LETTER

Entropy, cross-entropy, relative entropy: Deformation theory[(a)]

To cite this article: J. Zhang and H. Matsuzoe 2021 *EPL* **134** 18001

A Letters Journal Exploring the Frontiers of Physics

April 2021

EPL, **134** (2021) 18001
doi: 10.1209/0295-5075/134/18001

www.epljournal.org

Focus Article

Entropy, cross-entropy, relative entropy: Deformation theory[(a)]

J. Zhang[1(b)] and H. Matsuzoe[2(c)]

[1] *University of Michigan - Ann Arbor, USA*
[2] *Nagoya Institute of Technology - Nagoya, Japan*

received 13 November 2020; accepted in final form 22 March 2021
published online 14 May 2021

Abstract – Attempts at generalizing Shannon entropy and Kullback-Leibler divergence (relative entropy) led to a plenthora of deformation models in theoretical physics, including q-model, κ-model, etc. Naudts and Zhang (*Inf. Geom.*, **1** (2018) 79) established that these models can be unified under two notions: deformed ϕ-exponential family (Naudts, J., *J. Inequal. Pure Appl. Math.*, **5** (2004) 102) and conjugate (ρ, τ)-embedding (Zhang J., *Neural Comput.*, **16** (2004) 159) of probability functions. Conjugate (ρ, τ)-embedding has a gauge freedom which, upon its fixing, subsumes the U-model of Eguchi (*Sugaku Expositions*, **19** (2006) 197) proposed in a statistical machine learning context. The generalization by (ρ, τ)-entropy, (ρ, τ)-cross-entropy, (ρ, τ)-divergence, when applied to the ϕ-exponential family, yields either a Hessian structure or a conformal Hessian structure under different gauge selections —this "splitting" is the hallmark when deforming the exponential family with its dually flat (Hessian) geometry. This letter provides a unified information geometric perspective of deformation of the exponential model, with calculations for Tsallis q-model.

Introduction. – Shannon entropy and Kullback-Leibler divergence (also known as relative entropy) are widely used in science and engineering. It is well-known, *e.g.*, [1], that the principle of maximum entropy (or of minimum Kullback-Leibler divergence) with linear constraints leads to an exponential family of probability density functions, and that the corresponding manifold of parametric density functions is the dually flat Hessian manifold —in such manifolds, the Riemannian metric is the second derivative ("Hessian") of a potential function, with respect to which two sets of parameters, called natural and expectation parameters, form biorthogonal coordinates.

Systematic efforts abound to generalize the aforementioned framework from Shannon entropy $S[P] = -\int P \log P$ (of a probability density function P) and Kullback-Leibler divergence or relative entropy $D[P, Q] = \int P \log(P/Q)$ (of two probability density functions P and Q) to more general analytic forms (here and below $[\cdot]$ indicates a functional). An obvious approach is to generalize the logarithmic function log used in these expressions, or equivalently the exponential function exp, as shown in the following examples.

Example 1 (α-embedding). *Amari's α-embedding function $l^{(\alpha)} : \mathbb{R}^+ \to \mathbb{R}$, is defined as* [1]

$$l^{(\alpha)}(u) = \begin{cases} \log u, & \alpha = 1, \\ \dfrac{2}{1-\alpha}\, u^{(1-\alpha)/2}, & \alpha \neq 1. \end{cases}$$

Example 2 (q-logarithmic embedding). *Tsallis introduced* [2] *the q-dependent entropy, $q \in \mathbb{R}, q \neq 1$ through the q-logarithmic/exponential functions* [3]*:*

$$\log_q(u) = \frac{1}{1-q}(u^{1-q} - 1), \quad \exp_q(u) = [1 + (1-q)u]^{1/(1-q)}.$$

The q-logarithm reduces to the standard logarithm as q tends to 1. Note that q-embedding and α-embedding functions are slightly different but related:

$$\log_q(u) = l^{(\alpha)}(u) - \frac{2}{1-\alpha}, \quad \alpha = 2q - 1.$$

Example 3 (κ-logarithmic embedding). *Kaniadakis* [4] *introduced the κ-model, where $\kappa \in \mathbb{R}, \kappa \neq 0$:*

$$\log_\kappa(u) = \frac{1}{2\kappa}(u^\kappa - u^{-\kappa}), \quad \exp_\kappa(u) = (\kappa u + \sqrt{1 + \kappa^2 u^2})^{\frac{1}{\kappa}}.$$

Taking $\lim_{\kappa \to 0}$ yields the standard exponential/logarithm.

To go beyond these "parametric" approaches of extending the exponential/logarithmic function, there have been several approaches that adopt a "deformation function", namely, to use a function class that includes exp and log as

[(a)] Contribution to the Focus Issue *Progress on Statistical Physics and Complexity* edited by Roberta Citro, Giorgio Kaniadakis, Claudio Guarcello, Antonio Maria Scarfone and Davide Valenti.

[(b)] E-mail: junz@umich.edu (corresponding author)

[(c)] E-mail: matsuzoe@nitech.ac.jp

a special member: the ϕ-deformed exponential approach by Naudts [5–7], the conjugate (ρ,τ)-embedding approach by Zhang [8–10], and the U-model by Eguchi [11,12]. The ϕ-model and U-model are both one-function models, while the (ρ,τ)-model uses two free functions. The aforementioned parametric deformation such as α-, q-, κ-models can be viewed as special cases of the function class models, with their explicit function forms as given by the specific α-, q-, κ-embedding functions.

It was eventually demonstrated in 2018 by Naudts and Zhang [13] that i) the ϕ- and U-model turned out to be equivalent; ii) they are special cases of the (ρ,τ)-model upon a particular fixing of the gauge freedom; iii) the corresponding (ρ,τ)-geometry of the manifold of ϕ-exponential family can have different appearances, a Hessian geometry (under one type of gauge selection) and a conformal Hessian geometry (under another type of gauge selection). The authors of ref. [13] unified a continuous thread of explorations of deformation models with intermediary results [14–16]. This unification preserves the rigid interlock of various ingredients listed below, and therefore represents a true deformation (*i.e.*, away from exp or log) to classic information geometric expressions and theory [1,17] of Shannon entropy, Kullback-Leibler divergence, and exponential family:

i) the function form of entropy, cross-entropy, and relative entropy (divergence);

ii) the function form of the probability family with corresponding normalization and potential, and the duality between the natural and expectation parameterization;

iii) the expression of the Riemannian metric (Fisher-Rao metric in general and Hessian metric in particular) and of the conjugate connections.

This article will present the information geometric perspective for deforming the exponential/logarithmic function for generalizing the Shannon entropy and Kullback-Leibler relative entropy and the resulting Riemannian geometry. The "conjugate (ρ,τ)-embedding" framework [13] will be presented, in which the novel concept of "gauge" freedom for deformation is used to uncover two types of gauge selections and to capture a "splitting" of the conventional Hessian (dually flat) structure into two conformally related structures for the non-exponential case.

Deforming exponential and logarithmic functions. – We describe two approaches to deformed exponential/logarithmic function that "mimick" nice properties of exp and log. One is by [5], based on the fact that

$$\frac{\mathrm{d}\log(t)}{\mathrm{d}t}=\frac{1}{t},\quad \log(t)=\int_1^t \frac{1}{s}\,\mathrm{d}s$$

and

$$\frac{\mathrm{d}\exp(t)}{\mathrm{d}t}=\exp(t),\quad \exp(t)=1+\int_0^t \exp(s)\,\mathrm{d}s.$$

The other is by [8], built upon the fact that $\log t$ is the derivative of a strictly convex function $f(t)=t\log t-t$:

$$f'(t)=(t\log t-t)'=\log t,$$

where $'$ denotes taking the derivative. By convex analysis, $f(t)$ and $f^*(t)=\exp(t)$ are a pair of convex conjugate functions, with $(f^*)'=(\exp)'=(f')^{-1}$. The (negative) Shannon entropy $S[P]$ of a probability function $P(\zeta)$

$$-S[P]=\int_\zeta P\log P=1+\int_\zeta (P\log P-P)=1+\int_\zeta f(P)$$

is known to be dual (convex conjugate) to log-partition function for an exponential family P. Note that throughout the paper, we write $\int_\zeta$, which can be read as $\int \mathrm{d}\mu(\zeta)$ having assumed a reference (background) measure μ on the sample space (indexed by ζ), or simply read as $\int \mathrm{d}\zeta$.

ϕ-exponential/logarithmic approach (Naudts [5–7]*).* Given a strictly increasing and positive function $\phi:\mathbb{R}_+\to\mathbb{R}_+$, the ϕ-logarithm, $\log_\phi$, is *notationally* defined as

$$\log_\phi(t)=\int_1^t \frac{1}{\phi(s)}\mathrm{d}s \qquad (t>0), \tag{1}$$

with $\log_\phi(1)=0$. It is the function that satisfies

$$(\log_\phi)'(t)\equiv\frac{\mathrm{d}\log_\phi(t)}{\mathrm{d}t}=\frac{1}{\phi(t)} \tag{2}$$

which, when $\phi(t)=t$, simply means $(\log(t))'=1/t$.

The ϕ-exponential $\exp_\phi$ is defined as the inverse function of $\log_\phi$:

$$\log_\phi(\exp_\phi(t))\equiv t.$$

Taking the derivative of the above, using the chain rule of differentiation, and rearranging yields

$$(\exp_\phi)'(t)\equiv\frac{\mathrm{d}\exp_\phi(t)}{\mathrm{d}t}=\phi\left(\exp_\phi(t)\right)\equiv\bar{\phi}(t),$$

where $\bar{\phi}$ denotes the derivative of $\exp_\phi$. Then, we have the integral identity:

$$\exp_\phi(t)=1+\int_0^t \phi(\exp_\phi(s))\,\mathrm{d}s=1+\int_0^t \bar{\phi}(s)\,\mathrm{d}s,$$

with $\exp_\phi(0)=1$. So the deformed exponential function $\exp_\phi$ can be viewed as the solution $h(t)=\exp_\phi(t)$ to the following integral (and its equivalent differential) equation about the unknown function h:

$$h(t)=1+\int_0^t \phi(h(s))\mathrm{d}s \quad\Longleftrightarrow\quad \frac{\mathrm{d}h}{\mathrm{d}t}=\phi(h(t)).$$

Using $\phi(t)=\bar{\phi}\left(\log_\phi(t)\right)$, we can recast (1) as

$$\frac{\mathrm{d}\log_\phi(t)}{\mathrm{d}t}=\frac{1}{\bar{\phi}\left(\log_\phi(t)\right)}$$

or equivalently

$$\log_\phi(t) = \int_1^t \frac{1}{\bar{\phi}\left(\log_\phi(s)\right)} \, \mathrm{d}s.$$

Therefore, we can view the deformed logarithm function $\log_\phi(t)$ as the solution $h(t) = \log_\phi(t)$ to the following integral (and its equivalent differential) equation about the unknown function h:

$$h(t) = \int_1^t \frac{1}{\bar{\phi}(h(s))} \, \mathrm{d}s \quad \Longleftrightarrow \quad \frac{\mathrm{d}h}{\mathrm{d}t} = \frac{1}{\bar{\phi}(h(t))}.$$

Note that the ϕ and $\bar{\phi}$ functions are linked through

$$\bar{\phi}(t) = \phi\left(\exp_\phi(t)\right), \quad \phi(t) = \bar{\phi}\left(\log_\phi(t)\right). \tag{3}$$

So with respect to the four functions ϕ, $\log_\phi$, $\exp_\phi$, $\bar{\phi}$, specifying the functional form of any one of them will specify all four function forms. For this reason, we say that Naudts' deformed exponential/logarithmic approach is a "one-function" model.

For later convenience, we will introduce the notation U_ϕ and U_ϕ^* to denote the integral functions of $\exp_\phi$ and $\log_\phi$

$$(U_\phi)' = \exp_\phi, \qquad (U_\phi^*)' = \log_\phi .$$

In other words, we have the following chains of derivatives:

$$U_\phi \xrightarrow{\prime} \exp_\phi \xrightarrow{\prime} \bar{\phi} = \phi(\exp_\phi);$$
$$U_\phi^* \xrightarrow{\prime} \log_\phi \xrightarrow{\prime} \frac{1}{\phi}.$$

From convex analysis, U_ϕ and U_ϕ^* are a pair of strictly convex functions that are conjugate to one another, so $*$ is the convex "conjugate" operation. The notation U is in honor of Eguchi [11] who independently proposed the U-model to the statistical machine learning community; the U-model turns out to be identical to the ϕ-model, see [13].

The following identity can be verified:

$$\begin{aligned} U_\phi(t) + U_\phi^*(\exp_\phi(t)) &= t \cdot \exp_\phi(t), \\ U_\phi(\log_\phi(t)) + U_\phi^*(t) &= \log_\phi(t) \cdot t. \end{aligned}$$

Conjugate (ρ, τ)*-embedding approach* (*Zhang* [8,9,18]). Given any smooth strictly convex function $f : \mathbb{R} \to \mathbb{R}$, the convex conjugate function f^* is given by

$$f^*(u) = u \cdot (f')^{-1}(u) - f((f')^{-1}(u)),$$

with

$$(f^*)^* = f, \quad (f^*)' = (f')^{-1}.$$

Here we use $()'$ to indicate taking the derivative, and $()^{-1}$ to indicate taking the function inverse. Strict convexity of f, f^* means that f' and $(f^*)'$ are both strictly increasing functions.

Substituting u for $\tau(t)$, and introducing $\rho(t)$ to denote $(f')^{-1}(\tau(t))$, with ρ, τ as two strictly increasing functions, we have the identity

$$f(\rho(t)) + f^*(\tau(t)) - \rho(t)\tau(t) = 0,$$

with

$$f'(\rho(t)) = \tau(t), \qquad (f^*)'(\tau(t)) = \rho(t).$$

Note that among the four functions, f, f^*, ρ, τ, only two can be freely chosen (except that they cannot simultaneously be f and f^*, since one specifies the other). In other words, we can freely choose any pair of (f, ρ), (f, τ), (f^*, ρ), (f^*, τ), or (ρ, τ) to specify the remaining two —for this reason, it is a "two-function" model for deforming exponential/logarithmic functions. By convention, we will use the notation (ρ, τ) to index this two-function deformation model, where $\rho(\cdot), \tau(\cdot)$ are two strictly increasing functions but otherwise freely chosen. When they become fixed, we have

$$f'(u) = \tau(\rho^{-1}(u)), \quad (f^*)'(u) = \rho(\tau^{-1}(u)).$$

So f and f^* can be obtained simply by integration. Note that exchanging $(\rho, \tau) \longleftrightarrow (\tau, \rho)$ leads to an exchange of $f \longleftrightarrow f^*$. Because strictly increasing functions form a group under function composition, as observed in [8,10], when f and f^* are strictly convex, f' and $(f^*)'$ are nothing but a pair of mutually inverse elements of such group. The motivation of the (ρ, τ)-representation is from the generalization of α-divergence [8] (which itself is a generalization of Kullback Leibler divergence).

Relation between ϕ*- and* (ρ, τ)*-formulation.* In the language of ϕ-formulation, we have, from eq. (3),

$$\begin{aligned} \exp_\phi(t) &= \phi^{-1}\left(\bar{\phi}(t)\right) = (\phi^{-1} \circ \bar{\phi})(t), \\ \log_\phi(t) &= (\bar{\phi})^{-1}\left(\phi(t)\right) = (\bar{\phi}^{-1} \circ \phi)(t). \end{aligned}$$

This can be compared with the following relation in the language of (ρ, τ)-formulation:

$$\begin{aligned} f'(t) &= \tau(\rho^{-1}(t)) = (\tau \circ \rho^{-1})(t), \\ f^{*\prime}(t) &= \rho(\tau^{-1}(t)) = (\rho \circ \tau^{-1})(t). \end{aligned}$$

There is a key difference between the two formulations —in the former, there is only free function (whether taken to be ϕ or $\bar{\phi}$), whereas in the latter, both ρ and τ are taken as free functions. So there are a variety of ways ("gauge freedom") to reduce the two-function (ρ, τ)-model to a one-function ϕ-model.

i) Take $\tau = \mathrm{id}$. Then $(f^*)' = \rho$ can be taken to be $\log_\phi$. So $f' = ((f^*)')^{-1} = (\log_\phi)^{-1} = \exp_\phi$. That is to say, f, f^* are simply integrals of $\log_\phi$ and $\exp_\phi$, respectively. We call this *Type I gauge.*

ii) Take $\tau = \phi$. Then ρ^{-1} can be taken to be $\exp_\phi$, with $f' = \phi(\exp_\phi) = \bar{\phi} = (\exp_\phi)'$, so $f = \exp_\phi$. And $\rho = \log_\phi$, so the (ρ, τ)-relation is governed by the deformed log transformation: $\rho = \log_\tau$. We call this *Type II gauge.*

Type I and Type II gauges are motivated by the Riemannian metric of deformed exponential models (see later) though other gauge selections are possible. Table 1 provides the expressions of f, f^*, τ in terms of given ρ for both Type I and Type II gauges.

Table 1: Type I and Type II gauges in the (ρ, τ)-model.

Gauge	f	f^*	τ
Type I	$f' = \rho^{-1}$	$(f^*)' = \rho$	$\tau = \mathrm{id}$
Type II	$f = \rho^{-1}$	$(f^*)' = ((\rho^{-1})')^{-1}$	$\tau = 1/\rho'$

Deformed exponential family. – Within parametric families of probability functions $\theta \to p(\zeta\,|\,\theta)$ where $p(\zeta\,|\,\theta) \geq 0, \int_\zeta p(\zeta\,|\,\theta) = 1$, we consider a special family $p(\zeta\,|\,\theta) = P^\theta(\zeta)$ as described below. We fix an arbitrary monotone function ϕ along with a set of random functions $\{F_k(\zeta)\}_{k=1,\ldots,n}$. The so-called deformed exponential (or ϕ-exponential) model in statistical physics is defined as

$$P^\theta(\zeta) = \exp_\phi\left(\sum_k \theta^k F_k(\zeta) - \alpha(\theta)\right), \tag{4}$$

where one has assumed that the domain of θ is an open subset of $\mathbb{R}^n$ and $\alpha(\theta)$, called the *normalization function*, exists. Normalization of $P^\theta(\cdot)$ leads to

$$\frac{\partial\alpha(\theta)}{\partial\theta^i} = \int_\zeta \widetilde{P}^\theta F_i(\zeta),$$

where the so-called *escort* probability distribution is given by $\widetilde{P}^\theta(\cdot)$,

$$\widetilde{P}^\theta(\zeta) = \frac{1}{z(\theta)}\phi(P^\theta(\zeta)), \quad z(\theta) = \int_\zeta \phi(P^\theta(\zeta)). \tag{5}$$

Recall our notation $\phi \equiv 1/(\log_\phi)'$, the reciprocal of the derivative of $\log_\phi$ function.

Normalization function α. With respect to the normalization function $\alpha(\theta)$, we can define the dual variable

$$\lambda_i = \frac{\partial\alpha(\theta)}{\partial\theta^i} = \int_\zeta \widetilde{P}^\theta F_i(\zeta).$$

The variables $\{\lambda_k\}_{k=1,\cdots,n}$ can be called "escort expectation" coordinates of the probability family P^θ, because they are equal to the expected value of random functions $F_i(\zeta)$ with respect to the escort-transformed probability $\widetilde{P}^\theta$. The second derivative of $\alpha(\theta)$ is

$$\frac{\partial^2\alpha(\theta)}{\partial\theta^i\partial\theta^j} = \frac{1}{z(\theta)}\int_\zeta (\phi'\cdot\phi)(P^\theta)(F_i(\zeta)-\lambda_i(\theta))\,(F_j(\zeta)-\lambda_j(\theta)).$$

The Legendre dual of $\alpha(\theta)$ is calculated to be

$$\sum_k \theta^k \frac{\partial\alpha(\theta)}{\partial\theta^k} - \alpha(\theta) = \int_\zeta \widetilde{P}^\theta(\zeta)\log_\phi(P^\theta(\zeta)).$$

Potential function V. We define a potential function V by

$$V(\theta) = \alpha(\theta) + \int_\zeta U_\phi\left(\sum_k \theta^k F_k(\zeta) - \alpha(\theta)\right)$$

and calculate the dual variable with respect to $V(\theta)$

$$\eta_i = \frac{\partial V(\theta)}{\partial\theta^i} = \int_\zeta P^\theta(\zeta)F_i(\zeta), \tag{6}$$

where we have used $(U_\phi)' = \exp_\phi$, the definition of probability family $P^\theta(\zeta)$, and that $\int P^\theta(\zeta) = 1$. The variables $\{\eta_k\}_{k=1,\cdots,n}$ can be called "regular expectation" coordinates of the probability family P^θ. The second derivative of $V(\theta)$ is

$$\frac{\partial^2 V(\theta)}{\partial\theta^i\partial\theta^j} = \int_\zeta \phi(P^\theta)\left(F_i(\zeta) - \frac{\partial\alpha(\theta)}{\partial\theta^i}\right)\left(F_j(\zeta) - \frac{\partial\alpha(\theta)}{\partial\theta^j}\right).$$

So for a ϕ-exponential family P^θ with θ as natural coordinates, there are *two* sets of dual variables λ and η, with respect to the normalization function $\alpha(\theta)$ and potential function $V(\theta)$, respectively. The second derivatives of V and α are used to express two different Riemannian metrics under two different gauge selections (see below). This "splitting" of "normalization function" and "potential function" is a hallmark of ϕ-deformed exponential with $\phi(t) \neq t$.

Divergence, entropy and cross-entropy. – The conjugate (ρ, τ)-embedding framework [8,10,13] defines a (ρ, τ)-divergence $D_{\rho,\tau}[P, Q]$ such that it satisfies $D_{\rho,\tau}[P, Q] = D_{\tau,\rho}[Q, P]$. The suite of entropy, dual-entropy, cross-entropy, relative entropy is:

i) (ρ, τ)-cross-entropy $C_{\rho,\tau}[P, Q]$

$$C_{\rho,\tau}[P, Q] = -\int_\zeta \rho(P(\zeta))\,\tau(Q(\zeta)); \tag{7}$$

ii) (ρ, τ)-entropy $S_{\rho,\tau}[P]$

$$S_{\rho,\tau}[P] = -\int_\zeta f(\rho(P(\zeta))),$$

iii) (ρ, τ)-dual-entropy $S^*_{\rho,\tau}[P]$

$$S^*_{\rho,\tau}[P] = -\int_\zeta f^*(\tau(P(\zeta))),$$

iv) (ρ, τ)-divergence, or (ρ, τ)-relative-entropy $D_{\rho,\tau}[P, Q]$

$$\begin{aligned} D_{\rho,\tau}[P, Q] &= C_{\rho,\tau}[P, Q] - S_{\rho,\tau}[P] - S^*_{\rho,\tau}[Q] \\ &= S_{\rho,\tau}[Q] - S_{\rho,\tau}[P] + C_{\rho,\tau}[P, Q] - C_{\rho,\tau}[Q, Q] \\ &= S^*_{\rho,\tau}[P] - S^*_{\rho,\tau}[Q] + C_{\rho,\tau}[P, Q] - C_{\rho,\tau}[P, P]. \end{aligned}$$

Remark. Note that $S_{\rho,\tau}[P]$ is concave in $\rho(P)$, but not necessarily concave in P. And

$$C_{\rho,\tau}[P, Q] = C_{\tau,\rho}[Q, P], \qquad S^*_{\rho,\tau}[P] = S_{\tau,\rho}[P],$$
$$S_{\rho,\tau}[P] - C_{\rho,\tau}[P, P] + S^*_{\rho,\tau}[P] \equiv 0.$$

In general $C_{\rho,\tau}[Q, Q] \neq S_{\rho,\tau}[Q]$, which is related to the fact that $S^*_{\rho,\tau}[Q] \neq$ const. To compensate for this fact, we may define the *modified cross-entropy* or *U-cross-entropy*

$$\overline{C}_{\rho,\tau}[P, Q] = C_{\rho,\tau}[P, Q] - S^*_{\rho,\tau}[Q],$$

Table 2: Gauge types for ϕ-exponential family.

Functions	Type I gauge	Type II gauge
ρ, τ	$\rho = \log_\phi, \tau = \text{id}$	$\rho = \log_\phi,\ \tau = \phi$
f	$f = U_\phi,\ f' = \exp_\phi,\ f'' = \bar{\phi}$	$f = \exp_\phi,\ f' = \bar{\phi}$
f^*	$f^* = U_\phi^*,\ (f^*)' = \log_\phi$	$f^* \circ \phi = \phi \cdot \log_\phi - \text{id},\ (f^*)' = (\bar{\phi})^{-1}$

which has the desired property

$$\bar{C}_{\rho,\tau}[P,P] = S_{\rho,\tau}[P],$$

while

$$\bar{C}_{\rho,\tau}[P,Q] \neq \bar{C}_{\tau,\rho}[Q,P].$$

We have

$$\begin{aligned} D_{\rho,\tau}[P,Q] &= \bar{C}_{\rho,\tau}[P,Q] - \bar{C}_{\rho,\tau}[P,P] \\ &= \bar{C}_{\rho,\tau}[P,Q] - S_{\rho,\tau}[P]. \end{aligned}$$

Proposition 1. *The rho-tau-divergence function of a ϕ-exponential family and the associated entropy, dual-entropy, and cross-entropy take the form (note that we have chosen $\rho = \log_\phi$, so gauge selection is to select a τ; of course, the role of ρ and τ can be switched with only notational change):*

1) For Type I gauge:

$$\begin{aligned} D_{\rho,\tau}[P,Q] &= \int_\zeta U_\phi(\log_\phi P) - \int_\zeta U_\phi(\log_\phi Q) \\ &\quad - \int_\zeta Q \cdot (\log_\phi P - \log_\phi Q), \\ S_{\rho,\tau}[P] &= -\int_\zeta U_\phi(\log_\phi P), \\ S^*_{\rho,\tau}[P] &= \int_\zeta U_\phi(\log_\phi P) - \int_\zeta P \log_\phi P, \\ C_{\rho,\tau}[P,Q] &= -\int_\zeta Q \log_\phi P, \\ \bar{C}_{\rho,\tau}[P,Q] &= \int_\zeta Q \cdot (\log_\phi Q - \log_\phi P) - \int_\zeta U_\phi(\log_\phi Q). \end{aligned}$$

2) For Type II gauge:

$$\begin{aligned} D_{\rho,\tau}[P,Q] &= \int_\zeta \phi(Q)(\log_\phi Q - \log_\phi P), \\ S_{\rho,\tau}[P] &= \text{const}, \\ S^*_{\rho,\tau}[P] &= -\int_\zeta \phi(P) \log_\phi P + \text{const}, \\ C_{\rho,\tau}[P,Q] &= -\int_\zeta \phi(Q) \log_\phi P, \\ \bar{C}_{\rho,\tau}[P,Q] &= \int_\zeta \phi(Q)(\log_\phi Q - \log_\phi P) + \text{const}. \end{aligned}$$

Remark. A highlight of our analysis is that the (ρ,τ)-divergence $D_{\rho,\tau}$ provides a clear distinction between entropy and cross-entropy as *two* distinct quantities *without* requiring thelatter to degenerate to the former. On the other hand, fixing the gauge $f = \rho^{-1}$ renders entropy S constant. So it is also called constant-S gauge. In this case, $\tau \longleftrightarrow \rho$ is akin to the $\log_\phi \longleftrightarrow \phi$ transformation encountered in studying the ϕ-exponential family (see table 2).

Example 4. *In the q-exponential case, we take*

$$\rho(t) = \log_q(t) = \frac{t^{1-q} - 1}{1-q},$$

and Tsallis entropy S_q^T has been defined as [2,3]

$$S_q^T[P] = -\int_\zeta \frac{P - P^q}{1-q}.$$

The Type I gauge corresponds to the choice of

$$f(t) = U_\phi(t) = \int \exp_q(t) = \frac{(1 + (1-q)t)^{\frac{2-q}{1-q}}}{2-q}.$$

Under Type I gauge, rho-tau-entropy $S_{\rho,\tau}$ is essentially-equivalent to the Tsallis entropy (with parameter $2-q$):

$$\begin{aligned} S_{\rho,\tau}[P] &= -\int_\zeta f(\rho(P)) = -\frac{1}{2-q}\int_\zeta P^{2-q} \\ &= \frac{q-1}{2-q} S^T_{2-q}[P] + \frac{1}{q-2}, \end{aligned}$$

and so is dual rho-tau-entropy $S^*_{\rho,\tau}$

$$\begin{aligned} S^*_{\rho,\tau}[P] &= -\frac{1}{2-q}\int_\zeta P(\log_q(P) - 1) \\ &= \frac{1}{2-q} S^T_{2-q}[P] + \frac{1}{2-q}. \end{aligned}$$

The rho-tau-divergence (known as the U-divergence for the Tsallis model) can be explicitly given as follows:

$$\begin{aligned} D_{\rho,\tau}[P,Q] &= \int_\zeta \frac{Q^{2-q}}{2-q} - \int_\zeta \frac{P^{2-q}}{2-q} + \int_\zeta P \frac{Q^{1-q} - P^{1-q}}{1-q} \\ &= \frac{1}{(1-q)(2-q)}\int_\zeta P^{2-q} + \frac{1}{2-q}\int_\zeta Q^{2-q} - \frac{1}{1-q}\int_\zeta PQ^{1-q}. \end{aligned}$$

This diverence coincides with the density power divergence [19] *and the β-divergence with $\beta = 1-q$.*

On the other hand, the Type II gauge corresponds to the choice of

$$f(t) = \rho^{-1}(t) = \exp_q(t).$$

Under Type II gauge, dual rho-tau-entropy $S^*_{\rho,\tau}$ is essentially the Tsallis q-entropy:

$$\begin{aligned} S^*_{\rho,\tau}[P] &= 1 - \int_\zeta P^q \log_q(P) = 1 - \int_\zeta \frac{P - P^q}{1-q} \\ &= S^T_q[P] + 1. \end{aligned}$$

The rho-tau-entropy $S_{\rho,\tau}$ is constant:

$$S_{\rho,\tau}[P] = -\int_\zeta f(\rho(P)) = -\int_\zeta P = -1.$$

The rho-tau-divergence can be given by

$$\begin{aligned} D_{\rho,\tau}[P,Q] &= \int_\zeta Q^q(\log_q(Q) - \log_q(P)) \\ &= \frac{1}{1-q}\left(1 - \int_\zeta P^{1-q}Q^q\right). \end{aligned}$$

This divergence coincides with the α-divergence with $\alpha = 2q-1$ (except for a constant multiple).

Riemannian geometry of deformed exponential family. – Information geometry [1,17] provides a now-standard script for producing a Riemannian manifold structure based on a properly defined divergence function $D[P,Q]$. The resulting "statistical structure" not only comes with a positive semi-definite Riemannian metric g, but also a pair of torsion-free affine connections Γ, Γ^* that are conjugate with respect to g. These connections usually carry non-vanishing curvature. In the special case when $D[P,Q]$ takes the form of a canonical divergence function (related to Bregman divergence), a Hessian metric g would result, with a pair of dually flat connections (with vanishing curvature as well as torsion) and a pair of dual affine coordinates.

Specifically, for a parametric family of density functions denoted $p(\zeta \mid \theta)$, with θ_1, θ_2 indexing P, Q, respectively, the Riemannian metric $g(\theta)$ can be calculated from $D[P,Q] = D[p(\cdot \mid \theta_1), p(\cdot \mid \theta_2)] = D(\theta_1, \theta_2)$ via

$$g_{ij}(\theta) = -\left.\frac{\partial^2 D(\theta_1,\theta_2)}{\partial\theta_1^i \partial\theta_2^j}\right|_{\theta_2=\theta_1} = -\left.\frac{\partial^2 D(\theta_1,\theta_2)}{\partial\theta_2^i \partial\theta_1^j}\right|_{\theta_2=\theta_1},$$

$$\Gamma_{ij,k}(\theta) = -\left.\frac{\partial^3 D(\theta_1,\theta_2)}{\partial\theta_2^i \partial\theta_2^j \partial\theta_1^k}\right|_{\theta_2=\theta_1}, \Gamma^*_{ij,k}(\theta) = -\left.\frac{\partial^3 D(\theta_1,\theta_2)}{\partial\theta_1^i \partial\theta_1^j \partial\theta_2^k}\right|_{\theta_2=\theta_1}.$$

One verifies that Γ and Γ^* as induced above always satisfy

$$\frac{\partial g_{ij}(\theta)}{\partial\theta^k} = \Gamma_{ik,j}(\theta) + \Gamma^*_{kj,i}(\theta),$$

and are, by definition, "conjugate" with respect to the induced g; each behaves, under coordinate transform, in the same manner as an affine connection does. Moreover, $\frac{1}{2}(\Gamma + \Gamma^*)$ can be shown to be the Levi-Civita connection associated to g. Torsion-freeness of Γ and Γ^* is demonstrated by $\Gamma_{ik,j} = \Gamma_{ki,j}, \Gamma^*_{kj,i} = \Gamma^*_{jk,i}$. This relationship between a divergence function $D[P,Q]$ and the statistical structure (g, Γ, Γ^*) it induces forms the cornerstone of classical information geometry [1,17].

We carry out the explict calculations using the divergence function $D_{\rho,\tau}[P,Q]$, which can be equivalently carried out on the cross-entropy $C_{\rho,\tau}[P,Q]$ of eq. (7), to obtain the Riemannian (ρ,τ)-metric (calculations for conjugate (ρ,τ)-connections are omitted, see [8]):

$$g_{ij}(\theta) = \int_\zeta \frac{\partial\rho(p(\zeta|\theta))}{\partial\theta^i}\frac{\partial\tau(p(\zeta|\theta))}{\partial\theta^j} = \int_\zeta \frac{1}{\psi(p)}\frac{\partial p}{\partial\theta^i}\frac{\partial p}{\partial\theta^j},$$

with $\psi(t) = 1/(\rho'(t)\tau'(t))$. In general, $g_{ij}(\theta)$ is not a Hessian metric (the second derivative of some smooth convex function). When $p(\zeta|\theta)$ takes the form of a deformed exponential family P^θ in eq. (4), which includes the exponential family as a special case, then g may take the form of a Hessian metric through the potential function $V(\theta)$, or a conformal Hessian metric normalization function $\alpha(\theta)$ (along with a conformal factor).

Proposition 2. (*Naudts and Zhang* [13]) *The (ρ,τ)-metric g^ϕ of a ϕ-exponential family takes the form of a:*

1) (*Under Type I gauge*) *Hessian metric, to which the escort metric $\tilde{g}_{ij}$ is conformally related:*

$$g^\phi_{ij}(\theta) = g^V_{ij}(\theta) := \frac{\partial^2 V(\theta)}{\partial\theta^i\partial\theta^j} = z(\theta)\tilde{g}_{ij}(\theta);$$

here the escort metric $\tilde{g}_{ij}(\theta)$ is defined by

$$\int_\zeta \widetilde{P}^\theta\left(F_i(\zeta) - \int_\zeta \widetilde{P}^\theta F_i(\zeta)\right)\left(F_j(\zeta) - \int_\zeta \widetilde{P}^\theta F_j(\zeta)\right),$$

and $\widetilde{P}^\theta$ is the escort probability given by eq. (5).

2) (*Under Type II gauge*) *conformal Hessian metric, i.e., conformal to the Hessian metric $g^\alpha_{ij}(\theta)$:*

$$g^\phi_{ij}(\theta) = z(\theta)g^\alpha_{ij}(\theta) = z(\theta)\frac{\partial^2\alpha(\theta)}{\partial\theta^i\partial\theta^j}.$$

Note that, previously, various Riemannian metrics for the deformed exponential model were considered [14,15,20]. It was the work of Naudts and Zhang that provided a unified treatment using (ρ,τ)-metric (associated to the (ρ,τ)-divergence) and the concept of gauge freedom [13].

Example 5. *Consider the q-Gaussian distribution* ($1 < q < 3$) *of a random variable ζ over the real line $\mathbb{R}$*

$$P^\theta(\zeta) = \exp_q\left(\theta^1 f_1(\zeta) + \theta^2 f_2(\zeta) - \alpha(\theta)\right)$$

with $\theta^1 \in \mathbb{R}, \theta^2 < 0$ and $f_1(\zeta) = \zeta, f_2(\zeta) = (\zeta)^2$ for $\zeta \in \mathbb{R}$, see [21]. *The normalization function $\alpha(\theta)$ is*

$$\alpha(\theta) = -\frac{(\theta^1)^2}{4\theta^2} - \log_q\left(\frac{1}{C_q(\theta^2)}\right),$$

where ($B(\cdot,\cdot)$ stands for the beta function):

$$c_q = \sqrt{\frac{1}{q-1}}B\left(\frac{3-q}{2(q-1)}, \frac{1}{2}\right), \quad C_q(\theta^2) = \left(\frac{c_q}{\sqrt{-\theta^2}}\right)^{\frac{2}{3-q}}.$$

The Hessian metric g^α under the Type II gauge is

$$g^\alpha(\theta) = -\frac{1}{2\theta^2}\begin{pmatrix} 1 & -\theta^1/\theta^2 \\ -\theta^1/\theta^2 & (\theta^1)^2/(\theta^2)^2 - C_q(\theta^2)^{q-1}/\theta^2 \end{pmatrix}.$$

The rho-tau-metric g^ϕ is

$$g^\phi(\theta) = z(\theta)g^\alpha(\theta), \quad z(\theta) = \frac{3-q}{2}C_q(\theta^2)^{1-q}.$$

Discussions and perspectives. – This article provides a unified framework on generalizing (*i.e.*, deforming) Shannon entropy, cross-entropy and Kullback-Leibler divergence (relative entropy). Our analysis is from the perspective of information geometry, which affords a suite of differential geometric tools to study the manifold of probability density functions through divergence functions that characterize their proximty. The Riemannian geometry comes with a pair of conjugate connections linked to the induced metric, the average of which is the Levi-Civita connection. This is the "Standard Model" [1,17] of classical information geometry, the cornerstone of which is the dualistic affine connection structure generated from properly-defined divergence functions. The Standard Model includes the well understood case of Hessian geometry which rigidly interlocks 1) the exponential family of probability functions; 2) the Legendre duality of convex functions as coordinate transform on the Hessian manifold; 3) the dually flat nature of the induced affine connections. Our framework for deformation, which culminates from nearly two decades of exploration within the information geometry community, still follows the script of the Standard Model and hence is constrained by the rigid triangulation above. On the other hand, this article has not devoted much (if at all) attention to schemes by theoretical physicists and information theorists in generalizing Shannon entropy based on axiomatic approach, for example, [22–26]. Those schemes involve decomposing a complex system into multiple sub-components typically with certain scaling properties. While our framework is purely mathematical, relying on geometrical understanding of the relations among 1) through 3) above, those complementary, physics-based schemes provide contexts to view entropy, cross-entropy, relative entropy as information entities and to reveal their respective roles in aggregation, learning, self-organization, emergence, and other adaptive processes.

This so-called "conjugate (ρ, τ) embedding" approach clarifies various phenomena that emerge as a result of adopting general embedding functions —these phenomena have been largely obscured in the Standard Model due to its use of the standard logarithm/exponential function. In general, the deformed ϕ-exponential family always has *two* potentials, V and α, which are not equal unless there is no deformation. They induce a Hessian and a conformal Hessian metric under Type I and Type II gauge, respectively. This is the "splitting" phenomenon for deformed exponential families, and is worth exploring in interactive physical systems. One limitation is that the ϕ-exponential family only deals with substractive normalization but not divisive normalization [27]; the latter is used in the construction of the so-called $F^{(\alpha)}$-family [28] shown to be related to the Rényi entropy. Likewise, (ρ, τ) conjugate embedding is based on Legendre duality rather than the more general abstract c-duality used in [28]. Work in progress by the first author and T. K. L. Wong will overcome these limitations and establish the unique status of Tsallis and Rényi deformation (paper in preparation).

* * *

JZ is supported by DARPA/ARO Grant W911NF-16-1-0383 and AFOSR Grant FA9550-19-1-0213.

REFERENCES

[1] Amari S. and Nagaoka H., *Methods of information geometry, Translations of Mathematical Monographs*, Vol. **191** (AMS, 2000; Oxford University Press, 2000).

[2] Tsallis C., *J. Stat. Phys.*, **52** (1988) 479.

[3] Tsallis C., *Quim. Nova*, **17** (1994) 468.

[4] Kaniadakis G., *Physica A*, **296** (2001) 405.

[5] Naudts J., *J. Inequal. Pure Appl. Math.*, **5** (2004) 102.

[6] Naudts J., *Entropy*, **10** (2008) 131.

[7] Naudts J., *Generalised Thermostatistics* (Springer) 2011.

[8] Zhang J., *Neural Comput.*, **16** (2004) 159.

[9] Zhang J., *Entropy*, **15** (2013) 5384.

[10] Zhang J., *Entropy*, **17** (2015) 4485.

[11] Eguchi S., *Sugaku Expositions*, **19** (2006) 197; *Sūgaku*, **56** (2004) 380 (in Japanese).

[12] Murata N. *et al.*, *Neural Comput.*, **16** (2004) 1437.

[13] Naudts J. and Zhang J., *Inf. Geom.*, **1** (2018) 79.

[14] Amari S. I. *et al.*, *Physica A*, **391** (2012) 4308.

[15] Matsuzoe H., *Differ. Geom. Appl.*, **35** (2014) 323.

[16] Ohara A. *et al.*, *Mod. Phys. Lett. B*, **26** (2012) 1250063.

[17] Amari S., *Differential Geometric Methods in Statistics, Lect. Notes Stat.*, Vol. **28** (Springer) 1985.

[18] Zhang J., in *Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan, 2005* (Tokyo U. Press) 2006, pp. 58–67.

[19] Basu A. *et al.*, *Biometrika*, **85** (2020) 549.

[20] Matsuzoe H. *et al.*, in *International Conference on Geometric Science of Information* (Springer, Cham) 2017, pp. 223–230.

[21] Tanaya D. *et al.*, in *Recent Progress in Differential Geometry and Its Related Fields* (World Scientific Publ.) 2011, pp. 137–149.

[22] Hanel R. and Thurner S., *EPL*, **93** (2011) 20006.

[23] Jizba P. and Arimitsu T., *Ann. Phys.*, **312** (2004) 17.

[24] Jizba P. and Korbel J., *Phys. Rev. Lett.*, **122** (2019) 120601.

[25] Jizba P. and Korbel J., *Phys. Rev. E*, **101** (2020) 042126.

[26] Thurner S. *et al.*, *Introduction to the Theory of Complex Systems* (Oxford University Press) 2018.

[27] Zhang J. and Hästö P., *J. Math. Psychol.*, **50** (2006) 60.

[28] Wong T. K. L., *Inf. Geom.*, **1** (2018) 39.