Chapter 3
Hessian Structures and Divergence Functions on Deformed Exponential Families

Hiroshi Matsuzoe and Masayuki Henmi

Abstract A Hessian structure (∇, h) on a manifold is a pair of a flat affine connection ∇ and a semi-Riemannian metric h which is given by a Hessian of some function. In information geometry, it is known that an exponential family naturally has dualistic Hessian structures and their canonical divergences coincide with the Kullback-Leibler divergences, which are also called the relative entropies. A deformed exponential family is a generalization of exponential families. A deformed exponential family naturally has two kinds of dualistic Hessian structures and conformal structures of Hessian metrics. In this paper, geometry of such Hessian structures and conformal structures are summarized. In addition, divergence functions on these Hessian manifolds are constructed from the viewpoint of estimating functions. As an application of such Hessian structures to statistics, a generalization of independence and geometry of generalized maximum likelihood method are studied.

Keywords Hessian manifold · Statistical manifold · Deformed exponential family · Divergence · Information geometry · Tsallis statistics

3.1 Introduction

In information geometry, an exponential family is a useful statistical model and it is applied to various fields of statistical sciences (cf. [1]). For example, the set of Gaussian distributions is an exponential family. It is known that an exponential family can be naturally regarded as a Hessian manifold [28], which is also called a dually flat

H. Matsuzoe (✉)
Department of Computer Science and Engineering, Graduate School of Engineering,
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
e-mail: matsuzoe@nitech.ac.jp

M. Henmi
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: henmi@ism.ac.jp

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_3,

space [1] or a flat statistical manifold [12]. A pair of dually flat affine connections has essential roles in geometric theory of statistical inferences. In addition, a Hessian manifold has an asymmetric squared-distance like function, called the canonical divergence. On an exponential family, the canonical divergence coincides with the Kullback-Leibler divergence or the relative entropy. (See Sect. 3.3.)

A deformed exponential family is a generalization of exponential families, which was introduced in anomalous statistical physics [22]. (See also [23, 32] and [33].) A deformed exponential family naturally has two kinds of dualistic Hessian structures, and such geometric structures are independently studied in machine learning theory [21] and statistical physics [3, 26], etc. For example, a q-exponential family is a typical example of deformed exponential families. One of Hessian structures on a q-exponential family is related to geometry of β-divergences (or density power divergences [5]). The other Hessian structure is related to geometry of α-divergences. (In the q-exponential case, these geometry are studied in [18].) In addition, conformal structures of statistical manifolds play important roles in geometry of deformed exponential families.

In this paper, we summarize such Hessian structures and conformal structures on deformed exponential families. Then we construct a generalized relative entropy from the viewpoint of estimating functions. As an application, we consider generalization of independence of random variables, then elucidate geometry of the maximum q-likelihood estimator. This paper is written based on the proceeding [19].

3.2 Preliminaries

In this paper, we assume that all objects are smooth, and a manifold M is an open domain in $\boldsymbol{R}^n$.

Let (M, h) be a semi-Riemannian manifold, that is, h is assumed to be nondegenerate, which is not necessary to be positive definite (e.g. the Lorentzian metric in relativity). Let ∇ be an affine connection on M. We define the *dual connection* ∇^* of ∇ with respect to h by

$$Xh(Y, Z) = h(\nabla_X Y, Z) + h(Y, \nabla^*_X Z),$$

where X, Y and Z are arbitrary vector fields on M. It is easy to check that $(\nabla^*)^* = \nabla$.

For an affine connection ∇, we define the *curvature tensor field* R and the *torsion tensor field* T by

$$\begin{aligned} R(X, Y)Z &:= \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \\ T(X, Y) &:= \nabla_X Y - \nabla_Y X - [X, Y], \end{aligned}$$

where $[X, Y] := XY - YX$. We say that ∇ is *curvature-free* if R vanishes everywhere on M, and the one is *torsion-free* if T vanishes everywhere.

For pair of dual affine connections, the following proposition holds (cf. [16]).

Proposition 1 *Consider the conditions below:*

1. ∇ *is torsion-free.*
2. ∇^* *is torsion-free.*
3. $\nabla^{(0)} = (\nabla + \nabla^*)/2$ *is the Levi-Civita connection with respect to h.*
4. ∇h *is totally symmetric, where* ∇h *is the* $(0, 3)$*-tensor field defined by*

$$(\nabla_X h)(Y, Z) := Xh(Y, Z) - h(\nabla_X Y, Z) - h(Y, \nabla_X Z).$$

Assume any two of the above conditions, then the others hold.

From now on, we assume that an affine connection ∇ is torsion-free.

We say that an affine connection ∇ is *flat* if ∇ is curvature-free. For a flat affine connection ∇, there exists a coordinate system $\{\theta^i\}$ on M locally such that the connection coefficients $\{\Gamma_{ij}^{\nabla\ k}\}$ $(i, j, k = 1, \ldots, n)$ of ∇ vanish on its coordinate neighbourhood. We call such a coordinate system $\{\theta^i\}$ an *affine coordinate system*.

Let (M, h) be a semi-Riemannian manifold, and let ∇ be a flat affine connection on M. We say that the pair (∇, h) is a *Hessian structure* on M if there exists a function ψ, at least locally, such that $h = \nabla d\psi$ [28]. In the coordinate form, the following formula holds:

$$h_{ij}(p(\theta)) = \frac{\partial^2}{\partial\theta^i \partial\theta^j}\psi(p(\theta)),$$

where p is an arbitrary point in M and $\{\theta^i\}$ is a ∇-affine coordinate system around p. Under the same assumption, we call the triplet (M, ∇, h) a *Hessian manifold*. For a Hessian manifold (M, ∇, h), we define a totally symmetric $(0, 3)$-tensor field C by $C := \nabla h$. We call C the *cubic form* for (M, ∇, h).

For a semi-Riemannian manifold (M, h) with a torsion-free affine connection ∇, the triplet (M, ∇, h) is said to be a *statistical manifold* if ∇h is totally symmetric [12]. Originally, the triplet (M, g, C) is called a statistical manifold [14], where (M, g) is a Riemannian manifold and C is a totally symmetric $(0, 3)$-tensor field on M. From Proposition 1, these definitions are essentially equivalent. In fact, for a semi-Riemannian manifold (M, h) with a totally symmetric $(0, 3)$-tensor field C, we can define mutually dual torsion-free affine connections ∇ and ∇^* by

$$h(\nabla_X Y, Z) := h(\nabla_X^{(0)} Y, Z) - \frac{1}{2}C(X, Y, Z), \tag{3.1}$$

$$h(\nabla_X^* Y, Z) := h(\nabla_X^{(0)} Y, Z) + \frac{1}{2}C(X, Y, Z), \tag{3.2}$$

where $\nabla^{(0)}$ is the Levi-Civita connection with respect to h. In this case, ∇h and $\nabla^* h$ are totally symmetric. Hence (M, ∇, h) and (M, ∇^*, h) are statistical manifolds.

A triplet (M, ∇, h) is a flat statistical manifold if and only if it is a Hessian manifold (cf. [28]). Suppose that R and R^* are curvature tensors of ∇ and ∇^*, respectively. Then we have

$$h(R(X, Y)Z, V) = -h(Z, R^*(X, Y)V).$$

Hence the condition that the triplet (M, ∇, h) is a Hessian manifold is equivalent to that the quadruplet (M, h, ∇, ∇^*) is a *dually flat space* [1].

For a Hessian manifold (M, ∇, h), we suppose that $\{\theta^i\}$ is a ∇-affine coordinate system on M. Then there exists a ∇^*-affine coordinate system $\{\eta_i\}$ such that

$$h\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\eta_j}\right) = \delta^i_j.$$

We call $\{\eta_i\}$ the *dual coordinate system* of $\{\theta^i\}$ with respect to h.

Proposition 2 *Let (M, ∇, h) be a Hessian manifold. Suppose that $\{\theta^i\}$ is a ∇-affine coordinate system, and $\{\eta_i\}$ is the dual coordinate system of $\{\theta^i\}$. Then there exist functions ψ and ϕ on M such that*

$$\frac{\partial\psi}{\partial\theta^i} = \eta_i, \quad \frac{\partial\phi}{\partial\eta_i} = \theta^i, \quad \psi(p) + \phi(p) - \sum_{i=1}^{n} \theta^i(p)\eta_i(p) = 0, \quad (p \in M), \tag{3.3}$$

$$h_{ij} = \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j}, \quad h^{ij} = \frac{\partial^2\phi}{\partial\eta_i\partial\eta_j},$$

where (h_{ij}) is the component matrix of a semi-Riemannian metric h with respect to $\{\theta^i\}$, and (h^{ij}) is the inverse matrix of (h_{ij}). Moreover,

$$C_{ijk} = \frac{\partial^3\psi}{\partial\theta^i\partial\theta^j\partial\theta^k} \tag{3.4}$$

is the cubic form of (M, ∇, h).

For proof, see [1] and [28]. The functions ψ and ϕ are called the *θ-potential* and the *η-potential*, respectively. From the above proposition, the Hessians of θ-potential and η-potential coincide with the semi-Riemannian metric h:

$$\frac{\partial\eta_i}{\partial\theta^j} = \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j} = h_{ij}, \qquad \frac{\partial\theta^i}{\partial\eta_j} = \frac{\partial^2\phi}{\partial\eta_i\partial\eta_j} = h^{ij}. \tag{3.5}$$

In addition, we obtain the original flat connection ∇ and its dual ∇^* from the potential function ψ. From Eq. (3.4), we have the cubic form of Hessian manifold (M, ∇, h). Then we obtain two affine connections ∇ and ∇^* by Eqs. (3.1), (3.2) and (3.4).

Under the same assumptions as in Proposition 2, we define a function D on $M \times M$ by

$$D(p, r) := \psi(p) + \phi(r) - \sum_{i=1}^{n} \theta^i(p)\eta_i(r), \quad (p, r \in M).$$

We call D the *canonical divergence* of (M, ∇, h). The definition is independent of choice of an affine coordinate system. The canonical divergence is an asymmetric squared distance like function on M. In particular, the canonical divergence D is non-negative if the metric h is positive definite. However, we assumed that h is a semi-Riemannian metric, hence D can take negative values. (cf. [12] and [15].)

We remark that the canonical divergence induces the original Hessian manifold (M, ∇, h) by Eguchi's relation [7]. Suppose that D is a function on $M \times M$. We define a function on M by the following formula:

$$D[X_1, \ldots, X_i | Y_1, \ldots, Y_j](p) := (X_1)_p \cdots (X_i)_p (Y_1)_r \cdots (Y_j)_r D(p, r)|_{p=r},$$

where $X_1, \ldots, X_i$ and $Y_1, \cdots, Y_j$ are vector fields on M. We say that D is a *contrast function* on $M \times M$ if

$$\begin{aligned} &1.\ \ D[\ |\ ](p) = D(p, p) = 0, \\ &2.\ \ D[X|\ ](p) = D[\ |X](p) = 0, \\ &3.\ \ h(X, Y) := -D[X|Y] \\ &\qquad \text{is a semi-Riemannian metric on } M. \end{aligned} \tag{3.6}$$

For a contrast function D on $M \times M$, we define a pair of affine connections by

$$\begin{aligned} h(\nabla_X Y, Z) &= -D[XY|Z], \\ h(Y, \nabla^*_X Z) &= -D[Y|XZ]. \end{aligned}$$

By differentiating Eq. (3.6), two affine connections ∇ and ∇^* are mutually dual with respect to h. We can check that ∇ and ∇^* are torsion-free, and ∇h and $\nabla^* h$ are totally symmetric. Hence triplets (M, ∇, h) and (M, ∇^*, h) are statistical manifolds. We call (M, ∇, h) the *induced statistical manifold* from a contrast function D. If (M, ∇, h) is a Hessian manifold, we say that (M, ∇, h) is the *induced Hessian manifold* from D.

Proposition 3 *Suppose that D is the canonical divergence on a Hessian manifold (M, ∇, h). Then D is a contrast function on $M \times M$ which induces the original Hessian manifold (M, ∇, h).*

Proof From the definition and Eq. (3.3), we have $D[\ |\ ] = 0$ and $D[X|\ ] = D[\ |X] = 0$. Let $\{\theta^i\}$ be a ∇-affine coordinate and $\{\eta_j\}$ the dual affine coordinate of $\{\theta^j\}$. Set $\partial_i = \partial/\partial\theta^i$. From Eqs. (3.3) and (3.5), we have

$$\begin{aligned} D[\partial_i|\partial_j](p) = (\partial_i)_p(\partial_j)_r D(p, q)|_{p=r} \ &= \ (\partial_j)_r \left\{\eta_i(p) - \eta_i(r)\right\}|_{p=r} \\ = -(\partial_j)_r \eta_i(r)|_{p=r} \ &= \ -h_{ij}(p). \end{aligned}$$

This implies that the canonical divergence D is a contrast function on $M \times M$. Induced affine connections are given by

$$\begin{aligned} \Gamma_{ij,k} = -D[\partial_i\partial_j|\partial_k] \ &= \ (\partial_i)_p(\partial_k)_r \left\{\eta_j(p) - \eta_j(r)\right\}|_{p=r} \\ = -(\partial_i)_p(\partial_k)_r \eta_j(r)|_{p=r} \ &= \ 0, \end{aligned}$$

$$\begin{aligned}\Gamma^*_{ik,j} &= -D[\partial_j|\partial_i\partial_k] \;=\; (\partial_i)_r(\partial_k)_r\left\{\eta_j(p)-\eta_j(r)\right\}|_{p=r}\\ &= -(\partial_i)_r(\partial_k)_r\eta_j(r)|_{p=r} \;=\; -(\partial_i)_r(\partial_k)_r(\partial_j)_r\psi(r)|_{p=r}\\ &= C_{ikj},\end{aligned}$$

where $\Gamma_{ij,k}$ and $\Gamma^*_{ik,j}$ are Christoffel symbols of the first kind of ∇ and ∇^*, respectively. From Eqs. (3.1) and (3.2), since h is nondegenerate, the affine connection ∇ coincides with the original one of (M, ∇, h). □

At the end of this section, we review generalized conformal equivalence for statistical manifolds. Fix a number $\alpha \in \boldsymbol{R}$. We say that two statistical manifolds (M, ∇, h) and $(M, \bar{\nabla}, \bar{h})$ are *α-conformally equivalent* if there exists a function φ on M such that

$$\begin{aligned}\bar{h}(X,Y) &= e^{\varphi}h(X,Y),\\ \bar{\nabla}_X Y &= \nabla_X Y - \frac{1+\alpha}{2}h(X,Y)\mathrm{grad}_h\varphi + \frac{1-\alpha}{2}\left\{d\varphi(Y)\,X + d\varphi(X)\,Y\right\},\end{aligned}$$

where $\mathrm{grad}_h\varphi$ is the gradient vector field of φ with respect to h, that is,

$$h(\mathrm{grad}_h\varphi, X) := X\varphi.$$

(The vector field $\mathrm{grad}_h\varphi$ is often called the *natural gradient* of φ in neurosciences, etc.) We say that a statistical manifold (M, ∇, h) is *α-conformally flat* if it is locally α-conformally equivalent to some Hessian manifold [12].

Suppose that D and $\bar{D}$ are contrast functions on $M \times M$. We say that D and $\bar{D}$ are *α-conformally equivalent* if there exists a function φ on M such that

$$\bar{D}(p,r) = \exp\left[\frac{1+\alpha}{2}\varphi(p)\right]\exp\left[\frac{1-\alpha}{2}\varphi(r)\right]D(p,r).$$

In this case, induced statistical manifolds (M, ∇, h) and $(M, \bar{\nabla}, \bar{h})$ from D and $\bar{D}$, respectively, are α-conformally equivalent.

Historically, conformal equivalence of statistical manifolds was introduced in asymptotic theory of sequential estimation [27]. (See also [11].) Then it is generalized in affine differential geometry (e.g. [10, 12, 13] and [17]). As we will see in Sects. 3.5 and 3.6, conformal structures on a deformed exponential family play important roles. (See also [2, 20, 24] and [25].)

3.3 Statistical Models

Let $(\Omega, \mathcal{F}, P)$ be a probability space, that is, Ω is a sample space, $\mathcal{F}$ is a completely additive class on Ω, and P is a probability measure on Ω. Let Ξ be an open subset in $\boldsymbol{R}^n$. We say that S is a *statistical model* if S is a set of probability density functions on Ω with parameter $\xi = {}^t(\xi^1, \ldots, \xi^n) \in \Xi$ such that

$$S := \left\{ p(x;\xi) \,\middle|\, \int_{\Omega} p(x;\xi)dx = 1,\ p(x;\xi) > 0,\ \xi \in \Xi \subset \boldsymbol{R}^n \right\}.$$

Under suitable conditions, S can be regarded as a manifold with local coordinate system $\{\xi^i\}$ [1]. In particular, we assume that we can interchange differentials and integrals. Hence, the equation below holds

$$\int_{\Omega} \left(\frac{\partial}{\partial \xi^i} p(x;\xi)\right) dx = \frac{\partial}{\partial \xi^i} \int_{\Omega} p(x;\xi)dx = \frac{\partial}{\partial \xi^i} 1 = 0.$$

For a statistical model S, we define the Fisher information matrix $g^F(\xi) = (g^F_{ij}(\xi))$ by

$$\begin{aligned} g^F_{ij}(\xi) &:= \int_{\Omega} \left(\frac{\partial}{\partial \xi^i} \log p(x;\xi)\right)\left(\frac{\partial}{\partial \xi^j} \log p(x;\xi)\right) p(x;\xi)\, dx \qquad (3.7) \\ &= E_p[\partial_i l_\xi \partial_j l_\xi], \end{aligned}$$

where $\partial_i = \partial/\partial\xi^i$, $l_\xi = l(x;\xi) = \log p(x;\xi)$, and $E_p[f]$ is the expectation of $f(x)$ with respect to $p(x;\xi)$. The Fisher information matrix g^F is semi-positive definite in general. Assuming that g^F is positive definite and all components are finite, then g^F can be regarded as a Riemannian metric on S. We call g^F the *Fisher metric* on S. The Fisher metric g^F has the following representations:

$$\begin{aligned} g^F_{ij}(\xi) &= \int_{\Omega} \left(\frac{\partial}{\partial \xi^i} p(x;\xi)\right)\left(\frac{\partial}{\partial \xi^j} \log p(x;\xi)\right) dx \qquad (3.8) \\ &= \int_{\Omega} \frac{1}{p(x;\xi)} \left(\frac{\partial}{\partial \xi^i} p(x;\xi)\right)\left(\frac{\partial}{\partial \xi^j} p(x;\xi)\right) dx. \qquad (3.9) \end{aligned}$$

Next, let us define an affine connection on S. For a fixed $\alpha \in \boldsymbol{R}$, an α-*connection* $\nabla^{(\alpha)}$ on S is defined by

$$\Gamma^{(\alpha)}_{ij,k}(\xi) := E_p\left[\left(\partial_i\partial_j l_\xi + \frac{1-\alpha}{2}\partial_i l_\xi \partial_j l_\xi\right)(\partial_k l_\xi)\right],$$

where $\Gamma^{(\alpha)}_{ij,k}$ is the Christoffel symbol of the first kind of $\nabla^{(\alpha)}$.

We remark that $\nabla^{(0)}$ is the Levi-Civita connection with respect to the Fisher metric g^F. The connection $\nabla^{(e)} := \nabla^{(1)}$ is called the the *exponential connection* and $\nabla^{(m)} := \nabla^{(-1)}$ is called the *mixture connection*. Two connections $\nabla^{(e)}$ and $\nabla^{(m)}$ are expressed as follows:

$$\Gamma^{(e)}_{ij,k} = E_p[(\partial_i \partial_j l_\xi)(\partial_k l_\xi)] \;=\; \int_\Omega \partial_i \partial_j \log p(x;\xi) \partial_k p(x;\xi) dx, \tag{3.10}$$

$$\Gamma^{(m)}_{ij,k} = E_p[((\partial_i \partial_j l_\xi + \partial_i l_\xi \partial_j l_\xi)(\partial_k l_\xi)] \;=\; \int_\Omega \partial_i \partial_j p(x;\xi) \partial_k \log p(x;\xi) dx. \tag{3.11}$$

We can check that the α-connection $\nabla^{(\alpha)}$ is torsion-free and $\nabla^{(\alpha)} g^F$ is totally symmetric. These imply that $(S, \nabla^{(\alpha)}, g^F)$ forms a statistical manifold. In addition, it is known that the Fisher metric g^F and the α-connection $\nabla^{(\alpha)}$ are independent of choice of dominating measures on Ω. Hence we call the triplet $(S, \nabla^{(\alpha)}, g^F)$ an *invariant statistical manifold*. The cubic form C^F of the invariant statistical manifold $(S, \nabla^{(e)}, g^F)$ is given by

$$C^F_{ijk} = \Gamma^{(m)}_{ij,k} - \Gamma^{(e)}_{ij,k}.$$

A statistical model S_e is said to be an *exponential family* if

$$S_e := \left\{ p(x;\theta) \;\middle|\; p(x;\theta) = \exp\left[\sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) \right],\ \theta \in \Theta \subset \boldsymbol{R}^n \right\},$$

under a choice of suitable dominating measure, where $F_1(x), \ldots, F_n(x)$ are functions on the sample space Ω, $\theta = (\theta^1, \ldots, \theta^n)$ is a parameter, and $\psi(\theta)$ is a function of θ for normalization. The following proposition is well-known in information geometry [1].

Theorem 1 (cf. [1]) *For an exponential family S_e, the following hold:*

1. *$(S_e, \nabla^{(e)}, g^F)$ and $(S_e, \nabla^{(m)}, g^F)$ are mutually dual Hessian manifolds, that is, $(S_e, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{(e)}$-affine coordinate system on S_e.*
3. *For the Hessian structure $(\nabla^{(e)}, g^F)$ on S_e, $\psi(\theta)$ is the potential of g^F and C^F with respect to $\{\theta^i\}$:*

$$g^F_{ij}(\theta) = \partial_i \partial_j \psi(\theta), \quad (\partial_i = \partial/\partial\theta^i),$$
$$C^F_{ijk}(\theta) = \partial_i \partial_j \partial_k \psi(\theta).$$

4. *Set the expectation of $F_i(x)$ by $\eta_i := E_p[F_i(x)]$. Then $\{\eta_i\}$ is the dual affine coordinate system of $\{\theta^i\}$ with respect to g^F.*
5. *Set $\phi(\eta) := E_p[\log p(x;\theta)]$. Then $\phi(\eta)$ is the potential of g^F with respect to $\{\eta_i\}$.*

Since $(S_e, \nabla^{(e)}, g^F)$ is a Hessian manifold, the formulas in Proposition 2 hold.

For a statistical model S, we define a *Kullback-Leibler divergence* (or a *relative entropy*) by

$$
\begin{aligned}
D_{KL}(p,r) &:= \int_{\Omega} p(x) \log \frac{p(x)}{r(x)} dx \\
&= E_p[\log p(x) - \log r(x)], \quad (p(x), r(x) \in S).
\end{aligned}
$$

The Kullback-Leibler divergence D_{KL} on an exponential family S_e coincides with the canonical divergence D on $(S_e, \nabla^{(m)}, g^F)$.

We define an $\boldsymbol{R}^n$ valued function $s(x;\xi) = (s^1(x;\xi), \dots, s^n(x;\xi))^T$ by

$$
s^i(x;\xi) := \frac{\partial}{\partial \xi^i} \log p(x;\xi).
$$

We call $s(x;\xi)$ the *score function* of $p(x;\xi)$ with respect to ξ. In information geometry, $s^i(x;\xi)$ is called the *e-(exponential) representation* of $\partial/\partial\xi^i$, and $\partial/\partial\xi^i p(x;\xi)$ is called the *m-(mixture) representation*. The duality of e- and m-representations is important. In fact, Eq. (3.8) implies that the Fisher metric g^F is nothing but an L^2 inner product of e- and m-representations.

Construction of the Kullback-Leibler divergence is as follows. We define a *cross entropy* $d_{KL}(p,r)$ by

$$
d_{KL}(p,r) := -E_p[\log r(x)].
$$

A cross entropy $d_{KL}(p,r)$ gives a bias of information $-\log r(x)$ with respect to $p(x)$. A cross entropy is also called a *yoke* on S [4]. Intuitively, a yoke measures a dissimilarity of two probability density functions on S. We should also note that the cross entropy is obtained by taking the expectation with respect to $p(x)$ of the integrated score function at $r(x)$. Then we have the Kullback-Leibler divergence by

$$
\begin{aligned}
D_{KL}(p,r) &= -d_{KL}(p,p) + d_{KL(}p,r) \\
&= E_p[\log p(x) - \log r(x)].
\end{aligned}
$$

The Kullback-Leibler divergence D_{KL} is a normalized yoke on S, which satisfies $D_{KL}(p,p) = 0$. This argument suggests how to construct divergence functions. Once a function like the cross entropy is defined, we can construct divergence functions in the same way.

3.4 The Deformed Exponential Family

In this section, we review the deformed exponential family. For more details, see [3, 22, 23] and [26]. Geometry of deformed exponential families relates to so-called U-geometry [21].

Let χ be a strictly increasing function from $(0,\infty)$ to $(0,\infty)$. We define a *deformed logarithm function* (or a χ-*logarithm function*) by

$$\log_\chi(s) := \int_1^s \frac{1}{\chi(t)}\,dt.$$

We remark that $\log_\chi(s)$ is strictly increasing and satisfies $\log_\chi(1) = 0$. The domain and the target of $\log_\chi(s)$ depend on the function $\chi(t)$. Set $U = \{s \in (0,\infty) \mid |\log_\chi(s)| < \infty\}$ and $V = \{\log_\chi(s) \mid s \in U\}$. Then $\log_\chi(s)$ is a function from U to V. We also remark that the deformed logarithm is usually called the ϕ-logarithm [23]. However, we use ϕ as the dual potential on a Hessian manifold.

A *deformed exponential function* (or a χ*-exponential function*) is defined by the inverse of the deformed logarithm function $\log_\chi(s)$:

$$\exp_\chi(t) := 1 + \int_0^t \lambda(s)ds,$$

where $\lambda(s)$ is defined by the relation $\lambda(\log_\chi(s)) := \chi(s)$.

When $\chi(s)$ is a power function $\chi(s) = s^q$, $(q > 0, q \neq 1)$, the deformed logarithm and the deformed exponential are given by

$$\begin{aligned} \log_q(s) &:= \frac{s^{1-q}-1}{1-q}, && (s > 0), \\ \exp_q(t) &:= (1 + (1-q)t)^{\frac{1}{1-q}}, && (1 + (1-q)t > 0). \end{aligned}$$

The function $\log_q(s)$ is called the *q-logarithm* and $\exp_q(t)$ the *q-exponential*. Taking the limit $q \to 1$, the standard logarithm and the standard exponential are recovered, respectively.

A statistical model S_χ is said to be a *deformed exponential family* (or a χ*-exponential family*) if

$$S_\chi := \left\{ p(x;\theta) \,\middle|\, p(x;\theta) = \exp_\chi\left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right],\ \theta \in \Theta \subset \boldsymbol{R}^n \right\},$$

under a choice of suitable dominating measure, where $F_1(x), \ldots, F_n(x)$ are functions on the sample space Ω, $\theta = \{\theta^1, \ldots, \theta^n\}$ is a parameter, and $\psi(\theta)$ is the function of θ for normalization. We assume that S_χ is a statistical model in the sense of [1]. That is, $p(x;\theta)$ has support entirely on Ω, there exits a one-to-one correspondence between the parameter θ and the probability distribution $p(x;\theta)$, and differentiation and integration are interchangeable. In addition, functions $\{F_i(x)\}$, $\psi(\theta)$ and parameters $\{\theta^i\}$ must satisfy the anti-exponential condition. For example, in the q-exponential case, these functions satisfy

$$\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta) < \frac{1}{q-1}.$$

Then we can regard that S_χ is a manifold with local coordinate system $\{\theta^i\}$. We also assume that the function ψ is strictly convex since we consider Hessian metrics on S_χ later. A deformed exponential family has several different definitions. See [30] and [34], for example.

For a deformed exponential probability density $p(x;\theta) \in S_\chi$, we define the *escort distribution* $P_\chi(x;\theta)$ of $p(x;\theta)$ by

$$P_\chi(x;\theta) := \frac{1}{Z_\chi(\theta)} \chi\{p(x;\theta)\},$$

where $Z_\chi(\theta)$ is the normalization defined by

$$Z_\chi(\theta) := \int_\Omega \chi\{p(x;\theta)\} dx.$$

The χ*-expectation* $E_{\chi,p}[f]$ of $f(x)$ with respect to $P_\chi(x;\theta)$ is defined by

$$E_{\chi,p}[f] := \int_\Omega f(x) P_\chi(x;\theta)\, dx \;=\; \frac{1}{Z_\chi(\theta)} \int_\Omega f(x)\chi\{p(x;\theta)\} dx.$$

When χ is a power function $\chi(s) = s^q$, $(q > 0, q \neq 0)$, we denote the escort distribution of $p(x;\theta)$ by $P_q(x;\theta)$, and the χ-expectation with respect to $p(x;\theta)$ by $E_{q,p}[*]$.

Example 1 (*discrete distributions* [3]) The set of discrete distributions S_n is a deformed exponential family for an arbitrary χ. Suppose that Ω is a finite set: $\Omega = \{x_0, x_1, \ldots, x_n\}$. Then the statistical model S_n is given by

$$S_n := \left\{ p(x;\eta) \;\middle|\; \eta_i > 0,\; p(x;\eta) = \sum_{i=0}^n \eta_i \delta_i(x),\; \sum_{i=0}^n \eta_i = 1 \right\},$$

where $\eta_0 := 1 - \sum_{i=1}^n \eta_i$ and

$$\delta_i(x) := \begin{cases} 1 & (x = x_i), \\ 0 & (x \neq x_i). \end{cases}$$

Set $\theta^i = \log_\chi p(x_i) - \log_\chi p(x_0) = \log_\chi \eta_i - \log_\chi \eta_0,\; F_i(x) = \delta_i(x)$ and $\psi(\theta) = -\log_\chi \eta_0$. Then the χ-logarithm of $p(x) \in S_n$ is written by

$$\begin{aligned}\log_\chi p(x) &= \sum_{i=1}^{n} \left(\log_\chi \eta_i - \log_\chi \eta_0\right) \delta_i(x) + \log_\chi(\eta_0)\\ &= \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta).\end{aligned}$$

This implies that S_n is a deformed exponential family.

Example 2 (*q-normal distributions* [20]) A q-normal distribution is the probability distribution defined by the following formula:

$$p_q(x;\mu,\sigma) := \frac{1}{Z_q(\sigma)}\left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]_+^{\frac{1}{1-q}},$$

where $[*]_+ := \max\{0, *\}$, $\{\mu, \sigma\}$ are parameters $-\infty < \mu < \infty, 0 < \sigma < \infty$, and $Z_q(\sigma)$ is the normalization defined by

$$Z_q(\sigma) := \begin{cases} \dfrac{\sqrt{3-q}}{\sqrt{1-q}}\mathrm{B}\left(\dfrac{2-q}{1-q}, \dfrac{1}{2}\right)\sigma, & (-\infty < q < 1),\\ \dfrac{\sqrt{3-q}}{\sqrt{q-1}}\mathrm{B}\left(\dfrac{3-q}{2(q-1)}, \dfrac{1}{2}\right)\sigma, & (1 \leq q < 3).\end{cases}$$

Here, $\mathrm{B}\,(*, *)$ is the beta function. We restrict ourselves to consider the case $q \geq 1$. Then the probability distribution $p_q(x;\mu,\sigma)$ has its support entirely on $\boldsymbol{R}$ and the set of q-normal distributions S_q is a statistical model. Set

$$\begin{aligned}\theta^1 &:= \frac{2}{3-q}\{Z_q(\sigma)\}^{q-1}\frac{\mu}{\sigma^2}, \quad \theta^2 := -\frac{1}{3-q}\{Z_q(\sigma)\}^{q-1}\frac{1}{\sigma^2},\\ \psi(\theta) &:= -\frac{(\theta^1)^2}{4\theta^2} - \frac{\{Z_q(\sigma)\}^{q-1}-1}{1-q},\end{aligned}$$

then we have

$$\begin{aligned}\log_q p_q(x;\theta) &= \frac{1}{1-q}(\{p_q(x;\theta)\}^{1-q} - 1)\\ &= \frac{1}{1-q}\left\{\frac{1}{\{Z_q(\sigma)\}^{1-q}}\left(1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right) - 1\right\}\\ &= \frac{2\mu\{Z_q(\sigma)\}^{q-1}}{(3-q)\sigma^2}x - \frac{\{Z_q(\sigma)\}^{q-1}}{(3-q)\sigma^2}x^2\\ &\quad - \frac{\{Z_q(\sigma)\}^{q-1}}{3-q}\cdot\frac{\mu^2}{\sigma^2} + \frac{\{Z_q(\sigma)\}^{q-1}-1}{1-q}\\ &= \theta^1 x + \theta^2 x^2 - \psi(\theta).\end{aligned}$$

This implies that the set of q-normal distributions S_q is a q-exponential family. For a q-normal distribution $p_q(x; \mu, \sigma)$, the q-expectation μ_q and a q-variance σ_q^2 are given by

$$\begin{aligned}\mu_q &= E_{q,p}[x] \;=\; \mu,\\ \sigma_q^2 &= E_{q,p}\left[(x-\mu)^2\right] \;=\; \sigma^2.\end{aligned}$$

We remark that a q-normal distribution is nothing but a three-parameter version of Student's t-distribution when $q \geq 1$. In fact, if $q = 1$, then the q-normal distribution is the normal distribution. If $q = 2$, then the distribution is the Cauchy distribution. We also remark that mathematical properties of q-normal distributions have been obtained by several authors. See [29, 31], for example.

3.5 Geometry of Deformed Exponential Families Derived from the Standard Expectation

In this section, we consider geometry of deformed exponential families by generalizing the e-representation with the deformed logarithm function. For more details, see [21, 26].

Let S_χ be a deformed exponential family. We define an $\boldsymbol{R}^n$ valued function $s^\chi(x;\theta) = \left((s^\chi)^1(x;\theta), \dots, (s^\chi)^n(x;\theta)\right)^T$ by

$$(s^\chi)^i(x;\theta) := \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta), \quad (i = 1, \dots, n). \tag{3.12}$$

We call $s^\chi(x;\theta)$ the χ-*score function* of $p(x;\theta)$. Using the χ-score function, we define a $(0,2)$-tensor field g^M on S_χ by

$$g^M_{ij}(\theta) := \int_\Omega \partial_i p(x;\theta)\partial_j \log_\chi p(x;\theta)\, dx, \quad \left(\partial_i = \frac{\partial}{\partial\theta^i}\right). \tag{3.13}$$

Lemma 1 *The tensor field g^M on S_χ is semi-positive definite.*

Proof From the definitions of g^M and $\log_\chi$, the tensor field g^M is written as

$$g^M_{ij}(\theta) = \int_\Omega \chi(p(x;\theta))\,(F_i(x) - \partial_i\psi(\theta))\left(F_j(x) - \partial_j\psi(\theta)\right) dx. \tag{3.14}$$

Since χ is strictly increasing, g^M is semi-positive definite. □

From now on, we assume that g^M is positive definite. Hence g^M is a Riemannian metric on S_χ. This assumption is same as in the case of Fisher metric. The Riemannian metric g^M is a generalization of the Fisher metric in terms of the representation (3.8).

We can consider other types of generalizations of the Fisher metric as follows.

$$\begin{aligned} g^E_{ij}(\theta) &:= \int_\Omega \left(\partial_i \log_\chi p(x;\theta)\right)\left(\partial_j \log_\chi p(x;\theta)\right) P_\chi(x;\theta)dx \\ &= E_{\chi,p}[\partial_i l_\chi(\theta)\partial_j l_\chi(\theta)], \\ g^N_{ij}(\theta) &:= \int_\Omega \frac{1}{P_\chi(x;\theta)}\left(\partial_i p(x;\theta)\right)\left(\partial_j p(x;\theta)\right) dx, \end{aligned}$$

where $l_\chi(\theta) = \log_\chi p(x;\theta)$. Obviously, g^E and g^N are generalizations of the Fisher metic with respect to the representations (3.7) and (3.9), respectively.

Proposition 4 *Let S_χ be a deformed exponential family. Then Riemannian metrics g^E, g^M and g^N are mutually conformally equivalent. In particular, the following formulas hold:*

$$Z_\chi(\theta)g^E(\theta) = g^M(\theta) = \frac{1}{Z_\chi(\theta)}g^N(\theta),$$

where $Z_\chi(\theta)$ is the normalization of the escort distribution $P_\chi(x;\theta)$.

Proof For a deformed exponential family S_χ, the differentials of probability density functions are given as follows:

$$\begin{aligned} \frac{\partial}{\partial\theta^i}p(x;\theta) &= \chi(p(x;\theta))\left(F_i(x) - \frac{\partial}{\partial\theta^i}\psi(\theta)\right), \\ \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) &= F_i(x) - \frac{\partial}{\partial\theta^i}\psi(\theta). \end{aligned}$$

From the above formula and the definitions of Riemannian metrics g^E and g^N, we have

$$\begin{aligned} g^E_{ij}(\theta) &= \frac{1}{Z_\chi(\theta)}\int_\Omega \chi(p(x;\theta))\left(F_i(x) - \partial_i\psi(\theta)\right)\left(F_j(x) - \partial_j\psi(\theta)\right)dx, \\ g^N_{ij}(\theta) &= Z_\chi(\theta)\int_\Omega \chi(p(x;\theta))\left(F_i(x) - \partial_i\psi(\theta)\right)\left(F_j(x) - \partial_j\psi(\theta)\right)dx. \end{aligned}$$

These equations and Eq. (3.14) imply that Riemannian metrics g^E, g^M and g^N are mutually conformally equivalent. □

Among the three possibilities of generalizations of the Fisher metric, g^M is especially associated with a Hessian structure on S_χ, as we will see below. Although the meaning of g^E is unknown, g^N gives a kind of Cramér-Rao lower bound in statistical inferences. (See [22, 23].)

By differentiating Eq. (3.13), we can define mutually dual affine connections $\nabla^{M(e)}$ and $\nabla^{M(m)}$ on S_χ by

$$\Gamma^{M(e)}_{ij,k}(\theta) := \int_\Omega \partial_k p(x;\theta)\partial_i\partial_j \log_\chi p(x;\theta)dx,$$
$$\Gamma^{M(m)}_{ij,k}(\theta) := \int_\Omega \partial_i\partial_j p(x;\theta)\partial_k \log_\chi p(x;\theta)dx.$$

From the definitions of the deformed exponential family and the deformed logarithm function, $\Gamma^{M(e)}_{ij,k}$ vanishes identically. Hence the connection $\nabla^{M(e)}$ is flat, and $(\nabla^{M(e)}, g^M)$ is a Hessian structure on S_χ. Denote by C^M the cubic form of $(S_\chi, \nabla^{M(e)}, g^M)$, that is,

$$C^M_{ijk} = \Gamma^{M(m)}_{ij,k} - \Gamma^{M(e)}_{ij,k} = \Gamma^{M(m)}_{ij,k}.$$

For $t > 0$, set a function $V_\chi(t)$ by

$$V_\chi(t) := \int_1^t \log_\chi(s)\, ds.$$

We assume that $V_\chi(0) = \lim_{t\to+0} V_\chi(t)$ is finite. Then the *generalized entropy functional* I_χ and the *generalized Massieu potential* Ψ are defined by

$$I_\chi(p_\theta) := -\int_\Omega \left\{V_\chi(p(x;\theta)) + (p(x;\theta)-1)V_\chi(0)\right\} dx,$$
$$\Psi(\theta) := \int_\Omega p(x;\theta)\log_\chi p(x;\theta)dx + I_\chi(p_\theta) + \psi(\theta),$$

respectively, where ψ is the normalization of the deformed exponential family.

Theorem 2 (cf. [21, 26]) *For a deformed exponential family S_χ, the following hold:*

1. *$(S_\chi, \nabla^{M(e)}, g^M)$ and $(S_\chi, \nabla^{M(m)}, g^M)$ are mutually dual Hessian manifolds, that is, $(S_\chi, g^M, \nabla^{M(e)}, \nabla^{M(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{M(e)}$-affine coordinate system on S_χ.*
3. *$\Psi(\theta)$ is the potential of g^M and C^M with respect to $\{\theta^i\}$, that is,*

$$g^M_{ij}(\theta) = \partial_i\partial_j\Psi(\theta),$$
$$C^M_{ijk}(\theta) = \partial_i\partial_j\partial_k\Psi(\theta).$$

4. *Set the expectation of $F_i(x)$ by $\eta_i := E_p[F_i(x)]$. Then $\{\eta_i\}$ is a $\nabla^{M(m)}$-affine coordinate system on S_χ and the dual of $\{\theta^i\}$ with respect to g^M.*
5. *Set $\Phi(\eta) := -I_\chi(p_\theta)$. Then $\Phi(\eta)$ is the potential of g^M with respect to $\{\eta_i\}$.*

Let us construct a divergence function which induces the Hessian manifold $(S_\chi, \nabla^{M(e)}, g^M)$. We define the *bias corrected χ-score function* $u_p^\chi(x;\theta)$ of $p(x;\theta)$ by

$$(u_p^\chi)^i(x;\theta) := \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right].$$

Set a function $U_\chi(t)$ by

$$U_\chi(s) := \int_0^s \exp_\chi(t)\,dt.$$

Then we have

$$\begin{aligned} V_\chi(s) &= s\log_\chi(s) - \int_1^s t\left(\frac{d}{dt}\log_\chi(t)\right)dt \\ &= s\log_\chi(s) - \int_0^{\log_\chi(s)} \exp_\chi(u)du \\ &= s\log_\chi(s) - U_\chi(\log_\chi(s)). \end{aligned}$$

Since $\partial/\partial\theta^i V_\chi(p(x;\theta)) = (\partial/\partial\theta^i p(x;\theta))\log_\chi p(x;\theta)$, we have

$$p(x;\theta)\left(\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right) = \frac{\partial}{\partial\theta^i}U_\chi(\log_\chi p(x;\theta)).$$

Hence, by integrating the bias corrected χ-score function at $r(x;\theta)\in S_\chi$ with respect to θ, and by taking the standard expectation with respect to $p(x;\theta)$, we define a *χ-cross entropy of Bregman type* by

$$d_\chi^M(p,r) = -\int_\Omega p(x)\log_\chi r(x)dx + \int_\Omega U_\chi(\log_\chi r(x))dx.$$

Then we obtain the *χ-divergence* (or *U-divergence*) by

$$\begin{aligned} D_\chi(p,r) &= -d_\chi^M(p,p) + d_\chi^M(p,r) \\ &= \int_\Omega \big\{U_\chi(\log_\chi r(x)) - U_\chi(\log_\chi p(x)) \\ &\qquad -p(x)(\log_\chi r(x) - \log_\chi p(x))\big\}\,dx. \end{aligned}$$

In the q-exponential case, the *bias corrected q-score function* is given by

$$\begin{aligned} u_q^i(x;\theta) &= \frac{\partial}{\partial\theta^i}\log_q p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_q p(x;\theta)\right] \\ &= \frac{\partial}{\partial\theta^i}\left\{\frac{1}{1-q}p(x;\theta)^{1-q} - \frac{1}{2-q}\int_\Omega p(x;\theta)^{2-q}dx\right\} \\ &= p(x;\theta)^{1-q}s^i(x;\theta) - E_p[p(x;\theta)^{1-q}s^i(x;\theta)]. \end{aligned}$$

This score function is nothing but a weighted score function in robust statistics. The χ-divergence constructed from the bias corrected q-score function coincides with the *β-divergence* ($\beta = 1-q$):

$$\begin{aligned} D_{1-q}(p,r) &= -d_{1-q}(p,p) + d_{1-q}(p,r) \\ &= \frac{1}{(1-q)(2-q)}\int_\Omega p(x)^{2-q}dx \\ &\quad - \frac{1}{1-q}\int_\Omega p(x)r(x)^{1-q}dx + \frac{1}{2-q}\int_\Omega r(x)^{2-q}dx. \end{aligned}$$

3.6 Geometry of Deformed Exponential Families Derived from the χ-Expectation

Since S_χ is linearizable by the deformed logarithm function, we can naturally define geometric structures from the potential function ψ.

A *χ-Fisher metric* g^χ and a *χ-cubic form* C^χ are defined by

$$\begin{aligned} g_{ij}^\chi(\theta) &:= \partial_i\partial_j\psi(\theta), \\ C_{ijk}^\chi(\theta) &:= \partial_i\partial_j\partial_k\psi(\theta), \end{aligned}$$

respectively [3]. In the q-exponential case, we denote the χ-Fisher metric by g^q, and the χ-cubic form by C^q. We call g^q and C^q a *q-Fisher metric* and a *q-cubic form*, respectively.

Let $\nabla^{\chi(0)}$ be the Levi-Civita connection with respect to the χ-Fisher metric g^χ. Then a *χ-exponential connection*$\nabla^{\chi(e)}$ and a *χ-mixture connection*$\nabla^{\chi(m)}$ are defined by

$$\begin{aligned} g^\chi(\nabla_X^{\chi(e)}Y,Z) &:= g^\chi(\nabla_X^{\chi(0)}Y,Z) - \frac{1}{2}C^\chi(X,Y,Z), \\ g^\chi(\nabla_X^{\chi(m)}Y,Z) &:= g^\chi(\nabla_X^{\chi(0)}Y,Z) + \frac{1}{2}C^\chi(X,Y,Z), \end{aligned}$$

respectively. The following theorem is known in [3].

Theorem 3 (cf. [3]) *For a deformed exponential family S_χ, the following hold:*

1. *$(S_\chi, \nabla^{\chi(e)}, g^\chi)$ and $(S_\chi, \nabla^{\chi(m)}, g^\chi)$ are mutually dual Hessian manifolds, that is, $(S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$ is a dually flat space.*
2. *$\{\theta^i\}$ is a $\nabla^{\chi(e)}$-affine coordinate system on S_χ.*
3. *$\psi(\theta)$ is the potential of g^χ and C^χ with respect to $\{\theta^i\}$.*
4. *Set the χ-expectation of $F_i(x)$ by $\eta_i := E_{\chi,p}[F_i(x)]$. Then $\{\eta_i\}$ is a $\nabla^{\chi(m)}$-affine coordinate system on S_χ and the dual of $\{\theta^i\}$ with respect to g^χ.*
5. *Set $\phi(\eta) := E_{\chi,p}[\log_\chi p(x;\theta)]$. Then $\phi(\eta)$ is the potential of g^χ with respect to $\{\eta_i\}$.*

Proof Statements 1, 2 and 3 are easily obtained from the definitions of χ-Fisher metric and χ-cubic form. From Eq. (3.3) and $\eta_i = E_{\chi,p}[F_i(x)]$, Statements 4 and 5 follow from the fact that

$$E_{\chi,p}[\log_\chi p(x;\theta)] = E_{\chi,p}\left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right] = \sum_{i=1}^n \theta^i \eta_i - \psi(\theta). \quad \square$$

Suppose that $s^\chi(x;\theta)$ is the χ-score function defined by (3.12). The χ-score is unbiased with respect to χ-expectation, that is, $E_{\chi,p}[(s^\chi)^i(x;\theta)] = 0$. Hence we regard that $s^\chi(x;\theta)$ is a generalization of unbiased estimating functions.

By integrating a χ-score function, we define the *χ-cross entropy* by

$$\begin{aligned} d^\chi(p,r) &:= -E_{\chi,p}[\log_\chi r(x)] \\ &= -\int_\Omega P(x)\log_\chi r(x)dx. \end{aligned}$$

Then we obtain the *generalized relative entropy $D^\chi(p,r)$* by

$$\begin{aligned} D^\chi(p,r) &:= -d^\chi(p,p) + d^\chi(p,r) \\ &= E_{\chi,p}[\log_\chi p(x) - \log_\chi r(x)]. \end{aligned} \tag{3.15}$$

The generalized relative entropy $D^\chi(p,r)$ coincides with the canonical divergence $D(r,p)$ for $(S_\chi, \nabla^{\chi(e)}, g^\chi)$. In fact, from (3.15), we can check that

$$\begin{aligned} D^\chi(p(\theta), p(\theta')) &= E_{\chi,p}\left[\left(\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right) - \left(\sum_{i=1}^n (\theta')^i F_i(x) - \psi(\theta')\right)\right] \\ &= \psi(\theta') + \sum_{i=1}^n \theta^i \eta_i - \psi(\theta) - \sum_{i=1}^n (\theta')^i \eta_i \ = \ D(p(\theta'), p(\theta)). \end{aligned}$$

Let us consider the q-exponential case. We assume that a q-exponential family S_q admits an invariant statistical manifold structure $(S_q, \nabla^{(\alpha)}, g^F)$.

Theorem 4 ([20]) *For a q-exponential family S_q, the invariant statistical manifold $(S_q, \nabla^{(2q-1)}, g^F)$ and the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$ are 1-conformally equivalent. In this case, the invariant statistical manifold $(S_q, \nabla^{(2q-1)}, g^F)$ is 1-conformally flat.*

Divergence functions for $(S_q, \nabla^{q(e)}, g^q)$ and $(S_q, \nabla^{(2q-1)}, g^F)$ are given as follows. The *α-divergence* $D^{(\alpha)}(p, r)$ with $\alpha = 1 - 2q$ is defined by

$$D^{(1-2q)}(p, r) := \frac{1}{q(1-q)} \left\{ 1 - \int_\Omega p(x)^q r(x)^{1-q} dx \right\}.$$

On the other hand, the *normalized Tsallis relative entropy* $D_q^T(p, r)$ is defined by

$$\begin{aligned} D_q^T(p, r) &:= \int_\Omega P_q(x) \left(\log_q p(x) - \log_q r(x) \right) dx \\ &= E_{q,p}[\log_q p(x) - \log_q r(x)]. \end{aligned}$$

We remark that the invariant statistical manifold $(S_q, \nabla^{(1-2q)}, g^F)$ is induced from the α-divergence with $\alpha = 1 - 2q$, and that the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$ is induced from the dual of the normalized Tsallis relative entropy. In fact, for a q-exponential family S_q, divergence functions have the following relations:

$$\begin{aligned} D(r, p) &= D_q^T(p, r) \\ &= \int_\Omega \frac{p(x)^q}{Z_q(p)} \left(\log_q p(x) - \log_q r(x) \right) dx \\ &= \frac{1}{Z_q(p)} \int_\Omega \left(\frac{p(x) - p(x)^q}{1-q} - \frac{p(x)^q r(x)^{1-q} - p(x)^q}{1-q} \right) dx \\ &= \frac{1}{(1-q)Z_q(p)} \left\{ 1 - \int_\Omega p(x)^q r(x)^{1-q} dx \right\} \\ &= \frac{q}{Z_q(r)} D^{(1-2q)}(p, r), \end{aligned}$$

where D is the canonical divergence of the Hessian manifold $(S_q, \nabla^{q(e)}, g^q)$.

3.7 Maximum q-Likelihood Estimators

In this section, we generalize the maximum likelihood method from the viewpoint of generalized independence. To avoid complicated arguments, we restrict ourselves to consider the q-exponential case. However, we can generalize it to the χ-exponential case (cf. [8, 9]).

Let X and Y be random variables which follow probability distributions $p_1(x)$ and $p_2(y)$, respectively. We say that two random variables X and Y are *independent* if the joint probability $p(x, y)$ is decomposed by a product of marginal distributions $p_1(x)$ and $p_2(Y)$:

$$p(x, y) = p_1(x)p_2(y).$$

When $p_1(x) > 0$ and $p_2(y) > 0$, the independence can be written with an exponential function and a logarithm function by

$$p(x, y) = \exp\left[\log p_1(x) + \log p_2(x)\right].$$

We generalize the notion of independence using the q-exponential and q-logarithm. Suppose that $x > 0, y > 0$ and $x^{1-q} + y^{1-q} - 1 > 0$ $(q > 0)$. We say that $x \otimes_q y$ is a *q-product* [6] of x and y if

$$\begin{aligned} x \otimes_q y &:= \left[x^{1-q} + y^{1-q} - 1\right]^{\frac{1}{1-q}} \\ &= \exp_q\left[\log_q x + \log_q y\right]. \end{aligned}$$

In this case, the following low of exponents holds:

$$\exp_q x \otimes_q \exp_q y = \exp_q(x + y),$$

in other words,

$$\log_q(x \otimes_q y) = \log_q x + \log_q y.$$

Let X_i be a random variable on $\mathcal{X}_i$ which follows $p_i(x)$ $(i = 1, 2, \dots, N)$. We say that $X_1, X_2, \dots, X_N$ are *q-independent with m-normalization* (mixture normalization) if

$$p(x_1, x_2, \dots, x_N) = \frac{p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N)}{Z_{p_1,p_2,\cdots,p_N}}$$

where $p(x_1, x_2, \dots, x_N)$ is the joint probability density of $X_1, X_2, \dots, X_N$ and $Z_{p_1,p_2,\cdots,p_N}$ is the normalization of $p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N)$ defined by

$$Z_{p_1,p_2,\cdots,p_N} := \int \cdots \int_{\mathcal{X}_1 \cdots \mathcal{X}_N} p_1(x_1) \otimes_q p_2(x_2) \otimes_q \cdots \otimes_q p_N(x_N) dx_1 \cdots dx_N.$$

Let $S_q = \{p(x;\xi) | \xi \in \Xi\}$ be a q-exponential family, and let $\{x_1, \dots, x_N\}$ be N-observations from $p(x;\xi) \in S_q$. We define a *q-likelihood function* $L_q(\xi)$ by

$$L_q(\xi) = p(x_1;\xi) \otimes_q p(x_2;\xi) \otimes_q \cdots \otimes_q p(x_N;\xi).$$

Equivalently, a *q-log-likelihood function* is given by

$$\log_q L_q(\xi) = \sum_{i=1}^{N} \log_q p(x_i;\xi).$$

In the case $q \to 1$, L_q is the standard likelihood function on Ξ.

The *maximum q-likelihood estimator* $\hat{\xi}$ is the maximizer of the q-likelihood functions, which is defined by

$$\hat{\xi} := \operatorname*{argmax}_{\xi \in \Xi} L_q(\xi) \quad \left(= \operatorname*{argmax}_{\xi \in \Xi} \log_q L_q(\xi) \right).$$

Let us consider geometry of maximum q-likelihood estimators. Let S_q be a q-exponential family. Suppose that $\{x_1, \dots, x_N\}$ are N-observations generated from $p(x;\theta) \in S_q$.

The q-log-likelihood function is calculated as

$$\begin{aligned}\log_q L_q(\theta) = \sum_{j=1}^{N} \log_q p(x_j;\theta) &= \sum_{j=1}^{N} \left\{ \sum_{i=1}^{n} \theta^i F_i(x_j) - \psi(\theta) \right\} \\ &= \sum_{i=1}^{n} \theta^i \sum_{j=1}^{N} F_i(x_j) - N\psi(\theta).\end{aligned}$$

The q-log-likelihood equation is

$$\partial_i \log_q L_q(\theta) = \sum_{j=1}^{N} F_i(x_j) - N\partial_i \psi(\theta) = 0.$$

Thus, the maximum q-likelihood estimator for η is given by

$$\hat{\eta}_i = \frac{1}{N} \sum_{j=1}^{N} F_i(x_j).$$

On the other hand, the canonical divergence for $(S_q, \nabla^{q(e)}, g^q)$ can be calculated as

$$\begin{aligned} D_q^T(p(\hat{\eta}), p(\theta)) &= D(p(\theta), p(\hat{\eta})) \\ &= \psi(\theta) + \phi(\hat{\eta}) - \sum_{i=1}^{n} \theta^i \hat{\eta}_i \\ &= \phi(\hat{\eta}) - \frac{1}{N} \log_q L_q(\theta). \end{aligned}$$

This implies that the q-likelihood attains the maximum if and only if the normalized Tsallis relative entropy attains the minimum.

Let M be a *curved q-exponential family* in S_q, that is, M is a submanifold in S_q and is a statistical model itself. Suppose that $\{x_1, \ldots, x_N\}$ are N-observations generated from $p(x; u) = p(x; \theta(u)) \in M$. The above arguments implies that the maximum q-likelihood estimator for M is given by the orthogonal projection of data with respect to the normalized Tsallis relative entropy.

We remark that the maximum q-likelihood estimator can be generalized by U-geometry. (See [8, 9] by Fujimoto and Murata.) However, their approach and ours are slightly different. They applied the χ-divergence (U-divergence) projection for a parameter estimation, whereas we applied the generalized relative entropy. As we discussed in this paper, the induced Hessian structures from those divergences are different.

3.8 Conclusion

In this paper, we considered two Hessian structures from the viewpoints of the standard expectation and the χ-expectation. Though the former and the later are known as U-geometry ([21, 26]) and χ-geometry ([3]), respectively, they turn out to be different Hessian structures in the same deformed exponential family through a comparison of each other.

We note that, from the viewpoint of estimating functions, the former is geometry of bias-corrected χ-score functions with the standard expectation, whereas the later is geometry of unbiased χ-score functions with the χ-expectation.

As an application to statistics, we considered generalization of maximum likelihood method for q-exponential family. We used the normalized Tsallis relative entropy for orthogonal projection, whereas the previous results used χ-divergences of Bregman type.

Acknowledgments The authors would like to express their sincere gratitude to the anonymous reviewers for constructive comments for preparation of this paper. The first named author is partially supported by JSPS KAKENHI Grant Number 23740047.

References

1. Amari, S., Nagaoka, H.: Method of Information Geometry. American Mathematical Society, Providence, Oxford University Press, Oxford (2000)
2. Amari, S., Ohara, A.: Geometry of q-exponential family of probability distributions. Entropy **13**, 1170–1185 (2011)
3. Amari, S., Ohara, A., Matsuzoe, H.: Geometry of deformed exponential families: invariant, dually-flat and conformal geometry. Phys. A. **391**, 4308–4319 (2012)
4. Barondorff-Nielsen, O.E., Jupp, P.E.: Statistics, yokes and symplectic geometry. Ann. Facul. Sci. Toulouse **6**, 389–427 (1997)
5. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998)
6. Borgesa, E.P.: A possible deformed algebra and calculus inspired in nonextensive thermostatistics. Phys. A **340**, 95–101 (2004)
7. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
8. Fujimoto, Y., Murata, N.: A generalization of independence in naive bayes model. Lect. Notes Comp. Sci. **6283**, 153–161 (2010)
9. Fujimoto Y., Murata N.: A generalisation of independence in statistical models for categorical distribution. Int. J. Data Min. Model. Manage. **2**(4), 172–187 (2012)
10. Ivanov, S.: On dual-projectively flat affine connections. J. Geom. **53**, 89–99 (1995)
11. Kumon, M., Takemura, A., Takeuchi, K.: Conformal geometry of statistical manifold with application to sequential estimation. Sequential Anal. **30**, 308–337 (2011)
12. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tôhoku Math. J. **46**, 427–433 (1994)
13. Kurose, T.: Conformal-projective geometry of statistical manifolds. Interdiscip. Inform. Sci. **8**, 89–100 (2002)
14. Lauritzen, S. L.: Statistical Manifolds, Differential Geometry in Statistical Inferences, IMS Lecture Notes Monograph Series, vol. 10, pp. 96–163. Hayward, California (1987)
15. Matsuzoe, H.: Geometry of contrast functions and conformal geometry. Hiroshima Math. J. **29**, 175–191 (1999)
16. Matsuzoe, H.: Geometry of statistical manifolds and its generalization. In: Proceedings of the 8th International Workshop on Complex Structures and Vector Fields, pp. 244–251. World Scientific, Singapore (2007)
17. Matsuzoe, H.: Computational geometry from the viewpoint of affine differential geometry. Lect. Notes Comp. Sci. **5416**, 103–113 (2009)
18. Matsuzoe, H.: Statistical manifolds and geometry of estimating functions, pp. 187–202. Recent Progress in Differential Geometry and Its Related Fields World Scientific, Singapore (2013)
19. Matsuzoe, H., Henmi, M.: Hessian structures on deformed exponential families. Lect. Notes Comp. Sci. **8085**, 275–282 (2013)
20. Matsuzoe, H., Ohara, A.: Geometry for q-exponential families. In: Recent progress in differential geometry and its related fields, pp. 55–71. World Scientific, Singapore (2011)
21. Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of u-boost and bregman divergence. Neural Comput. **16**, 1437–1481 (2004)
22. Naudts, J.: Estimators, escort probabilities, and ϕ-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
23. Naudts, J.: Generalised Thermostatistics, Springer, New York (2011)
24. Ohara, A.: Geometric study for the legendre duality of generalized entropies and its application to the porous medium equation. Euro. Phys. J. B. **70**, 15–28 (2009)
25. Ohara, A., Matsuzoe H., Amari S.: Conformal geometry of escort probability and its applications. Mod. Phys. Lett. B. **10**, 26:1250063 (2012)
26. Ohara A., Wada, T.: Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations. J. Phys. A: Math. Theor. **43**, 035002 (2010)
27. Okamoto, I., Amari, S., Takeuchi, K.: Asymptotic theory of sequential estimation procedures for curved exponential families. Ann. Stat. **19**, 961–961 (1991)

28. Shima, H.: The Geometry of Hessian Structures, World Scientific, Singapore (2007)
29. Suyari, H., Tsukada, M.: Law of error in tsallis statistics. IEEE Trans. Inform. Theory **51**, 753–757 (2005)
30. Takatsu, A.: Behaviors of φ-exponential distributions in wasserstein geometry and an evolution equation. SIAM J. Math. Anal. **45**, 2546–2546 (2013)
31. Tanaka, M.: Meaning of an escort distribution and τ-transformation. J. Phys.: Conf. Ser. **201**, 012007 (2010)
32. Tsallis, C.: Possible generalization of boltzmann—gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
33. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer, New York (2009)
34. Vigelis, R.F., Cavalcante, C.C.: On ϕ-families of probability distributions. J. Theor. Probab. **21**, 1–25 (2011)