Metadata of the chapter that will be visualized in SpringerLink

Book Title	Geometric Science of Information
Series Title	
Chapter Title	Information Geometry Under Monotone Embedding. Part I: Divergence Functions
Copyright Year	2017
Copyright HolderName	Springer International Publishing AG

Corresponding Author	Family Name	**Zhang**
	Particle	
	Given Name	**Jun**
	Prefix	
	Suffix	
	Division	
	Organization	University of Michigan
	Address	Ann Arbor, MI, USA
	Email	junz@umich.edu
Author	Family Name	**Naudts**
	Particle	
	Given Name	**Jan**
	Prefix	
	Suffix	
	Division	
	Organization	Universiteit Antwerpen
	Address	Antwerpen, Belgium
	Email	Jan.Naudts@uantwerpen.be

Abstract	The standard model of information geometry, expressed as Fisher-Rao metric and Amari-Chensov tensor, reflects an embedding of probability density by $\log$-transform. The standard embedding was generalized by one-parametric families of embedding function, such as α-embedding, q-embedding, κ-embedding. Further generalizations using arbitrary monotone functions (or positive functions as derivatives) include the deformed-log embedding (Naudts), U-embedding (Eguchi), and rho-tau dual embedding (Zhang). Here we demonstrate that the divergence function under the rho-tau dual embedding degenerates, upon taking $\rho = id$, to that under either deformed-log embedding or U-embedding; hence the latter two give an identical divergence function. While the rho-tau embedding gives rise to the most general form of cross-entropy with two free functions, its entropy reduces to that of deformed entropy of Naudts with only one free function. Fixing the gauge freedom in rho-tau embedding through normalization of dual-entropy function renders rho-tau cross-entropy to degenerate to U cross-entropy of Eguchi, which has the simpler property, not true for general rho-tau cross-entropy, of reducing to the deformed entropy upon setting the two pdfs to be equal. In Part I, we investigate monotone embedding in divergence function, entropy and cross-entropy, whereas in the sequel (Part II), in induced geometries and probability families.

Keywords (separated by '-')	phi-embedding - U-embedding - rho-tau embedding - rho-tau divergence - rho-tau cross-entropy - U cross-entropy - Deformed entropy - Gauge

Information Geometry Under Monotone Embedding. Part I: Divergence Functions

Jun Zhang[1(✉)] and Jan Naudts[2]

[1] University of Michigan, Ann Arbor, MI, USA
junz@umich.edu
[2] Universiteit Antwerpen, Antwerpen, Belgium
Jan.Naudts@uantwerpen.be

Abstract. The standard model of information geometry, expressed as Fisher-Rao metric and Amari-Chensov tensor, reflects an embedding of probability density by log-transform. The standard embedding was generalized by one-parametric families of embedding function, such as α-embedding, q-embedding, κ-embedding. Further generalizations using arbitrary monotone functions (or positive functions as derivatives) include the deformed-log embedding (Naudts), U-embedding (Eguchi), and rho-tau dual embedding (Zhang). Here we demonstrate that the divergence function under the rho-tau dual embedding degenerates, upon taking $\rho = id$, to that under either deformed-log embedding or U-embedding; hence the latter two give an identical divergence function. While the rho-tau embedding gives rise to the most general form of cross-entropy with two free functions, its entropy reduces to that of deformed entropy of Naudts with only one free function. Fixing the gauge freedom in rho-tau embedding through normalization of dual-entropy function renders rho-tau cross-entropy to degenerate to U cross-entropy of Eguchi, which has the simpler property, not true for general rho-tau cross-entropy, of reducing to the deformed entropy upon setting the two pdfs to be equal. In Part I, we investigate monotone embedding in divergence function, entropy and cross-entropy, whereas in the sequel (Part II), in induced geometries and probability families. AQ1

Keywords: phi-embedding · U-embedding · rho-tau embedding · rho-tau divergence · rho-tau cross-entropy · U cross-entropy · Deformed entropy · Gauge

1 Introduction: A Plethora of Probability Embeddings

One motivation to study probability embedding functions is to extend the framework of information geometry beyond the now-classic expressions of Fisher-Rao metric and Amari-Chensov tensor. Realizing that the standard α-geometry is based on log-embedding of probability functions, various approaches have been

J. Zhang and J. Naudts contributed equally to this paper.

F. Nielsen and F. Barbaresco (Eds.): GSI 2017, LNCS 10589, pp. 1–11, 2017.
https://doi.org/10.1007/978-3-319-68445-1_24

proposed to generalize such probability embedding, using a one-parameter family of specific functions at the first level of generality, and using arbitrarily chosen (monotone or positive) functions at the second level of generality.

(i). α-embedding. It was Amari [1] who first investigated the one-parameter family of embeddings $\log_\alpha : \mathbb{R}^+ \to \mathbb{R}$ defined by

$$\log_\alpha(u) = \begin{cases} \log u & \alpha = 1 \\ \frac{2}{1-\alpha}\, u^{(1-\alpha)/2} & \alpha \neq 1 \end{cases} \tag{1}$$

Under this α-embedding, α-divergence becomes canonical divergence, and α-connections have a simple Γ^1, Γ^{-1}-like characteristics [2].

(ii). q-exponential embedding. Tsallis [3], in investigating the equilibrium distribution of statistical physics which maximizes the Boltzmann-Gibbs-Shannon entropy under constraints, replaced the entropy function by a q-dependent entropy, resulting in a deformed version of statistical physics; here, $q \in \mathbb{R}$. The q-logarithmic/exponential functions were introduced [4]:

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

Note that q-embedding and α-embedding functions are different: $\log_q(\cdot) \neq \log_\alpha(\cdot)$, even after the identification $\alpha = 2q-1$. Like α-embedding, q-embedding reduces to the standard logarithm as $\lim_{q\to 1}$.

(iii). κ-exponential embedding. An alternative to the q-deformed exponential model for statistical physics is the κ-model [5], where

$$\log_\kappa(u) = \frac{1}{2\kappa}\left(u^\kappa - u^{-\kappa}\right), \qquad \exp_\kappa(u) = \left(\kappa u + \sqrt{1+\kappa^2 u^2}\right)^{\frac{1}{\kappa}}, \qquad \kappa \neq 0;$$

the case of $\lim_{\kappa\to 0}$ corresponds to the standard exponential/logarithm.

(iv). ϕ-, U-, and(ρ, τ)-embedding. Generalizing any parametric forms of embedding functions further leads to the consideration of probability embedding using arbitrary monotone (or after taking derivative, positive) functions. The prominent inventions are Naudts' phi-embedding [7], Eguchi's U-embedding [8], and Zhang's rho-tau embedding [6], though they have been re-invented/renamed by later authors, causing confusion and distraction. We discuss these in the next section.

Below we first review the deformed logarithm, $\log_\phi$, and deformed exponential, $\exp_\phi$, functions. Then we point out that $\log_\phi$ and $\exp_\phi$ are nothing but an arbitrary pair of mutually inverse monotone functions, and are representable as derivatives of a pair of conjugate convex functions f, f^*. The deformed divergence $D_\phi(p,q)$ is then precisely the Bregman divergence $D_f(p,q)$ associated with f. The construction of entropy and cross-entropy from this deformed approach is reviewed, as well as their construction from the U-embedding. Then, we review the rho-tau embedding, which provides two independently chosen embedding functions, and explicitly identify its entropy and cross-entropy. Our Main Theorem shows that the divergence function and entropy function of the rho-tau embedding reduce as a special case to those given by the phi-embedding and U-embedding, while the rho-tau cross-entropy reduces as another special case to the U cross-entropy.

2 Deformation Versus Embedding

2.1 "Deforming" Exponential and Logarithmic Functions

Naudts [7,9] defines the phi-deformed logarithm

$$\log_\phi(u) = \int_1^u \frac{1}{\phi(v)} \mathrm{d}v.$$

Here, $\phi(v)$ is a strictly positive function. In the context of discrete probabilities it suffices that it is strictly positive on the open interval $(0,1)$, possibly vanishing at the end points. In the case of a probability density function it is assumed to be strictly positive on the interval $(0,+\infty)$. Note that by construction one has $\log_\phi(1) = 0$. The inverse of the phi-logarithm is denoted $\exp_\phi(u)$, and called phi-exponential function:

$$\exp_\phi(\log_\phi(u)) = \log_\phi(\exp_\phi(u)) = u.$$

The phi-exponential has an integral expression

$$\exp_\phi(u) = 1 + \int_0^u \mathrm{d}v\, \psi(v),$$

where the function $\psi(u)$ is given by

$$\psi(u) = \frac{\mathrm{d}}{\mathrm{d}u} \exp_\phi(u) = \frac{\mathrm{d}}{\mathrm{d}u} (\log_\phi)^{-1}(u).$$

In terms of ϕ, ψ, we have the following relations:

$$\begin{aligned} \psi(u) &= \phi(\exp_\phi(u)), \\ \phi(u) &= \psi(\log_\phi(u)). \end{aligned}$$

We want to stress that all four functions, $\phi, \psi, \log_\phi, \exp_\phi$, arise out of choosing one positive-valued function ϕ.

As examples, $\phi(v) = v$ gives rise to the classic natural logarithm and exponential. Taking $\phi(u) = \frac{u}{1+u}$ in [13] leads to $\log_\phi(u) = u - 1 + \log(u)$. Taking $\phi(u) = u(1+\epsilon u)$ in [14] leads to

$$\log\left(\frac{(1+\epsilon)u}{1+\epsilon u}\right), \qquad \exp_\phi(u) = \frac{1}{(1+\epsilon)e^{-v} - \epsilon}.$$

2.2 Deformed Entropy and Deformed Divergence Functions

The phi-entropy of the probability distribution p is defined by [9]

$$S_\phi(p) = -\mathrm{E}_p \log_\phi p + \int_{\mathcal{X}} \mathrm{d}x \int_0^{p(x)} \mathrm{d}u \, \frac{u}{\phi(u)} + \text{ constant}. \tag{2}$$

By partial integration one obtains an equivalent expression

$$S_\phi(p) = -\int_{\mathcal{X}} \mathrm{d}x \int_1^{p(x)} \mathrm{d}u \, \log_\phi(u) + \text{ constant}. \tag{3}$$

For standard logarithm $\phi(u) = u$, the above expression is the well-known entropy of Boltzmann-Gibbs-Shannon

$$S(p) = -\mathrm{E}_p \log p.$$

The phi-divergence of two probability functions p and q is defined by [9]

$$D_\phi(p, q) = \int_{\mathcal{X}} \mathrm{d}x \int_{q(x)}^{p(x)} \mathrm{d}v \, \left[\log_\phi(v) - \log_\phi(q(x))\right], \tag{4}$$

which has another equivalent expression

$$D_\phi(p, q) = S_\phi(q) - S_\phi(p) - \int_{\mathcal{X}} \mathrm{d}x \, [p(x) - q(x)] \log_\phi(q(x)). \tag{5}$$

Now let us express these quantities in terms of a strictly convex function f, satisfying $f'(u) = \log_\phi(u)$. We have:

$$S_\phi(p) = -\int_{\mathcal{X}} \mathrm{d}x \, f(p(x)) + \text{ constant}, \tag{6}$$

$$D_\phi(p, q) = \int_{\mathcal{X}} \mathrm{d}x \, \{f(p(x)) - f(q(x)) - [p(x) - q(x)] f'(q(x))\}. \tag{7}$$

One can readily recognize that $D_\phi(p, q)$ is nothing but the Bregman divergence, whereas the function f itself determines the deformed entropy $S_\phi(p)$. Note that $p \mapsto S_\phi(p)$ is strictly concave while the map $p \mapsto D_\phi(p, q)$ is strictly convex.

2.3 U-embedding

Eguchi [8] introduces the U-embedding, which is essentially the Bregman divergence under a strictly convex function U coupled with an embedding using $\psi = (U')^{-1}$. The U cross-entropy $C_U(p, q)$ is defined as:

$$C_U(p, q) = \int_{\mathcal{X}} \mathrm{d}x \, \{U(\psi(q(x))) - p(x) \cdot \psi(q(x))\}, \tag{8}$$

whereas the U entropy H_U is defined as $H_U(p) = C_U(p, p)$. The U-divergence is

$$D_U(p, q) = C_U(p, q) - H_U(p, p)$$
$$= \int_{\mathcal{X}} \mathrm{d}x \left\{ U\left(\psi(q(x))\right) - U\left(\psi(p(x))\right) - p(x)\left[(\psi(q(x)) - \psi(p(x))\right] \right\}. \quad (9)$$

Note that the U-embedding only has one arbitrarily chosen function, as does phi-embedding.

2.4 Dual rho-tau Embedding

In contrast with the "single function" embedding of the phi-model and the U-model, Zhang's (2004) rho-tau framework uses *two* arbitrarily and independently chosen monotone functions. He starts with the observation that a pair of mutually inverse functions occurs naturally in the context of convex duality. Indeed, if f is strictly convex and f^* is its convex dual then the derivatives f' and $(f^*)'$ are inverse functions of each other:

$$f' \circ (f^*)'(u) = (f^*)' \circ f'(u) = u.$$

Here the definition of the convex dual f^* of f is:

$$f^*(u) = \sup\{uv - f(v)\}.$$

For u in the range of f' it is given by

$$f^*(u) = u(f')^{-1}(u) - f \circ (f')^{-1}(u).$$

Take the derivative of this expression to find $(f^*)' \circ f'(u) = u$. By convex duality then follows that also $f' \circ (f^*)'(u) = u$. Take an additional derivative to obtain

$$f''((f^*)'(u)) \cdot (f^*)''(u) = (f^*)''(f'(u)) \cdot f''(u) = 1. \quad (10)$$

This identity will be used further on.

Consider now a pair $(\rho(\cdot), \tau(\cdot))$ of strictly increasing functions. Then there exists a strictly convex function $f(\cdot)$ satisfying $f'(u) = \tau \circ \rho^{-1}(u)$. This is because the family of strictly increasing functions form a group, with function composition as group operation, an observation made in [6, 12]. In terms of the conjugate function f^*, the relation is $(f^*)'(u) = \rho \circ \tau^{-1}(u)$. The derivatives of $f(u)$ and of its conjugate $f^*(u)$ have the property that

$$f'(\rho(u)) = \tau(u) \quad \text{and} \quad (f^*)'(\tau(u)) = \rho(u). \quad (11)$$

Among the triple (f, ρ, τ), given any two functions, the third is specified. When we arbitrarily choose two strictly increasing functions ρ and τ as embedding functions, then they are automatically linked by a pair of conjugated convex functions f, f^*. On the other hand, we may also independently choose to specify

$(\rho, f), (\rho, f^*), (\tau, f)$, or (τ, f^*), with the others being fixed. Therefore, rho-tau embedding is a mechanism with *two* independently chosen functions; this differs from both the phi-embedding and the U-embedding. The following identities will be useful:

$$f''(\rho(u))\,\rho'(u) = \tau'(u)\;, \qquad (f^*)''(\tau(u))\,\tau'(u) = \rho'(u)\;, \tag{12}$$

$$f''(\rho(u))\,(\rho'(u))^2 = (f^*)''(\tau(u))\,(\tau'(u))^2\;, \tag{13}$$

$$f''(\rho(u))\,(f^*)''(\tau(u)) = 1. \tag{14}$$

2.5 Divergence of the rho-tau Embedding

Zhang (2004) introduces[1] the rho-tau divergence (see Proposition 6 of [6])

$$D_{\rho,\tau}(p,q) = \int_{\mathcal{X}} \mathrm{d}x\ \{f(\rho(p(x))) + f^*(\tau(q(x))) - \rho(p(x))\tau(q(x))\}\,, \tag{15}$$

where f is a strictly convex function satisfying $f'(\rho(u)) = \tau(u)$.

Lemma 1. *Expression (15) can be written as*

$$\begin{aligned} D_{\rho,\tau}(p,q) &= \int_{\mathcal{X}} \mathrm{d}x \left\{ f(\rho(p(x))) - f(\rho(q(x))) - [\rho(p(x)) - \rho(q(x))]\tau(q(x)) \right\} \\ &= \int_{\mathcal{X}} \mathrm{d}x \int_{q(x)}^{p(x)} [\tau(v) - \tau(q(x))]\,\mathrm{d}\rho(v) \\ &= \int_{\mathcal{X}} \mathrm{d}x \int_{\rho(q(x))}^{\rho(p(x))} \mathrm{d}u\ [f'(u) - f'(\rho(q(x)))]\,. \end{aligned} \tag{16}$$

In particular this implies that $D_{\rho,\tau}(p,q) \geq 0$, with equality if and only if $p = q$. We note the following identity:

$$f(\rho(p(x))) - \rho(p(x))\tau(p(x)) + f^*(\tau(p(x))) = 0. \tag{17}$$

The "reference-representation biduality" [6,10,12] reveals as

$$D_{\rho,\tau}(p,q) = D_{\tau,\rho}(q,p).$$

[1] The original definition as found in [6,12] uses the notation $D_{f,\rho}(p,q)$ and treats f and ρ as independent. In the present definition $D_{\rho,\tau}(p,q)$ the definition of f depends on ρ, τ. The difference in only notational and inconsequential.

2.6 Entropy and Cross-Entropy of rho-tau Embedding

It is now obvious to give the following definition of the rho-tau entropy

$$S_{\rho,\tau}(p) = -\int_{\mathcal{X}} \mathrm{d}x\, f(\rho(p(x))), \tag{18}$$

where $f(u)$ is a strictly convex function satisfying $f'(u) = \tau \circ \rho^{-1}(u)$. This can be written as

$$\begin{aligned} S_{\rho,\tau}(p) &= -\int_{\mathcal{X}} \mathrm{d}x \int^{\rho(p(x))} f'(v)\mathrm{d}v + \text{ constant} \\ &= -\int_{\mathcal{X}} \mathrm{d}x \int^{p(x)} \tau(u)\mathrm{d}\rho(u) + \text{ constant.} \end{aligned} \tag{19}$$

Note that the rho-tau entropy $S_{\rho,\tau}(p)$ is concave in $\rho(p)$, but not necessarily in p. This has consequences further on. We likewise define rho-tau cross-entropy

$$C_{\rho,\tau}(p,q) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x))\tau(q(x))$$

with $C_{\rho,\tau}(p,q) = C_{\tau,\rho}(q,p)$.

The rho-tau divergence can then be given by

$$\begin{aligned} D_{\rho,\tau}(p,q) &= S_{\rho,\tau}(q) - S_{\rho,\tau}(p) - \int_{\mathcal{X}} \mathrm{d}x\, [\rho(p(x)) - \rho(q(x))]\tau(q(x)). \\ &= [S_{\rho,\tau}(q) - C_{\rho,\tau}(q,q)] - [S_{\rho,\tau}(p) - C_{\rho,\tau}(p,q)] \end{aligned}$$

Note that in general $S_{\rho,\tau}(q) \neq C_{\rho,\tau}(q,q)$; this is because

$$S_{\rho,\tau}(p) - C_{\rho,\tau}(p,p) = \int_{\mathcal{X}} \mathrm{d}x\, f^*(\tau(p(x))).$$

So unless $f(u) = cu$ for constant c, f^* would not vanish. In fact, denote

$$S^*_{\rho,\tau}(p) = -\int_{\mathcal{X}} \mathrm{d}x\, f^*(\tau(p(x))). \tag{20}$$

Then $S^*_{\rho,\tau}(p) = S_{\tau,\rho}(p)$, and

$$S_{\rho,\tau}(p) - C_{\rho,\tau}(p,p) + S^*_{\rho,\tau}(p) = 0 \tag{21}$$

which is, after integrating $\int_{\mathcal{X}} \mathrm{d}x$, a re-write of (17). Therefore,

$$D_{\rho,\tau}(p,q) = S_{\rho,\tau}(p) - C_{\rho,\tau}(p,q) + S^*_{\rho,\tau}(q). \tag{22}$$

Because $D_{\rho,\tau}(p,q)$ is non-negative and vanishes if and only if $p = q$, the function $p \mapsto S_{\rho,\tau}(p) - C_{\rho,\tau}(p,q)$ has its unique maximum at $p = q$. Therefore, minimizing $p \mapsto D_{\rho,\tau}(p,q)$ is equivalent with maximizing $p \mapsto S_{\rho,\tau}(p) - C_{\rho,\tau}(p,q)$.

2.7 Gauge Freedom of the rho-tau Embedding

Because rho-tau embedding has the freedom of two functions, it reduces to the single-function embeddings (either phi- or U-embedding) upon fixing one embedding function.

Divergence. In the phi-embedding, Expression (15) of $D_{\rho,\tau}(p,q)$ reduces to the phi-divergence $D_\phi(p,q)$ for instance if $\rho = id$, the identity function; in this case, $\tau(u) = \log_\phi(u) = f'(u)$.

The U-embedding is also a special case of the rho-tau embedding, with $\rho = id$ identification: $U = f^*$, $\tau = (U')^{-1} = f'$. So phi-divergence (7) and U-divergence (9) are identical. U- and phi-embedding are the same, with $U' = \exp_\phi$, as noted in [11].

Entropy. By virtue of gauge selection $\rho = id$ in the rho-tau embedding, any phi-deformed entropy (3) is a special case of rho-tau entropy (18)

$$S_{\rho,\tau}(p) = S_\phi(\rho(p)).$$

On the other hand, though the rho-tau entropy (18) has two free functions in appearance, it is the result of their function composition that matters. So any rho-tau entropy is also a phi-entropy for a well-chosen ϕ.

The situation with the U-embedding is the same, because U-entropy is identical with phi-entropy:

$$\begin{aligned} H_U(p) &= \int_{\mathcal{X}} \mathrm{d}x \, \left[U((U')^{-1}(p(x))) - p(x)\cdot(U')^{-1}(p(x))\right] \\ &= \int_{\mathcal{X}} \mathrm{d}x \, \left[f^*(f'(p(x))) - p(x)\cdot f'(p(x))\right] = -\int_{\mathcal{X}} \mathrm{d}x f(p(x)) = S_\phi(p). \end{aligned}$$

Cross-entropy. The rho-tau embedding identifies $C_{\rho,\tau}(p,q)$ as the cross-entropy with a dual embedding mechanism, one free function for each of the p, q. In this most general form, however, we do not require that $C_{\rho,\tau}(p,q)$ reduce to either $S_{\rho,\tau}(p)$ or $S^*_{\rho,\tau}(p) \equiv S_{\tau,\rho}(p)$ when $p = q$. This is *different* from the approach of the U-embedding, where its cross-entropy $C_U(p,q)$ is such that $C_U(p,p) = H_U(p)$. It turns out that $C_U(p,q)$ given by (8) equals the rho-tau cross-entropy minus the dual rho-tau entropy (after adopting the $\rho = id$ gauge):

$$C_{\rho,\tau}(p,q) - S^*_{\rho,\tau}(q) = C_U(p,q). \tag{23}$$

Below, we *extend* Eguchi's definition of U cross-entropy by removing the $\rho = id$ restriction. In other words, we can call the left-hand side of (23) U cross-entropy, which depends on two free functions ρ, τ, and obtain from (22)

$$D_{\rho,\tau} = C_U(p,q) - C_U(p,p).$$

2.8 The Normalization Gauge

Let us fix the gauge by $f^* = \tau^{-1}$. In this case, $\int_{\mathcal{X}} \mathrm{d}x\, f^*(\tau(p(x))) = \int_{\mathcal{X}} p(x)\mathrm{d}x = 1$, so $S^*_{\rho,\tau}(p) = S^*_{\rho,\tau}(q) = -1$.

Adopting the $f^* = \tau^{-1}$ gauge (we call this "normalization gauge") implies that

$$\rho(p) = (f^*)'(\tau(p)) = (\tau^{-1})'(\tau(p)) = \frac{1}{\tau'(p)}.$$

So the transformation

$$\lambda : \tau(\cdot) \longrightarrow \frac{1}{\tau'(\cdot)} \equiv (\tau^{-1})'(\tau(\cdot))$$

reflects a transformation of embedding functions. In the phi-embedding language, $\tau \rightarrow \rho$ is simply $\log_\phi \rightarrow \phi$, or the phi-exponentiation operation. This transformation is important in studying phi-exponential family of pdfs (Part II).

Fixing the gauge freedom by normalization simplifies the form of $D_{\rho,\tau}$. Making use of (21), with $S^* = const$, implies that the rho-tau cross-entropy $C_{\rho,\tau}$ and U cross-entropy $C_U(\rho,\tau)$, as given by left-hand side of (23), are equal and are denoted C_0:

$$C_0(p,q) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x)) \cdot \tau(q(x)) = -\int_{\mathcal{X}} \mathrm{d}x\, (\tau^{-1})'(\tau(p(x))) \cdot \tau(q(x))$$

or, in terms of deformed-logarithm notation,

$$C_0(p,q) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x)) \log_\rho(q(x)).$$

Then

$$H_0(p) \equiv C_0(p,p) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x)) \log_\rho(p(x)),$$

with

$$\begin{aligned} D_0(p,q) &= C_0(p,q) - C_0(p,p) \\ &= \int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x)) \cdot (\log_\rho(p(x)) - \log_\rho(q(x))) \\ &= \int_{\mathcal{X}} \mathrm{d}x\, \frac{1}{\tau'(p(x))}\,(\tau(p(x)) - \tau(q(x))). \end{aligned} \tag{24}$$

Note that $D_0 \neq D_\phi$; they both degenerate from $D_{\rho,\tau}$ under different gauges.

We summarize the above conclusions in the following theorem:

Theorem 1. *The (ρ,τ) embedding reduces to special cases upon fixing the gauge as:*

(i) $\rho = id$: *rho-tau divergence $D_{\rho,\tau}$ reduces to deformed phi-divergence D_ϕ with $\tau = f' = \log_\phi$, and to U-divergence D_U with $U = f^*$ and $\tau = f' = (U')^{-1}$;*
(ii) $f^* = \tau^{-1}$: *rho-tau cross-entropy $C_{\rho,\tau}$ reduces to U-cross-entropy as redefined in (23). In this case, $\rho = \phi, \tau = \log_\phi$, i.e., $\tau \rightarrow \rho = (\tau^{-1})' \circ \tau \equiv 1/\tau'$ is taking phi-exponentiation operation;*
(iii) $\rho = \tau$: *rho-tau divergence $D_{\rho,\tau}$ becomes $\int dx(\rho(p(x)) - \rho(q(x)))^2/2$.*

3 Discussion

The main thesis of our paper is that the divergence function $D_{\rho,\tau}$ constructed from (ρ, τ)-embedding subsumes both the phi-divergence D_ϕ constructed from the deformed-log embedding and the U-divergence constructed from the U-embedding. A highlight of our analysis is that the rho-tau divergence $D_{\rho,\tau}$ provides a clear distinction between entropy and cross-entropy as *two* distinct quantities *without* requiring the latter to degenerate to the former. This is significant in terms of the resulting geometry generated by these two quantities (see Part II).

On the other hand, upon fixing the gauge $f^* = \tau^{-1}$ (normalization gauge) renders the rho-tau cross-entropy to be U cross-entropy, where the dual-entropy is constant. In this case, $\tau \leftrightarrow \rho$ is akin to $\log_\phi \leftrightarrow \phi$ transformation encountered in studying normalization of phi-exponential family. A thorough discussion of the geometries induced from the rho-tau divergence and from the phi-exponential family will be given in Part II.

Acknowledgement. The first author is supported by DARPA/ARO Grant W911NF-16-1-0383.

References

1. Amari, S.: Differential-Geometric Methods in Statistics. LNS, vol. 28. Springer, Heidelberg (1985). doi:10.1007/978-1-4612-5056-2
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. Translations of Mathematical Monographs, vol. 191. Am. Math. Soc., Oxford University Press, Oxford (2000). Originally in Japanese (Iwanami Shoten, Tokyo, 1993)
3. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
4. Tsallis, C.: What are the numbers that experiments provide? Quim. Nova **17**, 468 (1994)
5. Kaniadakis, G.: Non-linear kinetics underlying generalized statistics. Phys. A Stat. Mech. Appl. **296**(3), 405–425 (2001)
6. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
7. Naudts, J.: Estimators, escort probabilities, and phi-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004). arXiv:math-ph/0402005
8. Eguchi, S.: Information geometry and statistical pattern recognition. Sugaku Expo. (Amer. Math. Soc.) **19**, 197–216 (2006). Originally Sūgaku 56 (2004) 380 in Japanese
9. Naudts, J.: Generalised Thermostatistics. Springer, London (2011). ISBN: 978-0-85729-354-1
10. Zhang, J.: Nonparametric information geometry: from divergence function to referential-representational biduality on statistical manifolds. Entropy **15**, 5384–5418 (2013)
11. Naudts, J., Anthonis, B.: The exponential family in abstract information theory. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2013. LNCS, vol. 8085, pp. 265–272. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40020-9_28

12. Zhang, J.: On monotone embedding in information geometry. Entropy **17**, 4485–4499 (2015)
13. Newton, N.J.: Information geometric nonlinear filtering. Inf. Dim. Anal. Quantum Prob. Rel. Topics **18**, 1550014 (2015)
14. Zhou, J.: Information theory and statistical mechanics revisited, arXiv:1604.08739

Author Queries

Chapter 24

Query Refs.	Details Required	Author's response
AQ1	Please confirm if the corresponding author is correctly identified. Amend if necessary.	