold. When M is dually flat, there exists a local coordinate system $\boldsymbol{\theta} = (\theta^i)$ such that

$$\Gamma_{ijk}(\boldsymbol{\theta}) = 0. \tag{126}$$

A geodesic is linear in $\boldsymbol{\theta}$ and coordinate curve θ^i is a geodesic. We call $\boldsymbol{\theta}$ a primal affine coordinate system.

There also exists a coordinate system $\boldsymbol{\eta} = (\eta_i)$,

$$\Gamma^{*ijk}(\boldsymbol{\eta}) = 0. \tag{127}$$

We call $\boldsymbol{\eta}$ the dual affine coordinate system and any dual geodesic is linear in $\boldsymbol{\eta}$. We sometimes call the $\boldsymbol{\theta}$ coordinates the e-coordinates and $\boldsymbol{\eta}$ coordinates the m-coordinates. This is because the primal geodesic is an exponential family and a dual geodesic is a mixture family in the case of probability distributions.

We have the following fundamental theorem for a dually flat manifold.

Theorem 7. *When M is dually flat, the following holds:*

1) *There exist two affine coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ with respect to ∇ and ∇^*, respectively, in a local chart and two convex functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ such that the natural basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}^j$ with respect to the two coordinate systems*

$$\boldsymbol{e}_i = \frac{\partial}{\partial \theta^i}, \quad \boldsymbol{e}^j = \frac{\partial}{\partial \eta_j} \tag{128}$$

are bi-orthonormal,

$$\langle \boldsymbol{e}_i, \boldsymbol{e}^j \rangle = \delta_i^j. \tag{129}$$

2) *The metric g is given by*

$$g_{ij} = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \partial_i \partial_j \psi(\boldsymbol{\theta}) \tag{130}$$

in the e-coordinates and

$$g^{ij} = \langle \boldsymbol{e}^i, \boldsymbol{e}^j \rangle = \partial^i \partial^j \varphi(\boldsymbol{\eta}) \tag{131}$$

in the m-coordinates.

3) *The cubic tensor T is given by*

$$T_{ijk} = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}), \quad T^{ijk} = \partial^i \partial^j \partial^k \varphi(\boldsymbol{\eta}). \tag{132}$$

4) *There exists a unique divergence between $\boldsymbol{\theta}, \boldsymbol{\theta}' \in M$, called a canonical divergence,*

$$D[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}', \tag{133}$$

where $\boldsymbol{\eta}'$ denotes the $\boldsymbol{\eta}$-coordinates of $\boldsymbol{\theta}'$.

Proof. We consider the $\boldsymbol{\theta}$-coordinate system. From (60) and (126), we have

$$\partial_i g_{jk} = \Gamma^*_{ikj}. \tag{134}$$

Because Γ^*_{ikj} is symmetric (torsion-free) with respect to i and k, we have

$$\partial_i g_{jk} = \partial_k g_{ji}, \tag{135}$$

which is

$$\partial_i g_{k\cdot} = \partial_k g_{i\cdot} \tag{136}$$

by suppressing index j. Hence, there exists a function $\psi_\cdot$, satisfying

$$g_{i\cdot} = \partial_i \psi_\cdot \tag{137}$$

or

$$g_{ij} = \partial_i \psi_j. \tag{138}$$

Since g_{ij} is symmetric,

$$\partial_i \psi_j = \partial_j \psi_i, \tag{139}$$

which guarantees the existence of ψ such that

$$\psi_j = \partial_j \psi \tag{140}$$

and

$$g_{ij} = \partial_i \partial_j \psi. \tag{141}$$

Since $\nabla_i = \partial_i$ in this case, we have

$$T_{ijk} = \partial_i \partial_j \partial_k \psi. \tag{142}$$

By a similar argument, the existence of the dual potential $\varphi(\boldsymbol{\eta})$ is guaranteed in the dual coordinates $\boldsymbol{\eta}$.

Note that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are not unique, because any affine transformations for constant matrices $\boldsymbol{A}$, $\boldsymbol{A}'$ and vectors $\boldsymbol{b}, \boldsymbol{b}'$,

$$\tilde{\boldsymbol{\theta}} = A\boldsymbol{\theta} + \boldsymbol{b}, \tag{143}$$

$$\tilde{\boldsymbol{\eta}} = A'\boldsymbol{\eta} + \boldsymbol{b}' \tag{144}$$

give other affine coordinate systems for which (126) and (127) hold. Because of this, we may choose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ such that their natural bases are biorthogonal at one point, that is (129) holds, and then everywhere by virtue of (59). The canonical divergence (133) is constructed from $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ and is unique even though $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are not uniquely determined. □

4.3. Geometry of Legendre transformation

A dually flat manifold M has a convex function $\psi(\boldsymbol{\theta})$ in the affine coordinates satisfying (130) and (132). Conversely, a strictly convex function $\psi(\boldsymbol{\theta})$ generates a dually flat structure. When $\psi(\boldsymbol{\theta})$ is given, we define a flat affine connection

$$\Gamma_{ijk}(\boldsymbol{\theta}) = 0, \tag{145}$$

in terms of $\boldsymbol{\theta}$-coordinates. We also define a Riemannian metric by

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}). \tag{146}$$

The manifold is flat, $R = 0$ and thus $R^* = 0$. The dual affine coordinates $\boldsymbol{\eta}$ is given by the Legendre transformation

$$\eta_i = \partial_i \psi(\boldsymbol{\theta}). \tag{147}$$

There exists the dual potential $\varphi(\boldsymbol{\eta})$ defined by

$$\varphi(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}), \tag{148}$$

where

$$\theta^i = \partial^i \psi(\boldsymbol{\eta}) \tag{149}$$

is the inverse transformation of (147).

The canonical divergence (133) is known as the Bregman divergence ([BRE67])

$$D[\boldsymbol{\theta}, \boldsymbol{\theta}'] = \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}') - \nabla\psi(\boldsymbol{\theta}') \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}') \tag{150}$$

constructed from $\psi(\boldsymbol{\theta})$. The dually flat theory is regarded as the geometry of the Legendre transformation, when we supplement it with the canonical divergence.

The exponential family plays a guiding role in defining a dually flat manifold. Given a dually flat manifold M with convex $\psi(\boldsymbol{\theta})$, is it possible to have an exponential family that has the same geometric structure? The problem was affirmatively answered by [BAN05].

When $\psi(\boldsymbol{\theta})$ is given, we consider an exponential family

$$p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}\, d\mu(\boldsymbol{x}). \tag{151}$$

This is possible if we can find a measure $\mu(\boldsymbol{x})$ on $\boldsymbol{x} \in \Omega$ that satisfies, given $\psi(\boldsymbol{\theta})$,

$$\exp\{\psi(\boldsymbol{\theta})\} = \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x})\, d\mu(\boldsymbol{x}). \tag{152}$$

This is the problem of finding the inverse Laplace transform of $\exp\{\psi(\boldsymbol{\theta})\}$. It is possible to find $\mu(\boldsymbol{x})$ under a certain regularity condition on $\psi(\boldsymbol{\theta})$.

We immediately see that the canonical divergence of an exponential family is the KL-divergence. The KL-divergence is used frequently in statistics, information theory and other fields without any justification. The present theory shows that it is the canonical divergence when the underlying manifold is dually flat.

4.4. Generalized Pythagorean theorem and projection theorem

We have the following fundamental theorem, which is a generalization of the Pythagorean theorem applicable to a dually flat manifold M (Fig. 3).

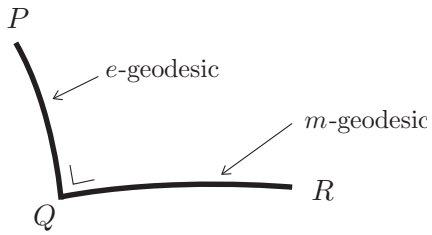

Fig. 3. Pythagorean theorem

Theorem 8. *For three points P, Q, R in a dually flat M and the canonical divergence D,*

$$D[P:Q] + D[Q:R] = D[P:R] \tag{153}$$

when the e-geodesic connecting P and Q is orthogonal at Q to the m-geodesic connecting Q and R. Dually,

$$D^*[P:Q] + D^*[Q:R] = D^*[P:R] \tag{154}$$

when the m-geodesic connecting P and Q is orthogonal to the e-geodesic connecting Q and R.

Proof. From (133), we have by calculations

$$D[P:Q] + D[Q:R] - D[P:R] = (\boldsymbol{\theta}_P - \boldsymbol{\theta}_Q) \cdot (\boldsymbol{\eta}_R - \boldsymbol{\eta}_Q), \tag{155}$$

where $\boldsymbol{\theta}_P$, $\boldsymbol{\eta}_P$, etc. are the $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-coordinates of P, etc. The e-geodesic connecting P and Q is

$$\boldsymbol{\theta}(t) = (1-t)\boldsymbol{\theta}_P + t\boldsymbol{\theta}_Q, \tag{156}$$

so its tangent at Q is

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}_Q - \boldsymbol{\theta}_P. \tag{157}$$

Similarly, the m-geodesic connecting Q and R is

$$\boldsymbol{\eta}(t) = (1-t)\boldsymbol{\eta}_Q + t\boldsymbol{\eta}_R, \tag{158}$$

and its tangent at Q is

$$\dot{\boldsymbol{\eta}} = \boldsymbol{\eta}_R - \boldsymbol{\eta}_Q. \tag{159}$$

Hence, the right-hand side of (155) vanishes. □

This is a generalization of the Pythagorean theorem in a Euclidean space, since this is a self-dual flat manifold and its canonical divergence is the half of the square of the Euclidean distance

$$D[P:Q] = \frac{1}{2}\sum_i (\theta_P^i - \theta_Q^i)^2. \tag{160}$$

As a consequence, we have the following projection theorem (Fig. 4).

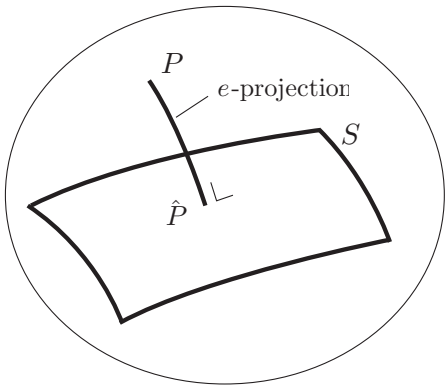

Fig. 4. Projection theorem

Theorem 9. *Let S be a smooth submanifold in a dually flat M. Given $P \in M$ outside of S, and letting $\hat{P} \in S$ be the minimizer of $D[P:Q]$, $Q \in S$. Then, the e-geodesic connecting P and $\hat{P}$ is orthogonal to S at $\hat{P}$. Dually, let $\hat{P}^* \in S$ be the minimizer of $D^*[P:Q]$, $Q \in S$. Then, the m-geodesic connecting P and $\hat{P}^*$ is orthogonal to S at $\hat{P}^*$.*

Proof. We prove only the former part. When $\hat{P}$ is an extreme point of $D[P:Q]$, $Q \in S$, we consider a small deviation $\hat{P} + dP \in S$. Then, dP is regarded as a tangent vector of S orthogonal to the e-geodesic connecting P and $\hat{P}$. Therefore,

$$D[P:\hat{P}+dP] = D[P:\hat{P}] + D[\hat{P}:\hat{P}+dP] \tag{161}$$

$$\geq D[P:\hat{P}], \tag{162}$$

proving the theorem. The $\hat{P}$ is called the e-projection of P to S and the $\hat{P}^*$ is called the m-projection of P to S.

When S is m-flat, the e-projection is unique. When S is e-flat, the m-projection is unique.

5. Semiparametric statistical model and estimating function

5.1. Rough sketch on function space of probability distributions

We give a rough sketch on the geometry of a function space of probability distributions. There are delicate problems for extending the geometry from a finite dimensional space to a function space of infinite dimensions. See, e.g. [CEP07], [PIS95], [AY17], etc. The theories in this section might not be rigorously founded, although they are useful in many applications.

Let $S = \{p(x)\}$ be a set of all probability distributions, where x is a random variable in $\boldsymbol{R}^n$, $p(x)$ is a density function equivalent to the Lebesgue measure. We assume that $p(x)$ is differentiable and x has moments of any orders. We attach random variables $w(x)$ to each $p(x) \in S$, which satisfy

$$\mathrm{E}_p[w(x)] = 0, \tag{163}$$

$$\mathrm{E}_p[\{w(x)\}^2] < \infty, \tag{164}$$

where E_p is the expectation with respect to $p(x)$. Since all $w(x)$'s form a linear space, we consider it as a tangent space at p,

$$T_p = \{w(x)\}. \tag{165}$$

Intuitively, a small deviation $w(x) = \delta \log p(x)$ of $p(x)$ is considered as an infinitesimally small tangent vector, since it satisfies (163) provided (164) is satisfied. We introduce the Fisher information metric defined by

$$ds^2 = \mathrm{E}_p[\{w(x)\}^2]\, dt^2, \tag{166}$$

where ds^2 is the square of the magnitude of $\delta \log p(x)$. Because of (164), the tangent space is a Hilbert space.

Let $c : p(x,t)$ be a smooth curve parametrized by t, where $p(x,0) = p(x)$. We define the tangent vector of the curve by

$$w_c(x) = \frac{d}{dt} \log p(x,t)\Big|_{t=0}. \tag{167}$$

From

$$w_c(x) = \frac{1}{p(x,t)} \frac{d}{dt} p(x,t)\Big|_{t=0}, \tag{168}$$

by differentiating

$$\int p(x,t)\, dx = 1, \tag{169}$$

we have

$$\mathrm{E}_p[w_c(x)] = 0. \tag{170}$$

We further assume that

$$\mathrm{E}_p[\{w_c(x)^2\}] < \infty, \tag{171}$$

excluding tangent vectors which do not satisfy (171). Then, the squared magnitude of $\delta \log p(x) = w_c(x)\, dt$ is

$$ds^2 = g\, dt^2, \tag{172}$$

$$g = \mathrm{E}_p[\{w_c(x)\}^2]. \tag{173}$$

The tangent space T_p is a Hilbert space attached to $p(x)$, consisting of all such $w(x)$. The inner product of two tangent vectors $w(x)$ and $v(x)$ is

$$\langle w, v \rangle = \mathrm{E}_p[w(x)v(x)]. \tag{174}$$

In order to define a dual pair of affine connections, we define two parallel transports of tangent vector $w(x)$ from $p(x)$ to $q(x)$ by

$$\prod_p^{e\ q} w(x) = w(x) - \mathrm{E}_q[w(x)], \tag{175}$$

$$\prod_p^{m\ q} w(x) = \frac{p(x)}{q(x)} w(x), \tag{176}$$

provided they belong to T_q. It is easy to confirm the following proposition, which shows the metric preservation by the pair of dual parallel transports. The proof is easy. (We do not use the term theorem but proposition, because we do not specify the exact conditions under which it holds.) See [AMK97] and also [AM16].

Proposition. *The inner product of two tangent vectors is kept constant by the two parallel transports,*

$$\langle w(x), v(x) \rangle_p = \left\langle \prod_p^{e\ q} w(x), \prod_p^{m\ q} v(x) \right\rangle_q. \tag{177}$$

5.2. Semiparametric statistical model

When we have interest in a specific set of parameters $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ concerning unknown probability densities $p(x)$, we single out them and denote the distribution by $p(x, \boldsymbol{\xi})$. Here, the densities $p(x, \boldsymbol{\xi})$ may have a function degrees of freedom. We call it a semiparametric statistical model, denoting $S = \{p(x, \boldsymbol{\xi})\}$. We have interest in estimating $\boldsymbol{\xi}$ from observed

iid data $D = \{x_1, \ldots, x_N\}$, not the density function $p(x, \boldsymbol{\xi})$ itself. Here, $\boldsymbol{\xi}$ is called the parameter of interest.

A simple example is the location model

$$S = \{p(x - \xi)\}, \tag{178}$$

where $p(x)$ is an arbitrary probability density function satisfying regularity conditions such as the continuity and existence of moments. We also request

$$\int xp(x)\, dx = 0, \tag{179}$$

which is necessary for identifying ξ. The statistical problem is to estimate the "mean" (or the "center") ξ from a number of independent observations D in spite that the exact form of $p(x)$ is unknown. When p is fixed, for example, to be a Gaussian distribution with variance 1,

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{x^2}{2} \Big\}, \tag{180}$$

S reduces to a simple parametric statistical model and the maximum-likelihood estimator $\hat{\xi}$ is given by the arithmetic mean of data

$$\hat{\xi} = \frac{1}{N} \sum x_i. \tag{181}$$

This is optimal, but it is not optimal for a general (unknown) p.

The location-scale model has two parameters of interest $\boldsymbol{\xi} = (\mu, \sigma^2)$, such that the semiparametric model is given by

$$p(x, \boldsymbol{\xi}) = p\Big\{ \frac{(x - \mu)^2}{\sigma^2} \Big\}, \tag{182}$$

where p satisfies

$$\int x^2 p(x)\, dx = 1 \tag{183}$$

in addition to (179).

The Neyman–Scott problem, which had been an unsolved problem for long years, is understood from the semiparametric point of view. Let us consider a parametric statistical model

$$M = \{p(x, \boldsymbol{\zeta}, \boldsymbol{\xi})\} \tag{184}$$

specified by two types of (finite-dimensional) parameters $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. The former is the parameter of interest which we want to estimate and $\boldsymbol{\zeta}$ is the nuisance parameters which we do not care about. We estimate $\boldsymbol{\xi}$ from N independently generated data $D = \{x_1, \ldots, x_N\}$. However, x_i is generated from distribution $p(x, \boldsymbol{\zeta}_i, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is common (but unknown)

for all data but $\boldsymbol{\zeta}_i$ are unknown and may be different for each i. Neyman and Scott presented the problem, showing that the maximum likelihood estimator is not necessarily consistent nor efficient. To search for the optimal estimator had been a long-standing unsolved problem, bothering theoretical statisticians.

We show a typical example. Let $\bar{y}$ be a random variable proportional to $\bar{z}$,

$$\bar{y} = \xi \bar{z}, \tag{185}$$

and we want to know ξ, the ratio of proportion from noisy observations of $(\bar{y}, \bar{z})$. Here, ξ is the parameter of interest. Let $x = (y, z)$ be a pair of random variables y and z, which are noisy observations of $\bar{y}$ and $\bar{z}$,

$$y_i = \bar{y}_i + \varepsilon_i, \tag{186}$$

$$z_i = \bar{z}_i + \varepsilon_i'. \tag{187}$$

We assume that ε_i and ε_i' are independent Gaussian variables subject to $N(0, 1)$. We observe $x_i = (y_i, z_i)$, $i = 1, 2, \ldots, N$, where $\bar{y}_i = \xi \bar{z}_i$ and $\bar{z}_1, \ldots, \bar{z}_N$ may take different values,

$$\bar{z}_i = \zeta_i, \tag{188}$$

which are the nuisance parameters. Hence, $x_i = (y_i, z_i)$ are subject to $z_i \sim N(\zeta_i, 1)$, $y_i \sim N(\xi\zeta_i, 1)$. We then have a model

$$M = \{p(x, \xi, \zeta_i)\}. \tag{189}$$

We assume that $\boldsymbol{\zeta}_i$ are selected from an unknown distribution $k(\boldsymbol{\zeta})$ each time i. Then, we consider a mixture of statistical models,

$$p(x, \boldsymbol{\xi}; k) = \int k(\boldsymbol{\zeta}) p(x, \boldsymbol{\xi}, \boldsymbol{\zeta})\, d\boldsymbol{\zeta}. \tag{190}$$

We may regard that all x_i are generated independently from it. The family $S_M = \{p(x, \boldsymbol{\xi}; k)\}$ is called a mixture-type semiparametric model, which is a submanifold of S. It is specified by two types of parameters: One is finite-dimensional $\boldsymbol{\xi}$ which is to be estimated and the other is $k(\boldsymbol{\zeta})$ called a mixing function which has function degrees of freedom. A semiparametric model S_M is used for estimating the parameter $\boldsymbol{\xi}$ of interest without caring about the nuisance function parameter $k(\boldsymbol{\zeta})$. We present useful results intuitively along the spirit of "experimental mathematics". The results elucidate mathematical structure of the geometry of the semiparametric model.

5.3. Decomposition of tangent space

Since S_M is included in S, the tangent space T_p of S includes tangent vectors of S_M, which are given by those in the directions of parameters $\boldsymbol{\xi}$ of interest and those in the directions of nuisance mixing parameter $k(\boldsymbol{\zeta})$. We denote the first one by T^U,

$$T^U = \{\boldsymbol{u}(x, \boldsymbol{\xi}, k) = \partial_{\boldsymbol{\xi}} \log p(x, \boldsymbol{\xi}; k)\} \tag{191}$$

and call the space generated by the components u_i of $\boldsymbol{u}$ the tangent space of interest. The second one is given by

$$T^V = \{v(x, \boldsymbol{\xi}, k) = \partial_k \log p(x, \boldsymbol{\xi}; k)\}, \tag{192}$$

where ∂_k is the Fréchet differentiation. It is called the nuisance tangent space. When we consider a curve $k(\boldsymbol{\zeta}, t)$ passing through $k(\boldsymbol{\zeta}) = k(\boldsymbol{\zeta}, 0)$, the tangent vector along this curve is

$$v(x, \boldsymbol{\xi}, k) = \frac{d}{dt} \log p\{x, \boldsymbol{\xi}; k(\zeta, t)\}\Big|_{t=0}. \tag{193}$$

We have the third tangent directions in S which are orthogonal to both T^U and T^V. Thus, the tangent space of S is decomposed into a direct sum,

$$T_p = (T_p^U \oplus T_p^V) \oplus T_p^A, \tag{194}$$

but it should be noted that T_p^U and T_p^V are not necessarily orthogonal.

5.4. Estimating function

An estimating function gives us a good means of estimating the parameters $\boldsymbol{\xi}$ of interest in a semiparametric model. An n-dimensional vector function $\boldsymbol{f}(x, \boldsymbol{\xi})$ is called an estimating function, when

$$\mathrm{E}_{p(x,\boldsymbol{\xi},k)}[\boldsymbol{f}(x, \boldsymbol{\xi})] = 0, \tag{195}$$

$$A = \mathrm{E}_{p(x,\boldsymbol{\xi},k)}[\partial_{\boldsymbol{\xi}} \boldsymbol{f}(x, \boldsymbol{\xi})] > 0, \tag{196}$$

where $A > 0$ implies that A is symmetric and positive-definite for any $(\boldsymbol{\xi}, k)$. The two conditions are equivalent to

$$\mathrm{E}_{\boldsymbol{\xi},\boldsymbol{\zeta}}[\boldsymbol{f}(x, \boldsymbol{\xi})] = 0, \tag{197}$$

$$\mathrm{E}_{\boldsymbol{\xi},\boldsymbol{\zeta}}[\partial_{\boldsymbol{\xi}} \boldsymbol{f}(x, \boldsymbol{\xi})] > 0, \tag{198}$$

for the finite-dimensional model

$$t\{p(x, \boldsymbol{\xi}, \boldsymbol{\zeta})\}, \tag{199}$$

where $\mathrm{E}_{\boldsymbol{\xi},\boldsymbol{\zeta}}$ is the expectation with respect to $p(x,\boldsymbol{\xi},\boldsymbol{\zeta})$. The latter conditions are easier to check. $\boldsymbol{u}(x,\boldsymbol{\xi},k)$ is a generalization of the score vector

$$\boldsymbol{u}(x,\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} \log p(x,\boldsymbol{\xi}) \tag{200}$$

of a usual parametric statistical model $p(x,\boldsymbol{\xi})$. When an estimating function exists, an estimator $\hat{\boldsymbol{\xi}}$ is obtained, by replacing the expectation by the empirical mean, as the solution of

$$\frac{1}{N}\sum \boldsymbol{f}(x_i,\boldsymbol{\xi}) = 0. \tag{201}$$

Let $\hat{\boldsymbol{\xi}}$ be the estimator for a mixture semiparametric model S_M that satisfies (201). The estimation error $\boldsymbol{e}$ is written as

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \boldsymbol{e}. \tag{202}$$

By expanding $\boldsymbol{f}(x_i,\hat{\boldsymbol{\xi}})$, we have

$$0 = \frac{1}{N}\sum f(x_i,\hat{\boldsymbol{\xi}}) \approx \frac{1}{N}\sum \boldsymbol{f}(x_i,\boldsymbol{\xi}) + \frac{1}{N}\partial_{\boldsymbol{\xi}}\boldsymbol{f}(x_i,\boldsymbol{\xi})\cdot\boldsymbol{e}. \tag{203}$$

When N is large, the central limit theorem guarantees that

$$\boldsymbol{r} = \frac{1}{\sqrt{N}}\sum \boldsymbol{f}(x_i,\boldsymbol{\xi}) \tag{204}$$

is asymptotically Gaussian subject to $N(0,\mathrm{E}[\boldsymbol{f}\boldsymbol{f}^T])$. The law of large numbers guarantees that $(1/N)\sum \partial_{\boldsymbol{\xi}}\boldsymbol{f}(x_i,\boldsymbol{\xi})$ converges to

$$A = \mathrm{E}[\partial_{\boldsymbol{\xi}}\boldsymbol{f}(x,\boldsymbol{\xi})] > 0. \tag{205}$$

Hence, we have

$$\boldsymbol{e} \approx -\frac{1}{\sqrt{N}}A^{-1}\boldsymbol{r}. \tag{206}$$

Proposition. *The estimator $\hat{\boldsymbol{\xi}}$ is consistent, asymptotically Gaussian and its asymptotic error covariance matrix is given by*

$$\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^T] \approx \frac{1}{N}\mathrm{E}[(A^{-1}\boldsymbol{f})(A^{-1}\boldsymbol{f})^T]. \tag{207}$$

The geometry of estimating functions is studied by using the tangent bundle structure. An estimating function is geometrically characterized in this approach to give all the estimating functions.

A semiparametric model S_M is a (curved) submanifold embedded in S. We consider parallel transport of a tangent vector from $p\{x,\boldsymbol{\xi},k_1(t)\}$ to $p\{x,\boldsymbol{\xi},k_2(t)\}$ in the manifold S of all distributions. We show that the estimating function is characterized in terms of the e-parallel transport.

Proposition. *An estimating function* $\boldsymbol{f}(x, \boldsymbol{\xi})$ *is orthogonal to* T^V *at any* $p(x, \boldsymbol{\xi}, k)$*, where* T^V *consists of tangent vectors along the nuisance parameters* $\boldsymbol{\zeta}$*,*

$$T^V = \Big\{ v(x, \boldsymbol{\xi}, \boldsymbol{\zeta}) = \frac{\partial}{\partial \boldsymbol{\zeta}} \log p(x, \boldsymbol{\xi}, \boldsymbol{\zeta}) \Big\}. \tag{208}$$

Moreover, it is invariant under the e-parallel transport,

$$\prod\nolimits_{k_1}^{e\ k_2} \boldsymbol{f}(x, \boldsymbol{\xi}) = \boldsymbol{f}(x, \boldsymbol{\xi}). \tag{209}$$

Outline of proof. From (175) and (197), it is immediate to see that $\boldsymbol{f}$ is e-invariant, that is,

$$\mathrm{E}[\boldsymbol{f}(x, \boldsymbol{\xi})] = 0 \tag{210}$$

for any k. By differentiating (210) in the direction of curve $k(\boldsymbol{\zeta}, t)$, we have

$$\langle v(x, \boldsymbol{\zeta}, \boldsymbol{\xi}), \boldsymbol{f}(x, \boldsymbol{\xi}) \rangle = 0, \tag{211}$$

which shows that $\boldsymbol{f}$ is orthogonal to T^V. □

By differentiating (210) with respect to $\boldsymbol{\xi}$, we have

$$\langle \boldsymbol{f}(x, \boldsymbol{\xi}), \boldsymbol{u}(x, \boldsymbol{\xi}, k) \rangle = -\mathrm{E}\Big[\frac{\partial}{\partial \boldsymbol{\xi}} \boldsymbol{f}(x, \boldsymbol{\xi}) \Big] \neq 0. \tag{212}$$

Hence, $\boldsymbol{f}(x, \boldsymbol{\xi})$ includes a non-zero component in the direction of the score vector $\partial_{\boldsymbol{\xi}} \log p(x, \boldsymbol{\xi}, k)$. Let $\boldsymbol{u}^I(x, \boldsymbol{\xi}, k)$ be the projection of the score $\boldsymbol{u}(x, \boldsymbol{\xi}, k)$ to the orthogonal subspace of T^V. We call it an information score. The following proposition follows immediately.

Proposition. *An estimating function exists when and only when the information score* $\boldsymbol{u}^I$ *does not vanish. Any estimation function is a sum of the information score and an ancillary vector,*

$$\boldsymbol{f}(x, \boldsymbol{\xi}) = \boldsymbol{u}^I(x, \boldsymbol{\xi}, k) + \boldsymbol{a}(x), \tag{213}$$

$$\boldsymbol{a}(x) \in T^A. \tag{214}$$

Proposition. *When the true distribution is* $p(x, \boldsymbol{\xi}, k)$*, the best estimating function is the information score at* k*.*

Since we do not know the true $k(\boldsymbol{\zeta})$, we cannot find the best estimating function. However, if we guess $k(\boldsymbol{\zeta})$ adequately, then the guessed information score is $\boldsymbol{f} = \boldsymbol{u}^I + \boldsymbol{a}$ and gives a good estimating function yielding an asymptotically unbiased estimator. Note that if we use a guessed $k(\boldsymbol{\zeta})$ and use the score function itself for estimation, it does not necessarily give an unbiased estimator.

For many semiparametric estimation problems, we can analyze the structure of the estimating functions, giving the optimal and semi-optimal solutions. See [AMK97] for details, where the Neyman–Scott problem is fully explored (see also [AM16]). The geometry of estimating functions have been applied to various problems such as independent component analysis ([AMC97], [AM00]) and estimation of the statistic of the inter-spike intervals of a neuron under unknown firing rate ([MOA06]).

6. α-geometry: conformally-projectively flat geometry in S_n

6.1. α-divergence

When $\{M, g, T\}$ is dually flat, the geometry of $\{M, g, \alpha T\}$ is called the α-geometry. Since we can construct an exponential family (151) from (g, T), we study the α-geometry of an exponential family. For simplicity, we mainly study S_n as a typical example because the α-geometry has not yet been fully explored.

The α-geometry is induced by the α-divergence. In S_n, it is given by

$$D_\alpha[\boldsymbol{p} : \boldsymbol{q}] = \frac{4}{1-\alpha^2}\Big(1 - \sum p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\Big), \quad \alpha \neq \pm 1, \tag{215}$$

where α is a real parameter. When $\alpha = \pm 1$, we define

$$D_{-1}[\boldsymbol{p} : \boldsymbol{q}] = D_{KL}[\boldsymbol{p} : \boldsymbol{q}], \tag{216}$$

$$D_1[\boldsymbol{p} : \boldsymbol{q}] = D_{KL}[\boldsymbol{q} : \boldsymbol{p}] \tag{217}$$

by considering limit $\alpha \to \pm 1$.

The Riemannian metric derived from D_α is the Fisher information matrix g, not depending on α. However, the α-covariant derivative derived from D_α satisfies

$$\overset{\alpha}{\nabla}_i g_{jk} = \alpha T_{ijk}. \tag{218}$$

The α-geometry $(M, g, \alpha T)$, $\alpha \neq \pm 1$, is not dually flat, because the Riemann–Christoffel curvature is

$$R^\alpha_{ijkl} = (1-\alpha^2)(T_{kmi}T_{jln} - T_{kmj}T_{iln})g^{mn}. \tag{219}$$

The two connections $\overset{\alpha}{\nabla}$ and $\overset{-\alpha}{\nabla}$ are dually coupled. [KUR99] proved that they are dually projectively flat in S_n from the affine geometry point of view.

6.2. Affine differential geometry

Affine differential geometry ([NOS94]) is closely related to information geometry. Let M be an n-dimensional manifold. We immerse it in an $(n+1)$-dimensional affine space $\boldsymbol{R}^{n+1}$ by f,

$$f(p) \in \boldsymbol{R}^{n+1}, \quad p \in M. \tag{220}$$

We attach a transversal vector field $\boldsymbol{\xi}(p)$ to $f(M)$. Affine differential geometry involves the study of the geometrical structure of M induced from $\{f(p), \boldsymbol{\xi}(p)\}$, see Fig. 5. An affine fundamental form $h = (h_{ij})$ is derived,

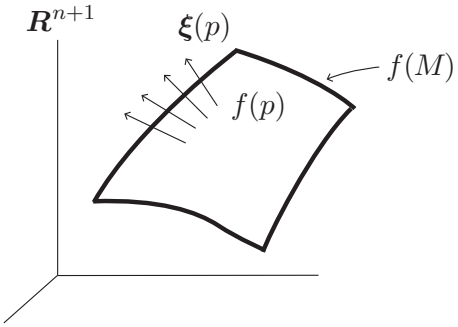

Fig. 5. Affine differential geometry

which is not necessarily positive-definite. When h is positive-definite, it gives a Riemannian metric. An affine connection is also induced in M.

Let $\boldsymbol{R}_{n+1}$ be the dual space of $\boldsymbol{R}^{n+1}$. We naturally consider an immersion $f_*(p) \in \boldsymbol{R}_{n+1}$ corresponding to $\{f(p), \boldsymbol{\xi}(p)\}$. They together induce a Riemannian metric and a dual pair of affine connections, when a certain condition is satisfied. In such a case, it is possible to study the dual geometry from the affine geometry point of view.

[SH07] studied the Hessian manifold, which is a dually flat manifold. [KUR94], [KUR02] defined a geometrical divergence when the induced connection is dually projectively flat. When the manifold is dually flat, it reduces to the canonical divergence. [MATS99] further studied immersion of M to $\boldsymbol{R}^{n+2}$, together with two transversal vector fields $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. A conformally-projectively flat connection is defined from such an immersion and a canonical divergence is also defined.

Not all n-dimensional statistical manifolds are realized by immersion to $\boldsymbol{R}^{n+1}$ or $\boldsymbol{R}^{n+2}$. The Lê theorem suggests that any statistical manifold is realized by immersion to S_N with finite N. It is interesting to characterize a statistical manifold by the number N that is the minimum for realizing it by immersion in S_N.

A statistical manifold is dually projectively flat when the curvature is constant and the reverse holds when $n \geq 3$ ([KUR99]). The geometry

constructed from the α-divergence is dually projectively flat. The following theorems are given by [KUR94].

Theorem 10. *When the α-geodesic connecting P and Q is orthogonal to the $-\alpha$-geodesic connecting Q and R for $P, Q, R, \in S_n$,*

$$D_\alpha[P:R] = D_\alpha[P:Q] + D_\alpha[Q:R] - \frac{1-\alpha^2}{4} D_\alpha[P:Q] D_\alpha[Q:R]. \quad (221)$$

From this, we have the projection theorem.

Theorem 11. *Let S be a smooth submanifold of S_n and $\hat{P}$ be the $-\alpha$-projection of P to S. Then the $-\alpha$-geodesic connecting P to $\hat{P}$ is orthogonal to S.*

6.3. Rényi divergence and Pythagorean theorem

We now show a theory given by [WON18], which is applicable to a general dually projectively flat manifold, although we state it only in S_n for simplicity. The key idea is exponential convexity (concavity) instead of convexity (concavity). Here, we study the case of $\alpha > 0$. Let us rewrite the probability distributions of S_n by using another parameterization $\boldsymbol{\xi}$,

$$p(x, \boldsymbol{\xi}) = (1 + \alpha \boldsymbol{\xi} \cdot \boldsymbol{x})^{-\frac{1}{\alpha}} e^{\varphi_\alpha(\boldsymbol{\xi})}, \quad (222)$$

$$x_i = \delta_i(x), \quad i = 1, \ldots, n, \quad (223)$$

where

$$\xi^i = \frac{1}{\alpha}\left\{\left(\frac{p_0}{p_i}\right)^\alpha - 1\right\}, \quad i = 1, \ldots, n, \quad (224)$$

$$\varphi_\alpha(\boldsymbol{\xi}) = \log p_0. \quad (225)$$

The potential function is written as

$$\varphi_\alpha(\boldsymbol{\xi}) = -\log \sum_{i=0}^{n} (1 + \alpha \xi^i)^{-\frac{1}{\alpha}}, \quad (226)$$

where we put $\xi^0 = 0$. The $\varphi_\alpha(\boldsymbol{\xi})$ is α-exponentially concave, when $\exp\{\alpha \varphi_\alpha(\boldsymbol{\xi})\}$ is concave, that is,

$$-\partial_i \partial_j \varphi_\alpha - \alpha \partial_i \varphi_\alpha \partial_j \varphi_\alpha > 0 \quad (227)$$

in the sense of matrix positive-definiteness.

We define a dual function by

$$\psi_\alpha(\boldsymbol{\eta}) = -\frac{1}{\alpha} \log \sum_{i=0}^{n} p_i^{1+\alpha}, \quad (228)$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is given by

$$\eta_i = \frac{p_i^{1+\alpha}}{\sum_{k=0}^{n} p_k^{1+\alpha}}. \tag{229}$$

A new α-divergence is defined by

$$\bar{D}_\alpha[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \frac{1}{\alpha}\log(1 + \alpha\boldsymbol{\xi}\cdot\boldsymbol{\eta}') - \varphi_\alpha(\boldsymbol{\xi}) - \psi_\alpha(\boldsymbol{\eta}'). \tag{230}$$

This is a Legendre-like duality, where $\log(1+\alpha\boldsymbol{\xi}\cdot\boldsymbol{\eta})$ is used instead of $\boldsymbol{\xi}\cdot\boldsymbol{\eta}$. It is easily proved that this is the Rényi $\hat{\alpha}$-divergence, where $\hat{\alpha} = 2\alpha + 1$,

$$\bar{D}_\alpha[\boldsymbol{p} : \boldsymbol{q}] = -\frac{2}{1-\hat{\alpha}}\log\Big(\sum p_i^{\frac{1-\hat{\alpha}}{2}} q_i^{\frac{1+\hat{\alpha}}{2}}\Big). \tag{231}$$

This is a function of the $\hat{\alpha}$-divergence,

$$\bar{D}_\alpha[\boldsymbol{p} : \boldsymbol{q}] = -\frac{2}{1-\hat{\alpha}}\log\Big\{1 - \frac{1-\hat{\alpha}^2}{4} D_{\hat{\alpha}}[\boldsymbol{p} : \boldsymbol{q}]\Big\}. \tag{232}$$

Therefore, the geometry induced by $\bar{D}_\alpha$ is essentially the same as that induced by $D_{\hat{\alpha}}$. [WON18] proved the following theorem.

Theorem 12. *Let P, Q, R be three points in S_n. When the $\hat{\alpha}$-geodesic connecting P and Q is orthogonal to the $-\hat{\alpha}$-geodesic connecting Q and R, we have*

$$\bar{D}_\alpha[P : Q] + \bar{D}_\alpha[Q : R] = \bar{D}_\alpha[P : R]. \tag{233}$$

It is surprising that Pythagorean and projection theorems hold even in a dually projectively flat manifold. However, it is possible to derive the Pythagrean relation from Kurose's formula (221). Recall that (221) is rewritten as

$$1 - \kappa D_{\hat{\alpha}}[P : R] = (1 - \kappa D_{\hat{\alpha}}[P : Q])(1 - \kappa D_{\hat{\alpha}}[Q : R]), \tag{234}$$

where

$$\kappa = \frac{1-\hat{\alpha}^2}{4} \tag{235}$$

is the scalar curvature. Hence, taking the logarithm of (234), we have the Pythagrean theorem, which is equivalent to (233).

6.4. Tsallis q-entropy and induced non-invariant dually flat structure

We consider the α-exponential family of probability distributions. Since this is closely related to Tsallis q-entropy ([TSA09], [NAU11]), we use the $q = (1+\alpha)/2$ instead of α. Let us define the q-logarithm by

$$\log_q u = \frac{1}{1-q}(u^{1-q} - 1), \tag{236}$$

and its inverse is

$$\exp_q(u) = \{1 + (1-q)u\}^{\frac{1}{1-q}}. \tag{237}$$

The q-exponential family is defined by

$$p(x, \boldsymbol{\theta}) = \exp_q\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta})\}, \tag{238}$$

where $\psi_q(\boldsymbol{\theta})$ corresponds to the normalization factor. In limit $q \to 1$, $\log_q u = \log u$, so this class includes the exponential family.

We first prove that $\psi_q(\boldsymbol{\theta})$ is a convex function.

Lemma 1. *ψ_q is a convex function of $\boldsymbol{\theta}$.*

Proof. By differentiating (238) with respect to $\boldsymbol{\theta}$, we have

$$\partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \theta)^q \{x_i - \partial_i \psi_q\} \tag{239}$$

and

$$\partial_i \partial_j p(\boldsymbol{x}, \boldsymbol{\theta}) = q p^{2q-1} \{x_i - \partial_i \psi_q\}\{x_j - \partial_j \psi_q\} - p^q \partial_i \partial_j \psi_q. \tag{240}$$

From

$$\partial_i \int p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}) = 0, \quad \partial_i \partial_j \int p(\boldsymbol{x}, \boldsymbol{\theta})\, d\mu(\boldsymbol{x}) = 0, \tag{241}$$

by putting

$$h_q(\boldsymbol{\theta}) = \int p(\boldsymbol{x}, \boldsymbol{\theta})^q\, d\mu(\boldsymbol{x}), \tag{242}$$

we have

$$\partial_i \psi_q(\boldsymbol{\theta}) = \eta_i = \frac{1}{h_q(\boldsymbol{\theta})} \int p(\boldsymbol{x}, \boldsymbol{\theta})^q x_i\, d\mu(\boldsymbol{x}) \tag{243}$$

and

$$\partial_i \partial_j \psi_q(\boldsymbol{\theta}) = \frac{q}{h_q(\boldsymbol{\theta})} \int \{x_i - \partial_i \psi_q\}\{x_j - \partial_j \psi_q\} p^{2q-1}\, d\mu(\boldsymbol{x}). \tag{244}$$

This latter shows that $\partial_i \partial_j \psi_q$ is positive-definite, that is, $\psi_q(\boldsymbol{\theta})$ is convex. □

We can construct a dually flat geometry from $\psi_q(\boldsymbol{\theta})$, which is not invariant and different from the α-geometry derived from the α-divergence or q-divergence ([AOM12]). The dual parameters derived from convex ψ_q are

$$\eta_i = \frac{1}{h_q} \int x_i p(\boldsymbol{x}, \boldsymbol{\theta})^q \, d\mu(\boldsymbol{x}), \tag{245}$$

which are different from the dual parameters in the case $\alpha = 1$ of an exponential family. The dual potential is

$$\varphi_q(\boldsymbol{\theta}) = \frac{1}{1-q} \Big\{ \frac{1}{h_q(\boldsymbol{\theta})} - 1 \Big\}. \tag{246}$$

We have the canonical divergence in S_n,

$$\tilde{D}_q[\boldsymbol{p} : \boldsymbol{q}] = \frac{q}{h_q(\boldsymbol{q})} D_\alpha[\boldsymbol{p} : \boldsymbol{q}]. \tag{247}$$

The Pythagorean theorem holds with respect to $\tilde{D}_q$, where the e-geodesic is linear in $\boldsymbol{\theta}$ and m-geodesic is linear in $\boldsymbol{\eta}$ defined by (245), which is different from the exponential family case of $q = 1$.

In the case of S_n, probability distributions can be represented in the following form

$$\log_q p(x, \boldsymbol{\theta}) = \sum_{i=1}^{n} \theta^i \delta_i(x) - \psi_q(\boldsymbol{\theta}), \tag{248}$$

for any q. Hence, the affine parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ due to $\tilde{D}_q$ are defined by

$$\theta^i = \frac{1}{1-q} (p_i^{1-q} - p_0^{1-q}), \tag{249}$$

$$\eta_i = \frac{1}{h_q} p_i^q. \tag{250}$$

They differ from (224) due to the α-divergence.

The canonical flat divergence $\tilde{D}_q$ is a conformal transformation of the α-divergence D_α, as shown in (247). Therefore, the two geometries derived from the $\tilde{D}_q$ divergence and D_α divergence are conformally related. We now use the α-representation instead of q. Since $\tilde{D}_q$ and its dual $\tilde{D}_q^*$ are flat, the α-geometry derived from D_α is dually conformally flat. Then, we see that the α-geometry is dually projectively flat (see [KUR94], [KUR02], [MATS98], [MATS10]). It is surprising that the Pythagorean relation holds in S_n for two divergences $\bar{D}_q$ and $\tilde{D}_q$ by using different geodesics.

7. Information geometry of Wasserstein distance

7.1. Wasserstein distance

A well-known non-invariant distance between two probability distributions is the Wasserstein distance. It has a long history of research ([VIL09], [SAN15], [PEC18]) and is still a hot topic in mathematics. Let Ω be a metric space and $p(x)$ and $q(x)$ be two probability densities on Ω. Let us consider two distributions of commodities subject to $p(x)$ and $q(x)$ on Ω. We consider the problem of transporting commodities in Ω such that the original distribution is $p(x)$ and the resultant distribution becomes $q(x)$, see Fig. 6. The distance between two points x and $y \in \Omega$ is denoted as $d(x, y)$. We assume that the cost of transporting a unit of commodity from x to y is $m(x, y)$, which is an increasing function of $d(x, y)$,

$$m(x, y) = f\{d(x, y)\}, \tag{251}$$

satisfying $f(0) = 0$, for example $f(d) = d^2$.

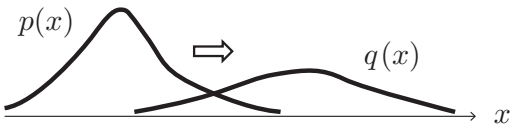

Fig. 6. Transportation of $p(x)$ to $q(x)$

Let $P(x, y)$ denote the amount of commodities transporting from x to y. This satisfies the sender and receiver conditions,

$$\int P(x, y)\, dy = p(x), \tag{252}$$

$$\int P(x, y)\, dx = q(y). \tag{253}$$

This is a stochastic matrix on $\Omega \times \Omega$ satisfying $P(x, y) \geq 0$ and $\int P(x, y)\, dx\, dy = 1$.

The total transportation cost is

$$C(P) = \langle m(x, y), P(x, y) \rangle = \int m(x, y) P(x, y)\, dx\, dy. \tag{254}$$

The minimum of C under the constraints (252) and (253) is called the Wasserstein distance between $p(x)$ and $q(x)$,

$$D_W(p, q) = \langle m(x, y), P^*(x, y) \rangle, \tag{255}$$

where $P^*(x, y)$ is the optimal solution that minimizes (254), when it exists.

The transportation problem searches for the minimizer of the linear function $\langle m, P\rangle$ under the linear constraints on P (252), (253); thus it is a LP (linear programming) problem. Therefore, the uniqueness of the optimal solution P^* is not guaranteed. The optimal solution is not necessarily continuous with respect to p and q, as one can easily see in the discrete LP case. It is generally difficult to obtain an analytical solution except for the case of $\Omega = \boldsymbol{R}^1$. The numerical computation is a burden when the size of the problem is large in the discrete case.

7.2. Entropy-regularized transportation plan

The Wasserstein distance takes the metric structure $d(x, y)$ of the underlying space Ω into account. Therefore, it is natural for certain applications such as computer vision. A visual pattern is represented by a distribution of brightness over a plane $\boldsymbol{R}^2$. Usually, $\boldsymbol{R}^2$ is discretized into n^2 pixels, and we solve the discrete LP problem for obtaining the transportation cost from $p(x)$ to $q(x)$ $x \in \boldsymbol{R}^2$. However, the computational cost is huge when n is large.

[CUT13] used the idea of modifying the problem by introducing a regularization such that the entropy of P should be larger than a constant, where the entropy is

$$H(P) = -\int P(x, y) \log P(x, y)\, dx\, dy. \tag{256}$$

Then the cost function to be minimized becomes

$$C_\lambda(P) = \langle m, P\rangle + \lambda H(P), \tag{257}$$

where λ is a Lagrange multiplier. The additional term relaxes the solution, letting the entropy become larger. Its degree is controlled by λ, and when $\lambda = 0$, the solution reduces to the original Wasserstein distance. Introduction of the $\lambda > 0$ term makes the optimal plan $P^*(p, q)$ be uniquely determined in the discrete case of the following subsections, and $P^*(p, q)$ is continuous with respect to p and q.

7.3. Manifold of optimal transportation plans

We concentrate on the discrete case $\Omega = \{0, 1, \ldots, n\}$, where distributions are patterns on Ω and belong to $\bar{S}_n$, where $\bar{S}_n$ is the closure of S_n. The problem is to transport $\boldsymbol{p} = (p_i) \in \bar{S}_n$ to $\boldsymbol{q} = (q_i) \in \bar{S}_n$, where the cost

is $M = (m_{ij})$, $m_{ii} = 0$. A transportation plan is a stochastic matrix $P = (P_{ij})$, satisfying

$$\sum_{j=0}^{n} P_{ij} = p_i, \quad i = 0, 1, \dots, n, \tag{258}$$

$$\sum_{i=0}^{n} P_{ij} = q_j, \quad j = 0, 1, \dots, n, \tag{259}$$

$$\sum_{i,j=0}^{n} P_{ij} = 1, \quad P_{ij} \geq 0. \tag{260}$$

We use the following Lagrange function

$$L = \frac{1}{1+\lambda}\langle M, P\rangle + \frac{\lambda}{1+\lambda}H(P) - \sum_{i,j=0}^{n} \alpha_i P_{ij} - \sum_{i,j=0}^{n} \beta_j P_{ij} \tag{261}$$

to be minimized, where α_i, β_j are Lagrange multipliers corresponding to constraints (258), (259). We may put $\alpha_0 = 0$, $\beta_0 = 0$. By differentiating L with respect to P_{ij}, we have

$$\frac{1+\lambda}{\lambda}\frac{\partial L}{\partial P_{ij}} = \frac{1}{\lambda}m_{ij} + \log P_{ij} - \frac{1+\lambda}{\lambda}(\alpha_i + \beta_j). \tag{262}$$

Therefore, we have the following theorem ([AKO18], [CP16]).

Theorem 13. *The optimal transportation plan sending* $\boldsymbol{p}$ *to* $\boldsymbol{q}$ *is given by*

$$P^*_{ij} = \exp\Big\{ -\frac{1}{\lambda}m_{ij} + \frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - \frac{\psi}{\lambda}\Big\}, \quad (i,j) \neq (0,0), \tag{263}$$

$$\psi = -\lambda \log P^*_{00}, \tag{264}$$

where ψ *is the normalization constant and* α_i *and* β_j *are to be determined from constraints (258), (259).*

The above theorem shows that the set of all optimal transportation plans form an exponential family ([AKO18]). By introducing random variables $x_{ij} = \delta_{ij}(x)$, where x denotes branches connecting two nodes of Ω, the optimal transportation plan is

$$P^*(x) = \exp\Big\{ \sum_{i,j}(\tilde{\alpha}_i + \tilde{\beta}_j)\delta_{ij}(x) - \tilde{\psi} - \sum_{i,j}\frac{m_{ij}}{\lambda}\delta_{ij}(x)\Big\}, \tag{265}$$

where random variables $\delta_{ij}(x)$ are

$$\delta_{ij}(x) = \begin{cases} 1, \text{ when } x = (i,j), \\ 0, \text{ otherwise,} \end{cases} \tag{266}$$

and $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ are canonical parameters,

$$\tilde{\alpha}_i = \frac{(1+\lambda)\alpha_i}{\lambda}, \quad \tilde{\beta}_j = \frac{(1+\lambda)\beta_j}{\lambda}, \quad \tilde{\alpha}_0 = \tilde{\beta}_0 = 0, \quad \tilde{\psi} = \frac{1}{\lambda}\psi. \tag{267}$$

The $\tilde{\psi}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is the potential function, and the dual parameters (expectation parameters) are $\boldsymbol{p}, \boldsymbol{q}$, because

$$\mathrm{E}\Big[\sum_j \delta_{ij}(x)\Big] = p_i, \quad \mathrm{E}\Big[\sum_i \delta_{ij}(x)\Big] = q_j. \tag{268}$$

Since $\tilde{\psi}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is a convex function, we have its dual $\tilde{\varphi}(\boldsymbol{p}, \boldsymbol{q})$, and the Legendre relations hold

$$\boldsymbol{p} = \frac{\partial}{\partial \tilde{\boldsymbol{\alpha}}}\tilde{\psi}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}), \quad \boldsymbol{q} = \frac{\partial}{\partial \tilde{\boldsymbol{\beta}}}\tilde{\psi}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}), \tag{269}$$

$$\tilde{\boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{p}}\tilde{\varphi}(\boldsymbol{p}, \boldsymbol{q}), \quad \tilde{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{q}}\tilde{\varphi}(\boldsymbol{p}, \boldsymbol{q}). \tag{270}$$

Theorem 14. *The optimal cost function $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is convex with respect to $(\boldsymbol{p}, \boldsymbol{q})$ and is the Legendre dual $\tilde{\varphi}_\lambda(\boldsymbol{p}, \boldsymbol{q})$ of the potential $\tilde{\psi}_\lambda(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$,*

$$C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \tilde{\varphi}_\lambda(\boldsymbol{p}, \boldsymbol{q}). \tag{271}$$

Proof. Since $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is rewritten as

$$C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \langle M, P^* \rangle + \lambda \sum P^*_{ij}\Big\{(\tilde{\alpha}_i + \tilde{\beta}_j) - \frac{m_{ij}}{\lambda} - \tilde{\psi}_\lambda(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})\Big\} \tag{272}$$

$$= \boldsymbol{p} \cdot \tilde{\boldsymbol{\alpha}} + \boldsymbol{q} \cdot \tilde{\boldsymbol{\beta}} - \psi_\lambda(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}), \tag{273}$$

we have

$$C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \tilde{\varphi}_\lambda(\boldsymbol{p}, \boldsymbol{q}). \tag{274}$$

This shows that $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is convex with respect to $(\boldsymbol{p}, \boldsymbol{q})$. □

7.4. Minimizer of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$, given $\boldsymbol{p}$

[CUT13] used the cost function $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ as a substitute of the transportation cost from $\boldsymbol{p}$ to $\boldsymbol{q}$ instead of the Wasserstein distance $D_W(\boldsymbol{p}, \boldsymbol{q})$. He solved various problems in vision research with remarkable success, because the entropy regularized approach is computationally tractable.

However, $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is not necessarily positive. More seriously, it is not minimized at $\boldsymbol{p} = \boldsymbol{q}$. Therefore, it is not adequate as a distance or divergence. The minimizer of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$, given $\boldsymbol{p}$, is obtained as follows ([AKOC19]).

Theorem 15. *For fixed $\boldsymbol{p}$, $C_\lambda(\boldsymbol{p},\boldsymbol{q})$ is minimized at*

$$\boldsymbol{q}^* = \tilde{K}\boldsymbol{p} = \Big(\sum_{i=0}^{n} \tilde{K}_{i|j} p_i\Big), \tag{275}$$

where $\tilde{K}$ is a linear operator defined by

$$K_{ij} = \exp\Big\{-\frac{m_{ij}}{\lambda}\Big\}, \tag{276}$$

$$\tilde{K}_{i|j} = \frac{K_{ij}}{\sum\limits_j} K_{ij} = \frac{K_{ij}}{k_i}. \tag{277}$$

Proof. Since $\boldsymbol{q}^*$ satisfies

$$\partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p},\boldsymbol{q}^*) = 0, \tag{278}$$

this implies $\tilde{\boldsymbol{\beta}} = 0$. Hence, the optimal plan from $\boldsymbol{p}$ to $\boldsymbol{q}^*$ is

$$P^*_{ij} = \exp\Big\{-\frac{m_{ij}}{\lambda} + \tilde{\alpha}_i - \tilde{\psi}\Big\} = K_{ij}\exp\{\tilde{\alpha}_i - \tilde{\psi}\}. \tag{279}$$

From

$$\sum_j P^*_{ij} = p_i, \tag{280}$$

we have

$$\exp\{\tilde{\alpha}_i - \tilde{\psi}\}k_i = p_i. \tag{281}$$

Hence,

$$P^*_{ij} = \frac{p_i}{k_i} K_{ij}. \tag{282}$$

From

$$\sum_i P^*_{ij} = q^*_j, \tag{283}$$

$$\boldsymbol{q}^* = \tilde{K}\boldsymbol{p} \tag{284}$$

is proved. □

7.5. New divergences introduced in S_n

It is possible to modify $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ to give a true divergence. We show two such divergences. One was derived by [AKOC19],

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = C_\lambda(\boldsymbol{p}, \tilde{K}\boldsymbol{q}) - C_\lambda(\boldsymbol{p}, \tilde{K}\boldsymbol{p}). \tag{285}$$

Since $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is a convex function because of Theorem 14, its modification $D_\lambda[\boldsymbol{p}, \boldsymbol{q}]$ is a convex function with respect to $(\boldsymbol{p}, \boldsymbol{q})$. This is a divergence but is not dually flat. The Riemannian metric is given by

$$g_{ij} = \partial_{ij} D_\lambda[\boldsymbol{p} : \boldsymbol{q}]_{\boldsymbol{q}=\boldsymbol{p}}. \tag{286}$$

The other divergence is simply given by

$$\tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}] = C_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \frac{1}{2}\{C_\lambda(\boldsymbol{p}, \boldsymbol{p}) + C_\lambda(\boldsymbol{q}, \boldsymbol{q})\}. \tag{287}$$

This was used in [GPC18] and [RTC17] without any justification. We need to show that this is convex with respect to $(\boldsymbol{p}, \boldsymbol{q})$.

Conjecture. When $m_{ij} = m_{ji}$, $\tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ is a divergence in S_n.

It is easy to see that

$$\tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{p}] = 0 \tag{288}$$

for any $\boldsymbol{p}$. We calculate the derivative of $\tilde{D}_\lambda$ with respect to $\boldsymbol{q}$,

$$\partial_{\boldsymbol{q}} \tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \frac{1}{2}\{\partial_{\boldsymbol{p}} C_\lambda(\boldsymbol{q}, \boldsymbol{q}) + \partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{q}, \boldsymbol{q})\}, \tag{289}$$

where $\partial_{\boldsymbol{p}}$ and $\partial_{\boldsymbol{q}}$ denote differentiation in $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ with respect to $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively, and we have

$$\partial_{\boldsymbol{q}} \tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}]_{\boldsymbol{q}=\boldsymbol{p}} = \frac{1}{2}\{\partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \boldsymbol{q})\}_{\boldsymbol{q}=\boldsymbol{p}} = 0, \tag{290}$$

because $C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = C_\lambda(\boldsymbol{q}, \boldsymbol{p})$, provided $m_{ij} = m_{ji}$. Hence, for fixed $\boldsymbol{p}$, $\boldsymbol{q} = \boldsymbol{p}$ is a critical point of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$. We need to prove that the second derivative $\partial_{\boldsymbol{qq}} \tilde{D}_\lambda[\boldsymbol{p}, \boldsymbol{q}]$ is positive-definite at $\boldsymbol{q} = \boldsymbol{p}$. See [FEY19], where this is proved in the case of Wasserstein divergence regularized by the Shannon mutual information.

7.6. Barycenter of patterns: shape-location separation theorem

We give an example in computational vision, showing the superiority of the Wasserstein distance to invariant divergences such as the KL-divergence. We consider a set of patterns $S = \{p_1(\boldsymbol{\xi}), \dots, p_n(\boldsymbol{\xi})\}$, where $\boldsymbol{\xi} \in \boldsymbol{R}^2$. A pattern $p(\boldsymbol{\xi})$ is a distribution over $\boldsymbol{R}^2$. For a divergence D, the D-barycenter of S is the pattern $q_D^*(\boldsymbol{\xi})$ that minimizes

$$F_S(q) = \sum_{i=1}^{n} D[p_i : q], \tag{291}$$

$$q_D^* = \arg\min F_S(q). \tag{292}$$

The D_λ-barycenter $q^*(\boldsymbol{\xi})$ captures a common shape of S as shown in the following shape-location separation theorem ([AKOC19]). The theorem holds for C_λ-, $\tilde{D}_\lambda$-, and D_λ-barycenters but does not hold for invariant divergences such as KL-divergence or Hellinger divergence.

We define the center $\boldsymbol{\xi}_p$ of pattern $p(\boldsymbol{\xi})$ by

$$\boldsymbol{\xi}_p = \int \boldsymbol{\xi} p(\boldsymbol{\xi})\, d\boldsymbol{\xi}. \tag{293}$$

A shift of pattern $p(\boldsymbol{\xi})$ by $\bar{\boldsymbol{\xi}}$ is

$$T_{\bar{\boldsymbol{\xi}}} p(\boldsymbol{\xi}) = p(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}). \tag{294}$$

We shift all $p_1(\boldsymbol{\xi}), \dots, p_n(\boldsymbol{\xi})$ such that their centers become $\boldsymbol{\xi} = 0$,

$$\bar{p}_i(\boldsymbol{\xi}) = p_i(\boldsymbol{\xi} - \boldsymbol{\xi}_{p_i}), \quad i = 1, \dots, n. \tag{295}$$

All $\bar{p}_1(\boldsymbol{\xi}), \dots, \bar{p}_n(\boldsymbol{\xi})$ are located at $\boldsymbol{\xi} = 0$, that is, their centers are 0, without changing the shapes.

Theorem 16. *The C_λ-barycenter $q_{C_\lambda}^*$ are located at the barycenter of the centers $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ of patterns $p_1(\boldsymbol{\xi}), \dots, p_n(\boldsymbol{\xi})$ and its shape is congruent to the shape of the barycenter of the co-located $\bar{p}_1(\boldsymbol{\xi}), \dots, \bar{p}_n(\boldsymbol{\xi})$. This holds for D_λ- and $\tilde{D}_\lambda$-barycenters.*

Proof. Let $P_{p,q}^*(\boldsymbol{\xi}, \boldsymbol{\xi}')$ be the optimal transportation plan from $p(\boldsymbol{\xi})$ to $q(\boldsymbol{\xi})$, and let

$$\bar{\boldsymbol{\xi}} = \boldsymbol{\xi}_q - \boldsymbol{\xi}_p. \tag{296}$$

Then $\bar{q}(\boldsymbol{\xi}) = T_{\bar{\boldsymbol{\xi}}} q(\boldsymbol{\xi})$ and $p(\boldsymbol{\xi})$ have the same center. Let the optimal plan sending $p(\boldsymbol{\xi})$ to $\bar{q}(\boldsymbol{\xi})$ be $P_{p,\bar{q}}^*$. Then, we easily have

$$C_\lambda(P_{p,q}^*) = C_\lambda(P_{p,\bar{q}}^*) + |\boldsymbol{\xi}_p - \boldsymbol{\xi}_q|^2, \tag{297}$$

since the entropy of P does not change by shift of patterns. This shows that the transportation cost $C_\lambda(P^*_{p,q})$ is decomposed into a sum of the costs due to their location difference and shape difference. Hence

$$\sum C_\lambda(P^*_{p_i,q}) = \sum C_\lambda(P^*_{p_i,\bar{q}}) + \sum |\boldsymbol{\xi}_{p_i} - \boldsymbol{\xi}_q|^2, \tag{298}$$

which proves the theorem. The same discussion holds for D_λ- and $\tilde{D}_\lambda$-barycenters. □

[CUT13] demonstrated that the C_λ-barycenter extracts a common shape from patterns. We confirmed this by proving the shape-location separation theorem for the C_λ-barycenter. However, since $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is not a divergence, there is a problem. Let $p_1(\boldsymbol{\xi}), \ldots, p_n(\boldsymbol{\xi})$ be shifted patterns of $p(\boldsymbol{\xi})$. Then, their C_λ-barycenter is the minimizer of $C_\lambda(p, q)$, that is

$$\boldsymbol{q}^*_{C_\lambda} = \tilde{K}\boldsymbol{p}, \tag{299}$$

which is a blurred pattern of $\boldsymbol{p}$. However, the D_λ- and $\tilde{D}_\lambda$-barycenters are the same as the original pattern

$$\boldsymbol{q}^*_{D_\lambda} = \boldsymbol{q}^*_{\tilde{D}_\lambda} = \boldsymbol{p}. \tag{300}$$

7.7. Information matrix connecting the Fisher metric and Wasserstein metric

Recently, [LIZ19] proposed an interesting idea of connecting the Fisher information metric with the Wasserstein Riemannian metric by using a one-parameter operator. We shortly introduce their idea. Let us consider a one-dimensional base space $\Omega = \boldsymbol{R}^1$. Let $M = \{p(x, \boldsymbol{\theta})\}$ be a regular statistical model parameterized by $\boldsymbol{\theta}$. Moreover, let g be a metric tensor over $\boldsymbol{R}^1$. We define the Riemannian metric $G(\boldsymbol{\theta})$ induced in M by the pullback of g, as

$$G(\boldsymbol{\theta}) = \int \frac{\partial}{\partial \boldsymbol{\theta}} p(x, \boldsymbol{\theta}) g\{p(x, \boldsymbol{\theta})\} \frac{\partial}{\partial \boldsymbol{\theta}} p(x, \boldsymbol{\theta})\, dx. \tag{301}$$

It is given in the components form as

$$G_{ij}(\boldsymbol{\theta}) = \int \partial_i p(x, \boldsymbol{\theta}) \{g(p(x, \boldsymbol{\theta})) \partial_j p(x, \boldsymbol{\theta})\}\, dx, \tag{302}$$

where

$$\partial_i = \frac{\partial}{\partial \theta^i}. \tag{303}$$

When g is a simple scalar function,

$$g(p) = \frac{1}{p(x, \boldsymbol{\theta})}, \tag{304}$$

the Riemannian metric is

$$G_{ij} = \int \frac{1}{p(x, \boldsymbol{\theta})} \partial_i p(x, \boldsymbol{\theta}) \partial_j p(x, \boldsymbol{\theta}) \, dx, \tag{305}$$

which is equal to the Fisher information matrix,

We define the Wasserstein metric tensor by using the inverse of differential operator as

$$g_W(p) = (-\Delta_p)^{-1}, \tag{306}$$

where the Laplacian Δ_p is defined by

$$\Delta_p = \nabla \cdot p \nabla, \tag{307}$$

$$\nabla = \frac{d}{dx}. \tag{308}$$

The induced metric is the Wasserstein information matrix

$$G_{ij}^W(\boldsymbol{\theta}) = -\int \partial_i p(x, \boldsymbol{\theta}) \Delta_p^{-1} \partial_j p(x, \boldsymbol{\theta}) \, dx. \tag{309}$$

Let $P(x, \boldsymbol{\theta})$ be the cumulative distribution of $p(x, \boldsymbol{\theta})$,

$$P(x, \boldsymbol{\theta}) = \int_{-\infty}^{x} p(u, \boldsymbol{\theta}) \, du. \tag{310}$$

Then, the Wasserstein information matrix is explicitly given by

$$G_{ij}^W(\boldsymbol{\theta}) = \int \frac{1}{p(x, \boldsymbol{\theta})} \partial_i P(x, \boldsymbol{\theta}) \partial_j P(x, \boldsymbol{\theta}) \, dx. \tag{311}$$

It is interesting to explore statistical inference based on the Wasserstein distance. See [AMM20] for the recent developments by another approach.

Conclusions and future perspectives

I discussed information geometry, which emerged from the invariant properties of a manifold of probability distributions. It gives a Riemannian geometry equipped with a third-order symmetric tensor T, from which a dual pair of affine connections are introduced. This is a natural geometry derived from an asymmetric divergence function of two points in a manifold and related to affine differential geometry. A dually flat statistical manifold has good properties such as existence of a canonical divergence,

the generalized Pythagorean theorem and projection theorem. They are particularly useful for applications in statistics, signal processing, game theory, computer vision, artificial intelligence, etc., although we do not touch upon applications in this paper, except for the semi-parametric estimation and the Wasserstein problem.

Geometry having dual affine connections has not yet been fully explored. There are lots of problems to be studied in future from the mathematical point of view. For example, a diffusion process or random walk in a manifold of dual affine connections is an interesting topic, where dually coupled Laplacians Δ and Δ^* and diffusion flows exist. It is interesting to know the role of duality in this setting. Another example is dual Ricci flows. It will be interesting to extend the various results in Riemannian geometry to these problems in the dual setting.

We studied the information geometry of the entropy-regularized Wasserstein problem, obtaining a new Wasserstein-motivated divergence in a manifold of probability distributions. It is also an interesting mathematical topic to explore the geometry derived from the divergence of the entropy-regularized Wasserstein problem.

I cannot touch two important subjects on information geometry. One is the geometry of a function space of probability distributions initiated by [PIS95]. See also [AY17]. The other is quantum information geometry which studies the geometry of quantum states ([AMN00], [HAY17]).

References

[AM85] S. Amari, Differential-Geometrical Methods in Statistics, Lect. Notes Stat., **28**, Springer-Verlag, 1985.

[AM00] S. Amari, Estimating functions of independent component analysis for temporally correlated signals, Neural Computation, **12** (2000), 2083–2107.

[AM16] S. Amari, Information Geometry and Its Applications, Appl. Math. Sci., **194**, Springer-Verlag, 2016.

[AMC97] S. Amari and J.-F. Cardoso, Blind source separation—Semiparametric statistical approach, IEEE Trans. Signal Process., **45** (1997), 2692–2700.

[AKO18] S. Amari, R. Karakida and M. Oizumi, Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem, Inf. Geom., **1** (2018), 13–37.

[AKOC19] S. Amari, R. Karakida, M. Oizumi and M. Cuturi, Information geometry for regularized optimal transport and barycenters of patterns, Neural Comput., **31** (2019), 827–848.

[AMK97] S. Amari and M. Kawanabe, Information geometry of estimating functions in semi-parametric statistical models, Bernoulli, **3** (1997), 29–54.

[AMM20] S. Amari and T. Matsuda, Wasserstein statistics in one-dimensional location-scale model, preprint, arXiv:2007.11401.

[AMN00] S. Amari and H. Nagaoka, Methods of Information Geometry, Transl. Math. Monogr., **191**, Amer. Math. Soc., Providence, RI; Oxford Univ. Press, 2000.

[AOM12] S. Amari, A. Ohara and H. Matsuzoe, Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries, Phys. A, **391** (2012), 4308–4319.

[AY17] N. Ay, J. Jost, H.V. Lê and L. Schwachhöfer, Information Geometry, Ergeb. Math. Grenzgeb. (3), **64**, Springer-Verlag, 2017.

[BAN05] A. Banerjee, S. Merugu, I.S. Dhillon and J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res., **6** (2005), 1705–1749.

[BAU16] M. Bauer, M. Bruveris and P.W. Michor, Uniqueness of the Fisher–Rao metric on the space of smooth densities, Bull. Lond. Math. Soc., **48** (2016), 499–506.

[BRE67] L.M. Brègman, The relaxation method of finding a common point of convex sets and its applications to the solution of problems in convex programming, U.S.S.R. Comput. Math. and Math. Phys., **7** (1967), 200–217.

[CEP07] A. Cena and G. Pistone, Exponential statistical manifold, Ann. Inst. Statist. Math., **59** (2007), 27–56.

[CHEN72] N.N. Chentsov, Statistical Decision Rules and Optimal Inference, Transl. Math. Monogr., **53**, Amer. Math. Soc., Providence, RI, 1982; Originally published in Russian, Nauka, 1972.

[CSI67] I. Csiszár, Information-type measures of difference of probability distributions and indirect observation, Studia Sci. Math. Hungar., **2** (1967), 299–318.

[CUT13] M. Cuturi, Sinkhorn distance: Lightspeed computation of optimal transport, Advances in Neural Information Processing Systems, **26** (2013), 2292–2300.

[CP16] M. Cuturi and G. Peyré, A smoothed dual approach for variational Wasserstein problems, SIAM J. Imaging Sci., **9** (2016), 320–343.

[DOW18] J.G. Dowty, Chentsov's theorem for exponential families, Inf. Geom., **1** (2018), 117–135.

[EG83] S. Eguchi, Second order efficiency of minimum contrast estimators in a curved exponential family, Ann. Statist., **11** (1983), 793–803.

[FEY19] J. Feydy, T. Séjourné, F.-X. Vialard, S. Amari, A. Trouvé and G. Peyré, Interpolating between optimal transport and MMD using Sinkhorn divergences, In: The 22nd International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. (PMLR), **89**, PMLR, 2019, pp. 2681–2690.

[FUJ15] A. Fujiwara, Foundations of Information Geometry, Makino Shoten, 2015.

[GPC18] A. Genevay, G. Peyré and M. Cuturi, Learning generative models with Shinkhorn divergences, In: International Conference on Artificial Intelligence, and Statistics, Proc. Mach. Learn. Res. (PMLR), **84**, PMLR, 2018, pp. 1608–1617.

[HAY17] M. Hayashi, Quantum Information Theory: Mathematical Foundation. 2nd ed., Grad. Texts Phys., Springer-Verlag, 2017.

[KUR94] T. Kurose, On the divergences of 1-conformally flat statistical manifolds, Tohoku, Math. J., **46** (1994), 427–433.

[KUR99] T. Kurose, Dual connections and projective geometry, Fukuoka Univ. Sci. Rep., **29** (1999), 221–224.

[KUR02] T. Kurose, Conformal-projective geometry of statistical manifolds, Interdiscip. Inform. Sci., **8** (2002), 89–100.

[LAU87] S.L. Lauritzen, Statistical manifolds, In: Differential Geometry in Statistical Inference, Institute of Mathematical Statistics, Lecture Notes Monograph Series, **10**, Institute of Mathematical Statistics, 1987, pp. 23–33.

[LE05] H.V. Lê, Statistical manifolds are statistical models, J. Geom., **84** (2005), 83–93.

[LIZ19] W. Li and J. Zhao, Wasserstein information matrix, preprint, arXiv:1910.11248.

[MATS98] H. Matsuzoe, On realization of conformally-projectively flat statistical manifolds and the divergences, Hokkaido Math. J., **27** (1998), 409–421.

[MATS99] H. Matsuzoe, Geometry of contrast functions and conformal geometry, Hiroshima Math. J., **29** (1999), 175–191.

[MATS10] H. Matsuzoe, Statistical manifolds and affine differential geometry, In: Probabilistic Approach to Geometry, Adv. Stud. Pure Math., **57**, Math. Soc. Japan, Tokyo, 2010, pp. 303–321.

[MAT93] T. Matumoto, Any statistical minifold has a contrast function—On the C^3-functions taking the minimum at the diagonal of the product manifold, Hiroshima Math. J., **23** (1993), 327–332.

[MOA06] K. Miura, M. Okada and S. Amari, Estimating spiking irregularities under changing environments, Neural Comput., **18** (2006), 2359–2386.

[MOR63] T. Morimoto, Markov processes and the H-theorem, J. Phys. Soc. Japan, **18** (1963), 328–331.

[NAU11] J. Naudts, Generalised Thermostatistics, Springer-Verlag, 2011.

[NOS94] K. Nomizu and T. Sasaki, Affine Differential Geometry, Cambridge Tracts in Math., **111**, Cambridge Univ. Press, Cambridge, 1994.

[PEC18] G. Peyré and M. Cuturi, Computational optimal transport, preprint, arXiv:1803.00567.

[PIS95] G. Pistone and C. Sempi, An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, Ann. Statist., **23** (1995), 1543–1561.

[RAO45] C. Radhakrishna Rao, Information and accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc., **37** (1945), 81–91.

[RTC17] A. Ramdas, N. García Trillos and M. Cuturi, On Wasserstein two-sample testing and related families of nonparametric tests, Entropy, **19** (2017), no. 47.

[SAN15] F. Santambrogio, Optimal Transport for Applied Mathematicians, Progr. Nonlinear Differential Equations Appl., **87**, Birkhäuser, 2015.

[SH07] H. Shima, The Geometry of Hessian Structures, World Sci. Publ., 2007.

[TSA09] C. Tsallis, Introduction to Nonextensive Statistical Mechanics. Approaching a Complex World, Springer-Verlag, 2009.

[VIL09] C. Villani, Optimal Transport. Old and New, Grundlehren Math. Wiss., **338**, Springer-Verlag, 2009.

[WON18] T.-K.L. Wong, Logarithmic divergences from optimal transport and Rényi geometry, Inf. Geom., **1** (2018), 39–78.