

Article

Duality of Maximum Entropy and Minimum Divergence

Shinto Eguchi ^{1,*}, Osamu Komori ² and Atsumi Ohara ³

¹ The Institute of Statistical Mathematics and The Graduate University of Advanced Studies, Tachikawa Tokyo 190-8562, Japan

² The Institute of Statistical Mathematics, Tachikawa Tokyo 190-8562, Japan;
E-Mail: komori@ism.ac.jp

³ Department of Electrical and Electronics Engineering, University of Fukui, Fukui 910-8507, Japan;
E-Mail: ohara@fuee.u-fukui.ac.jp

* Author to whom correspondence should be addressed; E-Mail: eguchi@ism.ac.jp;
Tel.: +81-50-5533-8500.

Received: 28 April 2014; in revised form: 19 June 2014 / Accepted: 24 June 2014 /

Published: 26 June 2014

Abstract: We discuss a special class of generalized divergence measures by the use of generator functions. Any divergence measure in the class is separated into the difference between cross and diagonal entropy. The diagonal entropy measure in the class associates with a model of maximum entropy distributions; the divergence measure leads to statistical estimation via minimization, for arbitrarily giving a statistical model. The dualistic relationship between the maximum entropy model and the minimum divergence estimation is explored in the framework of information geometry. The model of maximum entropy distributions is characterized to be totally geodesic with respect to the linear connection associated with the divergence. A natural extension for the classical theory for the maximum likelihood method under the maximum entropy model in terms of the Boltzmann-Gibbs-Shannon entropy is given. We discuss the duality in detail for Tsallis entropy as a typical example.

Keywords: β -divergence; dual connections; information geometry; MaxEnt; multivariate t -distribution; power exponential family; sufficiency

1. Introduction

Information divergence plays a central role in the understanding of integrating statistics, information science, statistical physics and machine learning. Let \mathcal{F} be the space of all the probability density functions with a common support with respect to a carrier measure Λ of a data space. Usually Λ is taken as the Lebesgue measure and the counting measure corresponding to continuous and discrete random variables, respectively. The most typical example of information divergence is the Kullback-Leibler divergence

$$D_0(f, g) = \int f(x)\{\log f(x) - \log g(x)\}d\Lambda(x)$$

on \mathcal{F} , which is decomposed into the difference of cross and diagonal entropy measures

$$C_0(f, g) = - \int f(x) \log g(x)d\Lambda(x)$$

and

$$H_0(f, g) = - \int f(x) \log f(x)d\Lambda(x).$$

The entropy $H_0(f)$ is nothing but Boltzmann-Gibbs-Shannon entropy. In effect, $D_0(f, g)$ connects the maximum likelihood [1,2], and the maximum entropy [3]. If we take a canonical statistic $t(X)$, then the maximum entropy distribution under a moment constraint for $t(X)$ belongs to the exponential model associated with $t(X)$,

$$M^{(e)} = \{f_\theta(x, \theta) := \exp\{\theta^\top t(x) - \kappa_0(\theta)\} : \theta \in \Theta\} \tag{1}$$

where $\kappa_0(\theta) = \log \int \exp\{\theta^\top t(x)\}d\Lambda(x)$ and $\Theta = \{\theta : \kappa_0(\theta) < \infty\}$. In this context, the statistic $t(X)$ is minimally sufficient in the model, in which the maximum likelihood estimator (MLE) for the parameter θ of the model is given by one-to-one correspondence with $t(X)$, see [4] for the convex geometry. If we consider the expectation parameter,

$$\mu = \mathbb{E}_{f_\theta(\cdot, \theta)}\{t(X)\}$$

in place of θ , then for a given random sample X_1, \dots, X_n , the MLE for μ is given by the sample mean of $t(X_i)$'s, that is

$$\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^n t(X_i).$$

We define two kinds of geodesic curves connecting f and g in \mathcal{F} . We call a curve

$$C^{(m)} = \{C_t^{(m)}(x) := (1 - t)f(x) + tg(x) : t \in (0, 1)\} \tag{2}$$

mixture-geodesic. Alternatively, we call a curve

$$C^{(e)} = \{C_t^{(e)}(x) := \exp\{(1 - t) \log f(x) + t \log g(x) - \kappa(t)\} : t \in (0, 1)\} \tag{3}$$

exponential geodesic, where $\kappa(t) = \log \int f(x)^{1-t}g(x)^td\Lambda(x)$. We denote $\Gamma^{(m)}$ and $\Gamma^{(e)}$ the two linear connections induced by the mixture and exponential geodesic curves on \mathcal{F} , which we call the mixture

connection and exponential connection on \mathcal{F} , respectively, see [5,6]. Thus all tangent vectors on a mixture geodesic curve are parallel to each other with respect to $\Gamma^{(m)}$; all tangent vectors on an exponential geodesic curve are parallel to each other with respect to $\Gamma^{(e)}$. It is well-known that $M^{(e)}$ is totally exponential-geodesic, that is, for any $f_0(x, \theta_0)$ and $f_0(x, \theta_1)$ in $M^{(e)}$ it holds that the exponential geodesic curve connecting $f_0(x, \theta_0)$ and $f_0(x, \theta_1)$ is in $M^{(e)}$. In effect we observe that $C_t^{(e)}(x) = f_0(x, \theta_t)$ with $\theta_t = (1-t)\theta_0 + t\theta_1$. Thus $C_t^{(e)}(x) \in M^{(e)}$ for all $t \in (0, 1)$ because Θ is a convex set. Alternatively, consider a parametric model

$$M^{(m)} = \{f_1(x, \pi) := \sum_{j=0}^d \pi_j f_j(x) : \pi_j > 0 (j = 0, \dots, d), \sum_{j=0}^d \pi_j = 1\}.$$

Then, $M^{(m)}$ is totally mixture-geodesic. Because a mixture geodesic curve $C_t^{(m)}(x) = (1-t)f_1(x, \pi_0) + tf_1(x, \pi_1)$ is in $M^{(m)}$ for any $t \in (0, 1)$ on account of $C_t^{(m)}(x) = f_1(x, \pi_t)$, where $(1-t)\pi_0 + t\pi_1$.

We discuss a generalized entropy and divergence measures with applications in statistical models and estimation. There have been recent developments for the generalization of Boltzmann-Shannon entropy and Kullback-Leibler divergence. We focus on U -divergence with a generator function U , in which U -divergence is separated into the differences between cross entropy and diagonal entropy. We observe a dualistic property associated with U -divergence between statistical model and estimation. The U -loss function is given by an empirical approximation for U -divergence based on a given dataset under a statistical model, in which the U -estimator is defined by minimization of the U -loss function on the parameter space. On the other hand, the diagonal entropy leads to a maximum entropy distribution with a mean equal space, where we call the family of distributions U -model. In accordance with this, the U -divergence leads to a pair of U -model and U -estimator as a statistical model and estimation. The typical example is that $U(t) = \exp(t)$, which is associated with the Kullback-Leibler divergence $D_0(f, g)$ generating a pair of an exponential family $M^{(e)}$ and the minus log-likelihood function.

This aspect is characterized as a minimax game between a decision maker and Nature.

The paper is organized as follows. Section 2 introduces the class of U -divergence measures. The information geometric framework associated with a divergence measure is given in Section 3. In Section 3 we discuss the maximum entropy model with respect to U -diagonal entropy. The minimum divergence method via U -divergence is discussed in Section 5. We next explore the duality between maximum U -entropy and minimum U -divergence in Section 6. Finally, we discuss the relation to the robust statistics by minimum divergence, and a future problem on MaxEnt in Section 7.

2. U -Divergence

A class of information divergence is constructed by a generator function U via a simple employment of conjugate convexity, see [7]. We introduce a class of generator functions by

$$\mathcal{U} = \{U : \mathbb{R} \rightarrow \mathbb{R}_+ : \frac{d}{ds}U(s) \geq 0, \frac{d^2}{ds^2}U(s) \geq 0\}.$$

Then we consider the conjugate convex function defined on \mathbb{R}_+ of U in \mathcal{U} as

$$U^*(t) = \max_{s \in \mathbb{R}} \{st - U(s)\},$$

and hence $U^*(t) = t\xi(t) - U(\xi(t))$, where $\xi(t)$ is the inverse function of the derivative of $U(s)$, or equivalently $(dU/ds)(\xi(t)) = t$. The existence for $\xi(t)$ is guaranteed from the assumption for U to be in \mathcal{U} , in which we observe an important property that the derivative of U^* is the inverse of the derivative of U , that is

$$\frac{d}{dt}U^*(t) = \xi(t). \tag{4}$$

The conjugate function U^* of U is reflexible, that is, $U^{**} = U$. By definition, for any $s \in \mathbb{R}$ and $t \in \mathbb{R}_+$,

$$U^*(t) \geq st - U(s) \tag{5}$$

with equality if and only if $s = \xi(t)$. We consider an information divergence functional using the generator function U as

$$D_U(f, g) = \int \{U^*(f) - f\xi(g) + U(\xi(g))\}d\Lambda, \tag{6}$$

called U -divergence. We can easily confirm that $D_U(f, g)$ satisfies the first axiom of a distance function since the integrand in Equation (6) is always nonnegative with equality of 0 if and only if $f(x) = g(x)$ because Equation (5). It follows from the construction that $D_U(f, g)$ is decomposed into $C_U(f, g)$ and $H_U(f)$ such that

$$D_U(f, g) = C_U(f, g) - H_U(f).$$

Here

$$C_U(f, g) = \int \{U(\xi(g)) - f\xi(g)\}d\Lambda,$$

is called U -cross entropy;

$$H_U(f) = - \int U^*(f)d\Lambda \tag{7}$$

is called U -diagonal entropy. We can write $H_U(f) = \int \{U(\xi(f)) - f\xi(f)\}d\Lambda$ by the definition for U^* , which equals the diagonal $C_U(f, f)$. We note that the U -divergence is expressed as

$$D_U(f, g) = \int \{U^*(f) - U^*(g) - \xi(g)(f - g)\}d\Lambda$$

because of Equation (4), which implies that U^* plays a role on a generator function in place of U . In fact, this is also called U^* -Bregman divergence, cf. [8,9]

The first example of U is $U_0(s) = \exp(s)$, which leads to $U_0^*(t) = t \log t - t$ and

$$\log(t) = \operatorname{argmax}_{s \in \mathbb{R}} \{st - \exp(s)\},$$

Thus U_0 -divergence, U_0 -cross entropy and U_0 -diagonal entropy equal $D_0(f, g)$, $C_0(f, g)$ and $H_0(f)$ as defined in Introduction, respectively. As for the second example we consider

$$U_\beta(s) = \frac{1}{\beta + 1} (1 + \beta s)^{\frac{1+\beta}{\beta}} \tag{8}$$

where β is a scalar. The conjugate function becomes

$$U_{\beta}^*(t) = \frac{1}{\beta(\beta + 1)}t^{\beta+1} - \frac{1}{\beta}t. \tag{9}$$

Then the generator function U_{β} associates with the β -power cross entropy

$$C_{\beta}(f, g) = \frac{1}{\beta + 1} \int g^{\beta+1}d\Lambda - \frac{1}{\beta} \int f(g^{\beta} - 1)d\Lambda,$$

β -diagonal power entropy

$$H_{\beta}(f) = -\frac{1}{\beta(\beta + 1)} \int f^{\beta+1}d\Lambda + \frac{1}{\beta}$$

and the β -power divergence $D_{\beta}(f, g) = C_{\beta}(f, g) - H_{\beta}(f)$, that is,

$$D_{\beta}(f, g) = \frac{1}{\beta(\beta + 1)} \int f^{\beta+1}d\Lambda - \frac{1}{\beta} \int fg^{\beta}d\Lambda + \frac{1}{\beta + 1} \int g^{\beta+1}d\Lambda.$$

We observe that

$$\lim_{\beta \rightarrow 0}(C_{\beta}(f, g), H_{\beta}(f)) = (C_0(f, g), H_0(f)).$$

The class of β -power divergence functionals includes the Kullback-Leibler divergence in the limiting sense of $\lim_{\beta \rightarrow 0} D_{\beta}(f, g) = D_0(f, g)$. If $\beta = 1$, then $D_{\beta}(f, g) = \frac{1}{2} \int (f - g)^2 d\Lambda$, which is a half of the squared L_2 norm. If we take a limit of β to -1 , then $D_{\beta}(f, g)$ becomes the Itakura-Saito divergence

$$D_{IS}(f, g) = \int \left(\log g - \log f + \frac{f}{g} - 1 \right) d\Lambda,$$

which is widely applied in signal processing and speech recognition, cf. [10–12].

The β -power divergence $D_{\beta}(p, q)$ is proposed in [13]; the β -power entropy H_{β} is equal to the Tsallis q -entropy with a relation $q = \beta + 1$, cf. [14–16]. Tsallis entropy is connected with spin glass relaxation, dissipative optical lattices and so on beyond the classical statistical physics associated with the Boltzmann-Shannon entropy $H_0(p)$. See also [17,18] for the power entropy in the field of ecology. We will discuss the statistical property for the minimum β divergence method in the presence of outliers departing from a supposed model, cf. [19–21]. A robustness performance is elucidated by appropriate selection for β . Beyond robustness perspective, a property of spontaneous learning to apply to clustering analysis is focused in [22], see also [23] for nonnegative matrix analysis.

The third example of a generator function is $U_{\eta}(s) = (1 - \eta) \exp(s) - \eta s$ with a scalar η . This generator function leads to the η -cross entropy

$$C_{\eta}(f, g) = - \int \{f(x) + \eta\} \log\{g(x) + \eta\} d\Lambda(x)$$

and the η -entropy

$$H_{\eta}(f) = \int \{f(x) + \eta\} \log\{f(x) + \eta\} d\Lambda(x),$$

so that the η -divergence is $D_{\eta}(f, g) = C_{\eta}(f, g) - H_{\eta}(f)$, see [24–27] for applications for pattern recognition. Obviously, if we take a limit of η to 0, then $C_{\eta}(f, g)$, $H_{\eta}(f)$ and $D_{\eta}(f, g)$ converge to $C_0(f, g)$, $H_0(f)$ and $D_0(f, g)$, respectively. A mislabeled model is derived by a maximum η -entropy distribution with momentary constraint if we consider a binary regression model. See [25,27] for a detailed discussion.

3. Geometry Associated with U -Divergence

We investigate geometric properties associated with U -divergence, which will help the discussion in subsequent sections. Let us arbitrarily fix a statistical model $M = \{f_\theta(x) : \theta \in \Theta\}$ embedded in the total space \mathcal{F} with mild regularity conditions. In fact, we consider the mixture geodesic curve $C^{(m)}$, the exponential geodesic curve $C^{(e)}$, the mixture model $M^{(m)}$ and the exponential model $M^{(e)}$ as typical examples of M . Here are difficult aspects to define \mathcal{F} as a differentiable manifold of infinite dimension because the constraint for positivity on the support is intractable in the sense of the topology, see Section 2 in [6] for detailed discussion and historical remarks. On the other hand, if we confine ourselves to a statistical model M , then we can formulate M as a finite dimensional manifold, as in the following discussion. Thus, we produce a path geometry in which for any two elements f and g of \mathcal{F} a class of geodesic curves connecting f and g including $C^{(m)}$ and $C^{(e)}$ is introduced so that the class of geodesic subspaces is derived as for $M^{(m)}$ and $M^{(e)}$.

3.1. Riemannian Metric and Linear Connections

We view the statistical model M as a d -dimensional differentiable manifold with the coordinate $\theta = (\theta^1, \dots, \theta^d)$. Any information divergence associates with a Riemannian metric and dual linear connections, see [28,29] for detailed discussion. We focus on the geometry generated by the U -divergence $D_U(f, g)$ as follows. The Riemannian metric at f_θ of M is given by

$$G_{ij}^{(U)}(\theta) = - \int \partial_i f_\theta \partial_j \xi(f_\theta) d\Lambda, \tag{10}$$

and linear connections are

$$\Gamma_{ij,k}^{(U)}(\theta) = - \int \partial_i \partial_j f_\theta \partial_k \xi(f_\theta) d\Lambda \tag{11}$$

and

$$*\Gamma_{ij,k}^{(U)}(\theta) = - \int \partial_k f_\theta \partial_i \partial_j \xi(f_\theta) d\Lambda, \tag{12}$$

where $\partial_i = \partial/\partial\theta^i$, see Appendix. for the derivation. Now we can assert the following theorem under an assumption for \mathcal{F} : Let f be arbitrarily fixed in \mathcal{F} . If $\int a(x)\{g(x) - f(x)\}d\Lambda(x) = 0$ for any g of \mathcal{F} , then $a(x)$ is constant in x almost everywhere with respect to Λ .

Theorem 1. Let $\Gamma^{(U)}$ be the linear connection defined in Equation (11). Then any $\Gamma^{(U)}$ -geodesic curve is equal to the mixture-geodesic curve defined in Equation (2).

Proof. Let $C^{(U)} := \{f_t(x) : t \in (0, 1)\}$ be a $\Gamma^{(U)}$ -geodesic curve with $f_0 = f$ and $f_1 = g$. We consider a 2-dimensional model defined by $f_\theta(x) = (1 - s + u)f_t(x) + (s - u)g(x)$, where $\theta = (s, t, u)$. Then we observe that if $u = s$, then

$$\Gamma_{11,2}^{(U)}(\theta) = - \int \left(\frac{d^2}{dt^2} f_t \right) \xi'(f_t)(g - f_t) d\Lambda \tag{13}$$

which identically 0 for any g of \mathcal{F} . It follows from the assumption for \mathcal{F} that $(d^2/dt^2)f_t(x) = c$ almost everywhere with respect to Λ , which solved by

$$f_t(x) = \frac{1}{2}ct(t - 1) + (1 - t)f(x) + tg(x)$$

from the endpoint condition for $C^{(U)}$. We observe that $c = 0$ because $f_t(x) \in \mathcal{F}$, which concludes that $C^{(U)}$ equals the mixture-geodesic. The proof is complete.

This property is elemental to characterize the U -divergence class, which is closely related with the empirical reducibility as discussed in a subsequent section. The assumption for \mathcal{F} holds if the carrier measure Λ is Lebesgue measure or the counting measure.

On the other hand, for a ${}^*\Gamma^{(U)}$ -geodesic curve ${}^*C^{(U)} := \{f_t^*(x) : t \in (0, 1)\}$ with $f_0 = f$ and $f_1 = g$ we consider an embedding into a 2-dimensional model,

$$f_\theta^*(x) = u((1 - s + t)\xi(f_t^*(x)) + (s - t)\xi(g(x)) - \kappa_\theta),$$

where $\theta = (s, t)$, where $u(s) = (d/dt)U(s)$ and κ_θ is a normalizing constant to satisfy $\int f_\theta^*(x)d\Lambda(x) = 1$. By definition

$${}^*\Gamma_{11,2}^{(U)}(\theta) = \int \left(\frac{d^2}{dt^2}\xi(f_t^*) \right) u'(\xi(f_t^*))\{\xi(g) - \xi(f_t^*)\}d\Lambda = 0 \tag{14}$$

if $s = t$. This leads to $(d^2/dt^2)\xi(f_t^*(x)) = c$ almost everywhere with respect to Λ , which is solved by

$$f_t^*(x) = u((1 - t)\xi(f(x)) - t\xi(g(x)) - \kappa_t), \tag{15}$$

We confirm that, if $U = \exp$, then ${}^*\Gamma^{(U)}$ -geodesic curve reduces to the exponential geodesic curve defined in Equation (3). \square

3.2. Generalized Pythagorean Theorems

We next consider the Pythagorean theorem based on the U -divergence as an extension of the result associated with the Kullback-Leibler divergence in [6].

Theorem 2. *Let p, q and r be in \mathcal{F} . We connect p with q by the mixture geodesic*

$$f_t^{(m)}(x) = (1 - t)p(x) + tq(x),$$

Alternatively we connect r and q by ${}^\Gamma^{(U)}$ -geodesic curve*

$$f_s^{(U)}(x) = u((1 - s)\xi(r(x)) + s\xi(q(x)) - \kappa(s)).$$

Two curves $\{f_t^{(m)}(x) : t \in [0, 1]\}$ and $\{f_s^{(U)}(x) : s \in [0, 1]\}$ orthogonally intersect at q with respect to the Riemannian metric $G^{(U)}$ defined in Equation (10) if and only if

$$D_U(p, r) = D_U(p, q) + D_U(q, r). \tag{16}$$

Proof. A straightforward calculus yields that

$$-\frac{\partial^2}{\partial t \partial s} D_U(f_t^{(m)}, f_s^{(U)}) \Big|_{t=1, s=1} = D_U(p, r) - \{D_U(p, q) + D_U(q, r)\}. \tag{17}$$

By the definition of $G^{(U)}$ we see that $G_{12}^{(U)}(\theta)$ is nothing but the left side of Equation (17) when

$$f_\theta(x) = (1 - t)p(x) + tf_s^{(U)}(x),$$

where $\theta = (t, s)$. Hence the orthogonality assumption is equivalent to Equation (16), which completes the proof. \square

Remark 1. We remark a further property such that, for any s and t in $[0, 1]$,

$$D_U(p_t, r) = D_U(p_t, q) + D_U(q, r_s).$$

If $U = \exp$, then Theorem 2 reduces to the Pythagoras theorem with the Kullback-Leibler divergence as shown in [6]. Consider two geodesic subspaces defined by

$$M^{(m)} = \{p_\pi(x) = \pi_0 q(x) + \sum_{j=1}^J \pi_j p_j(x) : \pi_j \geq 0 (j = 0, \dots, J), \sum_{j=0}^J \pi_j = 1\}$$

and

$$M^{(U)} = \{r_\epsilon(x) = u\left(\epsilon_0 \xi(q(x)) + \sum_{k=1}^K \epsilon_k \xi(r_k(x)) - \kappa(\epsilon)\right) : \epsilon_k \geq 0 (k = 0 \dots, K), \sum_{k=0}^K \epsilon_k = 1\}. \tag{18}$$

For any m -geodesic curve $C^{(m)}$ and U -geodesic curve $*C^{(U)}$ connecting q we assume that $C^{(m)}$ and $C^{(U)}$ orthogonally intersect at q in the sense of the Riemannian metric $G^{(U)}$. Then, for any $p \in M^{(m)}$ and $r \in M^{(U)}$

$$D_U(p, r) = D_U(p, q) + D_U(q, r),$$

in which two-way projection is associated with as

$$D_U(p, q) = \min_{r \in M_2} D_U(p, r) \quad \text{and} \quad D_U(q, r) = \min_{p \in M_1} D_U(p, r).$$

First we confirm a kind of reduction property for the Kullback-Leibler divergence to the framework in information geometry such that $(G^{(D_0)}, \Gamma^{(D_0)}, * \Gamma^{(D_0)}) = (G, \Gamma^{(m)}, \Gamma^{(e)})$, where G is the information metric. Second we return a case of the β -power divergence, which is reduced a special case of Theorem 2. Consider two curves $C^{(m)} = \{C_t^{(m)}(x) = (1 - t)p(x) + tq(x) : t \in [0, 1]\}$ and

$$C^{(\beta)} = \{C_s^{(\beta)}(x) = \{(1 - s)r(x)^\beta + tq(x)^\beta + c(s)\}^{\frac{1}{\beta}} : s \in [0, s]\}.$$

Then we observe for the Riemannian metric $G^{(\beta)}$ generated by β -power divergence that

$$G^{(\beta)}(\dot{C}_1^m, \dot{C}_1^\beta)(q) = D_\beta(p, r) - \{D_\beta(p, q) + D_\beta(q, r)\}, \tag{19}$$

which is $\int (p - q)(p^\beta - q^\beta) d\Lambda$. We observe that if $C^{(m)}$ and $C^{(\beta)}$ orthogonally intersect at q , then

$$D_\beta(p, r) = D_\beta(p, q) + D_\beta(q, r).$$

4. Maximum Entropy Distribution

The maximum entropy principle is based on the Boltzmann-Shannon entropy in which the maximum entropy distribution is characterized by an exponential model. The maximum entropy method has been widely enhanced in fields of natural language processing, ecological analysis and so forth. However, there are other types of entropy measures proposed as the Hill diversity index, the Gini-Simpson index, the Tsallis entropy and so on, *cf.* [14,17,18] in different fields. We introduced the class of U -entropy functionals, which include all the entropy measures mentioned above. In this subsection, we discuss the maximum entropy distribution based on an arbitrarily fixed U -entropy.

We check a finite discrete case with $K + 1$ cells as a special situation, where \mathcal{F} reduces to a K -dimensional simplex \mathcal{S}_K . The maximum U -entropy distribution is defined by

$$f^* = \operatorname{argmax}_{f \in \mathcal{S}_K} H_U(f).$$

The Lagrange function is

$$L(f, \lambda) = \sum_{i=1}^{K+1} \{-\xi(f_i)f_i + U(\xi(f_i))\} + \lambda \left(\sum_{i=1}^{K+1} f_i - 1 \right).$$

We observe that

$$\frac{\partial}{\partial f_i} L(f, \lambda) = -\xi(f_i) + \lambda = 0,$$

which implies $f_i^* = 1/(K + 1)$ for $i = 1, \dots, K + 1$. Therefore the maximum U -entropy distribution f^* is a uniform distribution on \mathcal{S}_K for any generator function U .

In general the U -entropy is an unbounded functional on \mathcal{F} unless \mathcal{F} is finite discrete. For this we introduce a moment constraint as follows. Let $t(X)$ be a k -dimensional statistic vector. Henceforth we assume that $\mathbb{E}_f\{\|t(X)\|^2\} < \infty$ for all f of \mathcal{F} . We consider a mean equal space for $t(X)$ as

$$\Gamma(\tau) = \{f \in \mathcal{F} : \mathbb{E}_f\{t(X)\} = \tau\},$$

where τ is a fixed vector in \mathbb{R}^k . By definition $\Gamma(\tau)$ is totally mixture geodesic, that is, if f and g are in $\Gamma(\tau)$, then $(1 - t)f + tg$ is also in $\Gamma(\tau)$ for any $t \in (0, 1)$.

Theorem 3. *Let $f_\tau^* = \operatorname{argmax}\{H_U(f) : f \in \Gamma(\tau)\}$, where $H_U(f)$ is U -diagonal entropy defined in Equation (7). Then the maximum U -entropy distribution is given by*

$$f_\tau^*(x) = u(\theta^\top t(x) - \kappa_U(\theta)), \tag{20}$$

where $\kappa_U(\theta)$ is the normalizing factor and θ is a parameter vector determined by the moment constraint

$$\int t(x)u(\theta^\top t(x) - \kappa_U(\theta))d\Lambda(x) = \tau.$$

Proof. The Euler-Lagrange functional is given by

$$\Phi(f, \theta, \lambda) = H_U(f) - \theta^\top [\mathbb{E}_f\{t(X)\} - \tau] - \lambda \left\{ \int f(x)d\Lambda(x) - 1 \right\}$$

If $g_\tau \in \Gamma(\tau)$ and $f_t(x) = (1 - t)f_\tau^*(x) + tg_\tau(x)$, then $f_t \in \Gamma(\tau)$, and

$$\frac{d}{dt}\Phi(f_t, \theta, \lambda)\Big|_{t=0} = 0, \quad \frac{d^2}{dt^2}\Phi(f_t, \theta, \lambda)\Big|_{t=0} < 0. \tag{21}$$

The equation in Equation (21) yields that

$$\int \{\xi(f_\tau^*(x)) - \theta^\top(t(x) - \tau) - \lambda\}\{g(x) - f_\tau^*(x)\}d\Lambda(x) = 0$$

for any $g_\tau(x)$ in $\Gamma(\tau)$, which concludes Equation (20). Since $\xi(t)$ is an increasing function, we observe that

$$\frac{d^2}{dt^2}\Phi(f_t, \theta, \lambda) = - \int \xi'(f_t(x))\{g(x) - f_\tau^*(x)\}^2d\Lambda(x) < 0 \tag{22}$$

for any $t \in [0, 1]$, which implies the inequality in Equation (21). Since $g_\tau \in \Gamma(\tau)$, we observe that

$$\mathbb{E}_{g_\tau}\{\xi(f_\tau^*(X))\} = \mathbb{E}_{f_\tau^*}\{\xi(f_\tau^*(X))\}$$

Therefore we can confirm that $H_U(f_\tau^*) \geq H_U(g_\tau)$ for any $g_\tau \in \Gamma(\tau)$ since

$$H_U(f_\tau^*) - H_U(g_\tau) = D_U(g_\tau, f_\tau^*),$$

which is nonnegative by the definition of U -divergence. The proof is complete. \square

Here we give a definition of the model of maximum U -entropy distributions as follows.

Definition 1. We define a k -dimensional model

$$M_U = \{f_U(x, \theta) := u(\theta^\top t(x) - \kappa_U(\theta)) : \theta \in \Theta\}, \tag{23}$$

which is called U -model, where $\Theta = \{\theta \in \mathbb{R}^k : \kappa_U(\theta) < \infty\}$.

The Naudts' deformed exponential family discussed from a statistical physical viewpoint as in [15] is closely related with U -model. The one-parameter family $\{r_s(x) : s \in [0, 1]\}$ as defined in Equation (15) is a one-dimensional U -model and $M^{(U)}$ defined in Equation (18) is a K -dimensional U -model. For a U -model M_U defined in Equation (23), the parameter θ is an affine parameter for the linear connection ${}^*\Gamma^{(U)}$ defined in Equation (12). In fact, we observe from the definition Equation (12) that

$${}^*\Gamma_{ij,k}^{(U)}(\theta) = \partial_j \partial_k \kappa_U(\theta) \int \partial_k f_U(\theta, x) d\Lambda(x)$$

which is identically 0 for all $\theta \in \Theta$. We have a geometric understanding for the U -model similar to the exponential model discussed in Introduction.

Theorem 4. Assume for $U(t)$ that $U'''(t) > 0$ for any t in \mathbb{R} . Then, the U -model is totally ${}^*\Gamma^{(U)}$ -geodesic.

Proof. For arbitrarily fixed θ_1 and θ_2 in Θ , we define the U -geodesic curve connecting between $f_U(x, \theta_1)$ and $f_U(x, \theta_2)$ such that, for $\lambda \in (0, 1)$,

$$f_\lambda(x) = u(\lambda\xi(f_U(x, \theta_1)) + (1 - \lambda)\xi(f_U(x, \theta_2)) - \kappa(\lambda))$$

with a normalizing factor $\kappa(\lambda)$, which is written by $f_\lambda(x) = f_U(x, \theta_\lambda)$, where $\theta_\lambda = \lambda\theta_1 + (1 - \lambda)\theta_2$. Hence it suffices to show $\theta_\lambda \in \Theta$ for all $\lambda \in (0, 1)$, where Θ is defined in Definition 1. We look at the identity $\int f_U(x, \theta)d\Lambda(x) = 1$ from a fact that $f_U(x, \theta)$ is a probability density function. This implies that the first derivative gives

$$\int u'(\theta^\top t(x) - \kappa_U(\theta)) \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\} d\Lambda(x) = 0$$

and the second derivative gives

$$\begin{aligned} & \int u''(\theta^\top t(x) - \kappa_U(\theta)) \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\} \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\}^\top d\Lambda(x) \\ & - \int u'(\theta^\top t(x) - \kappa_U(\theta)) d\Lambda(x) \frac{\partial^2}{\partial \theta \partial \theta^\top} \kappa_U(\theta) = 0 \end{aligned} \tag{24}$$

Since the identity Equation (24) shows that the Hessian of $\kappa_U(\theta)$ is proportional to a Gramian matrix, which implies that $\kappa_U(\theta)$ is convex in θ . Since $\kappa_U(\theta_\lambda) \leq (1 - \lambda)\kappa_U(\theta_1) + \lambda\kappa_U(\theta_2)$ and θ_1 and θ_2 in Θ , $\kappa_U(\theta_\lambda) \leq \infty$. This concludes that $\theta_\lambda \in \Theta$ for any $\lambda \in (0, 1)$, which completes the proof. \square

We discuss a typical example by the power entropy $H_\beta(f)$, see [15,30–34] from a viewpoint of statistical physics. First we consider a mean equal space of univariate distributions on $(0, \infty)$

$$\Gamma(\mu) = \{f : \mathbb{E}_f\{t(X)\} = \mu\}$$

where

$$t(x) = \left(x, \frac{x^{\beta(\kappa-1)} - 1}{\beta} \right)^\top$$

Note that $\lim_{\beta \rightarrow 0} t(x) = (x, (\kappa - 1) \log x)$. To get the maximum entropy distribution with H_β we consider the Euler-Lagrange function given by

$$E_\beta(f, \lambda) = \frac{1}{\beta(\beta + 1)} \int_0^\infty f(x)^{1+\beta} dx + \theta^\top \left\{ \int_0^\infty t(x)f(x)dx - \mu \right\} + \lambda \left\{ \int_0^\infty f(x)dx - 1 \right\},$$

where θ and λ are Lagrange multiplier parameters. This yields that the maximum entropy distribution is

$$\begin{aligned} f_\beta(x, \theta) &= Z_\beta(\theta)^{-1} (1 + \beta \theta^\top t(x))^{-\frac{1}{\beta}} \\ &= Z_\beta(\theta)^{-1} (\beta \theta_1 x + \theta_2 x^{\beta(\kappa-1)})^{-\frac{1}{\beta}} \\ &= Z_\beta(\theta)^{-1} x^{\kappa-1} (\theta_2 - \beta \theta_1 x^{1-\beta(\kappa-1)})^{-\frac{1}{\beta}}, \end{aligned}$$

where θ is determined by μ such that $\mathbb{E}_{f_\beta(\cdot, \theta)} t(X) = \mu$ and

$$Z_\beta(\theta) = \int_0^\infty x (\theta_2 - \beta \theta_1 x^{1-\beta})^{-\frac{1}{\beta}} dx.$$

A gamma distribution is defined by the density function

$$f(x, \kappa, \theta) = \frac{x^{\kappa-1} \exp(-\frac{x}{\theta})}{\Gamma(\kappa)\theta^\kappa}$$

Second, we consider a case of multivariate distributions, where the moment constraints are supposed that for a fixed p -dimensional vector μ and matrix V of size $p \times p$

$$\Gamma(\mu, V) = \{f \in \mathcal{F} : \mathbb{E}_f(X) = \mu, \mathbb{V}_f(X) = V\}.$$

Let

$$f_\beta(\cdot, \mu, V) = \operatorname{argmax}_{f \in \Gamma(\mu, V)} H_\beta(f).$$

If we consider a limit case of β to 0, then $H_\beta(f)$ reduces to the Boltzmann-Shannon entropy and the maximum entropy distribution is the Gaussian distribution with the density function

$$\varphi(x, \mu, V) = \{\det(2\pi V)\}^{p/2} \exp\left\{-\frac{1}{2}(x - \mu)^\top V^{-1}(x - \mu)\right\}.$$

In general we deduce that if $\beta > -\frac{2}{p+2}$, then the maximum β -power entropy distribution uniquely exists such that the density function is given by

$$f_\beta(x, \mu, V) = \frac{c_\beta}{\det(2\pi V)^{\frac{1}{2}}} \left\{1 - \frac{\beta}{2 + p\beta + 2\beta}(x - \mu)^\top V^{-1}(x - \mu)\right\}_+^{\frac{1}{\beta}},$$

where

$$c_\beta = \begin{cases} \left(\frac{2\beta}{2 + p\beta + 2\beta}\right)^{\frac{p}{2}} \Gamma\left(1 + \frac{p}{2} + \frac{1}{\beta}\right) \{\Gamma(1 + \frac{1}{\beta})\}^{-1} & \text{if } \beta \geq 0 \\ \left(\frac{-2\beta}{2 + p\beta + 2\beta}\right)^{\frac{p}{2}} \Gamma\left(-\frac{1}{\beta}\right) \{\Gamma(-\frac{1}{\beta} - \frac{p}{2})\}^{-1} & \text{if } -\frac{2}{p+2} < \beta \leq 0 \end{cases}$$

See [35,36] for the detailed discussion [37,38] for the discussion on group invariance. Thus, if $\beta > 0$, then the maximum β -power entropy distribution has a compact support

$$\{x \in \mathbb{R}^p : (x - \mu)^\top V^{-1}(x - \mu) \leq \frac{2}{\beta} + p + 2\}$$

The typical case is $\beta = 2$, which is called the Wigner semicircle distribution. On the other hand, if $-\frac{2}{p+2} < \beta < 0$, the maximum β -power entropy distribution has a full support of \mathbb{R}^p , and equals a p -variate t -distribution with a degree of freedom depending on β .

5. Minimum Divergence Method

We have shown a variety of U -divergence functionals using various generator functions in which the minimum divergence methods are applied to analyses in statistics and statistical machine learning. In effect the U -cross entropy $C_U(f, g)$ is convex-linear in f , that is,

$$C_U\left(\sum_{j=1}^J \lambda_j f_j, g\right) = \sum_{j=1}^J \lambda_j C_U(f_j, g)$$

for any $\lambda_j > 0$ with $\sum_{j=1}^J \lambda_j = 1$. It is closely related with a characteristic property that the linear connection $\Gamma^{(U)}$ associated with U -divergence is equal to the mixture connection $\Gamma^{(m)}$ as discussed in Theorem 1. Furthermore, for a fixed g , $C_U(f, g)$ can be viewed as a functional of F in place of f as follows:

$$C_U(F, g) = \int \left\{ \xi(g(x)) - \int U(\xi(g(x))) d\Lambda(x) \right\} dF(x),$$

where F is the probability distribution generated from $f(x)$. If we assume to have a random sequence X_1, \dots, X_n from a density function $f(x)$, then the U -cross entropy is approximated as

$$C_U(\bar{F}_n, g) = -\frac{1}{n} \sum_{i=1}^n \xi(g(X_i)) + \int U(\xi(g)) d\Lambda, \tag{25}$$

where \bar{F}_n is the empirical distribution based on the data X_1, \dots, X_n , that is, $\bar{F}_n(B) = \frac{1}{n} \sum_{i=1}^n I(X_i \in B)$ for any Borel measurable set B . By definition,

$$\int \xi(g(x)) d\bar{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \xi(g(X_i)).$$

Consequently, if we model g by a model function $f(\cdot, \theta)$, then the right side of Equation (25) depends only on the data set $(X_i)_{i=1}^n$ and parameter θ without any knowledge for the underlying density function $f(x)$. This gives the empirical approximation, which is advantageous over other classes of divergence measures. The minimum U -divergence method is directly applied to minimization of the empirical approximation with respect to θ . We note that the minimum divergence is equivalent to the minimum cross entropy, in which the diagonal entropy is just a constant in θ . In particular, in the classical case,

$$C_0(\bar{F}_n, f(\cdot, \theta)) = -\frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta) + 1,$$

which is equivalent to the minus log-likelihood function.

Let X_1, \dots, X_n be independently and identically distributed from an underlying density function $f(x)$ which is approximated by a statistical model $M = \{f(x, \theta) : \theta \in \Theta\}$. The U -loss function is introduced by

$$L_U(\theta) = -\frac{1}{n} \sum_{i=1}^n \xi(f(X_i, \theta)) + b_U(\theta),$$

where $b_U(\theta) = \int U(\xi(f(x, \theta))) d\Lambda(x)$. We call $\hat{\theta}_U = \operatorname{argmin}_{\theta \in \Theta} L_U(\theta)$ U -estimator for the parameter θ . By definition $\mathbb{E}_f\{L_U(\theta)\} = C_U(F, f(\cdot, \theta))$ for all θ in Θ , which implies that $L_U(\theta)$ almost surely converges to $C_U(F, f(\cdot, \theta))$ as n goes to ∞ . Let us define a statistical functional as

$$\theta_U(F) = \operatorname{argmin}_{\theta \in \Theta} C_U(F, f(\cdot, \theta)),$$

where $C_U(F, g)$ is written $C_U(f, g)$ placing f into the probability distribution F generated from f . Then $\theta_U(F)$ is model-consistent, or $\theta_U(F_\theta) = \theta$ for any $\theta \in \Theta$ because

$$C_U(F_\theta, f(\cdot, \theta')) \leq H_U(f(\cdot, \theta))$$

with equality if and if $\theta' = \theta$, where F_θ is the probability distribution induced form $f(x, \theta)$.

Hence U -estimator $\hat{\theta}_U$ is asymptotically consistent. The estimating function is given by

$$s_U(x, \theta) = \frac{\partial}{\partial \theta} \xi(f(x, \theta)) - \mathbb{E}_{f(\cdot, \theta)} \left\{ \frac{\partial}{\partial \theta} \xi(f(X, \theta)) \right\}. \quad (26)$$

Consequently we confirm that $s_U(x, \theta)$ is unbiased in the sense that $\mathbb{E}_{f(\cdot, \theta)} \{s_U(X, \theta)\} = 0$.

We next investigate the asymptotic normality for U -estimator. The estimating equation for the U -estimator is given by

$$\frac{1}{n} \sum_{i=1}^n s_U(X_i, \hat{\theta}_U) = 0,$$

of which the Taylor approximation gives

$$\frac{1}{n} \sum_{i=1}^n \left\{ s_U(X_i, \theta_U(F)) + \frac{\partial s_U}{\partial \theta^\top}(X_i, \theta_U) (\hat{\theta}_U - \theta_U(F)) \right\} = o(n_P^{-1}).$$

In accordance with this, we get the asymptotic approximation,

$$\sqrt{n} \{ \hat{\theta}_U - \theta_U(F) \} = \frac{1}{\sqrt{n}} J(\theta_U(F))^{-1} \sum_{i=1}^n s_U(X_i, \theta_U(F)) + o(n_P^{-\frac{1}{2}}),$$

where

$$J(\theta) = \mathbb{E}_{f(\cdot, \theta)} \left\{ \frac{\partial s_U}{\partial \theta^\top}(X, \theta) \right\}.$$

Because the strong law of large number gives

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial s_U}{\partial \theta^\top}(X_i, \theta_U(f)) \xrightarrow{\text{a.s.}} J(\theta_U(F))$$

as n goes to ∞ , where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence. If the underlying density function is in the model M , that is $f(x) = f(x, \theta)$, then it follows from the model consistency for $\theta_U(F)$ that

$$\sqrt{n}(\hat{\theta}_U - \theta) = \frac{1}{\sqrt{n}} J(\theta)^{-1} \sum_{i=1}^n s_U(X_i, \theta) + o(n_P^{-\frac{1}{2}}),$$

which implies that

$$\sqrt{n}(\hat{\theta}_U - \theta) \xrightarrow{D} N(0, J(\theta)^{-1} V(\theta) J(\theta)^{-1}),$$

where \xrightarrow{D} denotes convergence in distribution and

$$V(\theta) = \mathbb{V}_{f(\cdot, \theta)} \{s_U(X, \theta)\}.$$

If the generator function is taken as $U(s) = \exp(s)$, then the U -estimator reduces to the MLE with the asymptotic normality to $N(0, G(\theta)^{-1})$, where $G(\theta)$ is the Fisher information matrix for θ .

Consider U -estimator for the parameter θ of the exponential model $M^{(e)}$ in Equation (1). In particular we are concerned with a possible outlying contaminated in the exponential model, and hence a ϵ -contamination model is defined as

$$F_{\theta,\epsilon,y}(x) = (1 - \epsilon)F_0(x, \theta) + \epsilon\delta_y(x),$$

where $\epsilon, 0 < \epsilon < 1$ is a sufficiently small constant, $F_0(x, \theta)$ is the cumulative distribution function of the exponential model, and $\delta_y(x)$ denotes a degenerate distribution at y . The influence function for U -estimator is given by

$$\text{IF}(\hat{\theta}_U, y) := \lim_{\epsilon \rightarrow 0} \frac{\theta_U(F_{\theta,\epsilon,y}) - \theta}{\epsilon} = J(\theta)^{-1} s_U(y, \theta),$$

See [19,20,27]. Thus we can check the robustness for U -estimator whether the influence function is bounded in y or not. For example, if we adopt as $U(s) = (1 + \beta s)^{1/\beta}$, then

$$\text{IF}(\hat{\theta}_U, y) = J(\theta)^{-1} [\{t(y) - \mu\} f_0(y, \theta)^\beta - b(\theta, \beta)], \tag{27}$$

where $b(\theta, \beta) = \int \{t(x) - \mu\} f_0(x, \theta)^\beta d\Lambda(x)$. Thus, if $\beta > 0$, then the influence function is confirmed to be bounded in y for almost cases including a normal, exponential and Poisson distribution models since the term $\{t(y) - \mu\} f_0(y, \theta)^\beta$ in Equation (27) is bounded in y for these models. On the other hand, If $\beta = 0$, that is the maximum likelihood estimator entails the unbounded influence functions because the term $t(y) - \mu$ is unbounded in y for these models.

6. Duality of Maximum Entropy and Minimum Divergence

In this section, we discuss a dualistic interplay between statistical model and estimation. In statistical literature, the maximum likelihood estimation has a special position over other estimation methods in the sense of efficiency, invariance and sufficiency; while the statistical model has been explored various candidates in the presence of misspecification. For example, we frequently consider a Laplace distribution for estimating a Gaussian mean, which leads to the sample median as the maximum likelihood estimator for the mean of the Laplace distribution. In this sense, there is an unbalance in the employment for the model and estimator. In principle, we can select arbitrarily different generator functions U_0 and U_1 so that the U_1 -estimation gives consistency under the U_0 -model. There is a natural question which situation happens if we consider the U -estimation under the U -model?

Let M_U be a U -model defined by

$$M_U = \{f_U(x, \theta) := u(\theta^\top t(x) - \kappa_U(\theta)) : \theta \in \Theta\}, \tag{28}$$

where $\Theta = \{\theta \in \mathbb{R}^k : \kappa_U(\theta) < \infty\}$. The the U -loss function under the U -model for a given data set $\{X_1, \dots, X_n\}$ is defined by

$$L_U(\theta) = -\frac{1}{n} \sum_{i=1}^n \xi(f_U(X_i, \theta)) + \int U(\xi(f_U(x, \theta))) d\Lambda(x),$$

which is reduced to

$$L_U(\theta) = -\theta^\top \bar{t} + \kappa_U(\theta) + b_U(\theta), \tag{29}$$

where $\bar{t} = \frac{1}{n} \sum_{i=1}^n t(X_i)$ and

$$b_U(\theta) = \int U(\xi(\theta^\top t(x) - \kappa_U(\theta)))d\Lambda(x). \tag{30}$$

The estimating equation is given by

$$\frac{\partial}{\partial \theta} L_U(\theta) = -\bar{t} + \frac{\partial}{\partial \theta} \kappa_U(\theta) + \frac{\partial}{\partial \theta} b_U(\theta),$$

which is written by

$$\frac{\partial}{\partial \theta} L_U(\theta) = -\bar{t} + \mathbb{E}_{f(\cdot, \theta)}\{t(X)\}.$$

Hence, if we consider the U -estimator for a parameter η by the transformation of θ defined by $\varphi(\theta) = \mathbb{E}_{f(\cdot, \theta)}\{t(X)\}$, then the U -estimator $\hat{\eta}_U$ is nothing but the sample mean \bar{t} . Here we confirm that the transformation $\varphi(\theta)$ is one-to-one as follows. The Jacobian matrix of the transformation is given by

$$\frac{\partial}{\partial \theta} \varphi(\theta) = \int u'(\theta^\top t(x) - \kappa_U(\theta)) \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\} \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\}^\top d\Lambda(x),$$

since the first identity for M_U leads to

$$\frac{\partial}{\partial \theta} \int f_U(x, \theta) d\Lambda(x) = \int u'(\theta^\top t(x) - \kappa_U(\theta)) \left\{ t(x) - \frac{\partial}{\partial \theta} \kappa_U(\theta) \right\} d\Lambda(x) = 0.$$

Therefore, we conclude that the Jacobian matrix is symmetric and positive-definite since $u'(t)$ is a positive function from the assumption of the convexity for U , which implies that $\varphi(\theta)$ is one-to-one. Consequently, the estimator $\hat{\theta}_U$ for θ is given by $\varphi^{-1}(\bar{t})$. We summarize these results in the following theorem.

Theorem 5. *Let M_U be a U -model with a canonical statistic $t(X)$ as defined in Equation (28). Then the U -estimator for the expectation parameter η of $t(X)$ is always \bar{t} , where $\bar{t} = \frac{1}{n} \sum_{i=1}^n t(X_i)$.*

Remark 2. *We remark that the empirical Pythagorean theorem holds as in*

$$L_U(\theta) = L_U(\hat{\theta}_U) + D_U(\hat{\theta}_U, \theta),$$

since we observe that

$$L_U(\theta) - L_U(\hat{\theta}_U) = (\hat{\theta}_U - \theta)^\top \bar{t} + \kappa_U(\theta) + b_U(\theta) - \kappa_U(\hat{\theta}_U) + b_U(\hat{\theta}_U),$$

which gives another proof for which $\hat{\theta}_U$ is $\varphi^{-1}(\bar{t})$. The statistic \bar{t} is a sufficient statistic in the sense that the U -loss function $L_U(\theta)$ is a function of \bar{t} as in Equation (29). Accordingly, the U -estimator under U -model is a function only of \bar{t} from the observations X_1, \dots, X_n . In this extension, the MLE is a function of \bar{t} under the exponential model with the canonical statistic $t(X)$.

Let us look at the case of the β -power divergence. Under the β -power model given by

$$M_\beta = \{f_\beta(x, \theta) := \{\kappa_\beta(\theta) + \beta\theta^\top t(x)\}^{\frac{1}{\beta}} : \theta \in \Theta\},$$

the β -loss function is written by

$$L_\beta(\theta) = -\beta\theta^\top \bar{t} + \kappa_\beta(\theta) + b_\beta(\theta),$$

where

$$b_\beta(\theta) = \frac{1}{\beta + 1} \int \{\kappa_\beta(\theta) + \beta\theta^\top t(x)\}^{\frac{1+\beta}{\beta}} d\Lambda(x).$$

The β -power estimator for the expectation parameter of $t(X)$ is exactly given by \bar{t} .

7. Discussion

We concentrate on elucidating the dual structure of the U -estimator under the U -model, in which the perspective extends the relation of the maximum likelihood under the exponential model with a functional degree of freedom. Thus, we explore a rich and practical class of duality structures; however, there remains an unsolved problem when we directly treat the space \mathcal{F} as an differentiable manifold, see [39] for an infinite dimensional exponential family. The approach here is not a direct extension of an infinite dimensional manifold, but a path geometry in the following sense. For all pairs of elements of \mathcal{F} the geodesic curve connecting the pair is represented in an explicit form in the class of ${}^*\Gamma^{(U)}$ connections in our context.

The U -divergence approach was the first trial to introduce a dually flat structure to \mathcal{F} which is different from the alpha-geometry. However, there are many related studies. For example, a nonparametric information geometry on the space of all functions without constraints for positivity and normalizing is discussed in Zhang [40]. Amari [41] characterizes (ρ, τ) -divergence with decomposable dually flat structure, see also [42]. If ρ is an identity function and $\tau(s) = (d/ds)U(s)$, (ρ, τ) -divergence is no less than U -divergence. In effect we confine ourselves to discussing the U -divergence class for the sake of the direct estimability for U -estimator.

The duality between the maximum entropy and the minimum divergence has been explored in the minimax theorem for a zero-sum game between a decision maker and Nature. The pay-off function is taken by the cross U -entropy in which Nature tries to maximize the pay-off function under the mean equal constraint; the decision maker tries to minimize the pay-off function. The equilibrium is given by the minimax solution, which is the maximum U -entropy distribution, see [43] for the extensive discussion and the relation with Bayesian robustness. The observation explored in this paper is closely related with this minimax argument, however the duality between the statistical model and estimation is focused on, where the minimum U -divergence leads to projection onto the U -model.

In principle, the U -estimator is applicable for all the statistical model since U -loss function is written by a sample as well as the log-likelihood function. If the choice of the model is different from the U -model, then U -estimator has different performance from the present situation. For example, we consider an exponential model ($U(s) = \exp(s)$), and a β -estimator ($U(s) = (1 - \beta s)^{1/\beta}$) for getting a robustness property for outlying observations, cf. [19,20]. In such situations, the duality property is no longer valid, since the β -estimator for the parameter of the exponential model is not a function of the sufficient statistic \bar{t} defined in Theorem 5. Thus, we have to pay attention to another aspect than the duality structure in the presence of outlying, or misspecification for the statistical model. Furthermore,

another type of divergence measures including projective power divergence is recommended to perform super robustness, *cf.* [21,44].

We presented the method of generalized maximum entropy based on the proposed entropy measure, as an extension of the classical maximum entropy method based on the Boltzmann-Gibbs-Shannon entropy. Practical applications of MaxEnt are actively followed in ecological and computational linguistic researches based on the classical maximum entropy, *cf.* [45,46]. Difficult aspects are discussed, in which the MaxEnt is apt to be over-learning on data sets because it basically employs the maximum likelihood estimator. There is a great potential for the proposed method to implement these research fields in order to overcome these difficult aspects, by selecting an appropriate generator function. A detailed discussion is beyond the scope of the present paper; however, it will be challenged in the near future with concrete objectives motivated by real data analysis.

Acknowledgments

We thank to anonymous referees for their useful comments and suggestions for our revision. Shinto Eguchi and Osamu Komori were supported by Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology (CREST).

Author Contributions

Atsumi Ohara and Shinto Eguchi contributed to differential geometric parts associated with minimum divergence, and Osamu Komori and Shinto Eguchi contributed to statistical discussion for the maximum entropy model and minimum divergence estimation.

Appendix: Derivation for $G^{(U)}$, $\Gamma^{(U)}$ and $*\Gamma^{(U)}$

We apply the general formula for the Riemannian metric and the pair of linear connections discussed in [29] to U -divergence $D_U(f, g)$. The Riemannian metric is defined by

$$G_{ij}^{(U)}(\theta) = \frac{\partial^2}{\partial\theta^i\partial\theta^j} D_U(f_\theta, f_{\theta_1}) \Big|_{\theta_1=\theta}.$$

Hence $G_{ij}^{(U)}(\theta)$ is expressed by Equation (10). Next the pair of linear connections $\Gamma^{(U)}$ and $*\Gamma^{(U)}$ are defined by

$$\Gamma_{ij,k}^{(U)}(\theta) = \frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k} D_U(f_\theta, f_{\theta_1}) \Big|_{\theta_1=\theta}.$$

and

$$*\Gamma_{ij,k}^{(U)}(\theta) = \frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k} D_U(f_{\theta_1}, f_\theta) \Big|_{\theta_1=\theta}$$

which means Equations (11) and (12), respectively. We confirm the formula for $G^{(U)}$, $\Gamma^{(U)}$ and $*\Gamma^{(U)}$.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Fisher, R.A. On an Absolute Criterion for Fitting Frequency Curves. *Messenger Math.* **1912**, *41*, 155–160.
2. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **1922**, *222*, 309–368.
3. Jaynes, E.T. Information Theory and Statistical Mechanics. In *Statistical Physics*; Ford, K., Ed.; Benjamin: New York, NY, USA, 1963.
4. Barndorff-Nielsen, O. *Information and Exponential Families in Statistical Theory*; John Wiley: Chichester, UK, 1978.
5. Amari, S. *Differential-Geometrical Methods in Statistics*; Lecture Notes in Statistics, 28; Springer: New York, NY, USA, 1985.
6. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
7. Eguchi, S. Information divergence geometry and the application to statistical machine learning. In *Information Theory and Statistical Learning*; Emmert-Streib, F., Dehmer, M., Eds.; Springer US: New York, NY, USA, 2008; pp. 309–332.
8. Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217.
9. Barndorff-Nielsen, O.E.; Jupp, P.E. Statistics, yokes and symplectic geometry. *Ann. Fac. Sci. Toulouse Math.* **1997**, *3*, 389–427.
10. Scharf, L.L. *Statistical Signal Processing*; Addison-Wesley: Reading, MA, USA, 1991; Volume 98.
11. Févotte, C.; Bertin, N.; Durrieu, J.-L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
12. Cichocki, A.; Amari, S.I. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
13. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
14. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
15. Naudts, J. *Generalized Thermostatistics*; Springer: New York, NY, USA, 2011.
16. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics*; Springer: New York, NY, USA, 2009.
17. Simpson, E.H. Measurement of diversity. *Nature* **1949**, *163*, 688.
18. Hill, M.O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **1973**, *54*, 427–432.
19. Minami, M.; Eguchi, S. Robust blind source separation by beta divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
20. Fujisawa, H.; Eguchi, S. Robust estimation in the normal mixture model. *J. Stat. Plan. Inference* **2006**, *136*, 3989–4011.

21. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
22. Notsu, A.; Komori, O.; Eguchi, S. Spontaneous clustering via minimum gamma-divergence. *Neural Comput.* **2014**, *26*, 421–448.
23. Cichocki, A.; Cruces, S.; Amari, S. Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. *Entropy* **2011**, *13*, 134–170.
24. Eguchi, S.; Copas, J. A class of logistic-type discriminant functions. *Biometrika* **2002**, *89*, 1–22.
25. Takenouchi, T.; Eguchi, S. Robustifying AdaBoost by adding the naive error rate. *Neural Comput.* **2004**, *16*, 767–787.
26. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
27. Eguchi, S. Information geometry and statistical pattern recognition. *Sugaku Expo. Amer. Math. Soc.* **2006**, *19*, 197–216.
28. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
29. Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J* **1992**, *22*, 631–647.
30. Naudts, J. The q -exponential family in statistical Physics. *Cent. Eur. J. Phys.* **2009**, *7*, 405–413.
31. Naudts, J. Generalized exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
32. Ohara, A.; Wada, T. Information geometry of q -Gaussian densities and behaviors of solutions to related diffusion equations. *J. Phys. A: Math. Theor.* **2010**, doi:10.1088/1751-8113/43/3/035002.
33. Suyari, H. Mathematical structures derived from the q -multinomial coefficient in Tsallis statistics. *Phys. A: Stat. Mech. Appl.* **2006** *368*, 63–82.
34. Suyari, H.; Wada, T. Multiplicative duality, q -triplet and μ, ν, q -relation derived from the one-to-one correspondence between the (μ, ν) -multinomial coefficient and Tsallis entropy S_q . *Phys. A: Stat. Mech. Appl.* **2008**, *387*, 71–83.
35. Eguchi, S.; Kato, S. Entropy and divergence associated with power function and the statistical application. *Entropy* **2010**, *12*, 262–274.
36. Eguchi, S.; Komori, O.; Kato, S.; Projective Power Entropy and Maximum Tsallis Entropy Distributions. *Entropy* **2011**, *13*, 1746–1764.
37. Ohara, A.; Eguchi, S. Geometry on positive definite matrices deformed by V-potentials and its submanifold structure. In *Geometric Theory of Information*; Nielsen, F., Eds.; Springer: New York, NY, USA, 2014; Chapter 2, pp. 31–55.
38. Ohara, A.; Eguchi, S. Group invariance of information geometry on q -Gaussian distributions induced by beta-divergence. *Entropy* **2013**, *15*, 4732–4747.
39. Pistone, G.; Sempì, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **1995**, *33*, 1543–1561.
40. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on Statistical Manifolds. *Entropy* **2013**, *15*, 5384–5418.
41. Amari, S.-I. Information Geometry of Positive Measures and Positive-Definite Matrices: Decomposable Dually Flat Structure. *Entropy* **2014**, *16*, 2131–2145.

42. Harsha K.V.; Subrahmanian, M.K.S. F -Geometry and Amari's α -Geometry on a Statistical Manifold. *Entropy* **2014**, *16*, 2472–2487.
43. Grunwald, P.D.; Dawid, A.P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Stat.* **2004**, *32*, 1367–1433.
44. Chen, P.-W.; Hung, H.; Komori, O.; Huang, S.-Y.; Eguchi, S. Robust independent component analysis via minimum gamma-divergence estimation. *IEEE J. Sel. Top. Signal Process.* **2013**, *7*, 614–624.
45. Phillips, S.J.; Dudik, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **2008**, *31*, 161–175.
46. Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).