Information Geometry (2018) 1:79–115
https://doi.org/10.1007/s41884-018-0004-6

RESEARCH PAPER

Rho–tau embedding and gauge freedom in information geometry

Jan Naudts[1] · Jun Zhang[2]

Received: 1 November 2017 / Revised: 27 June 2018 / Published online: 20 August 2018

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 present paper studies parametrized statistical models and the induced geometry using arbitrary embedding functions, comparing single-function approaches (Eguchi's U-embedding and Naudts' deformed-log or phi-embedding) and a two-function embedding approach (Zhang's conjugate rho-tau embedding). In terms of geometry, the rho-tau embedding of a parametric statistical model defines both a Riemannian metric, called "rho-tau metric", and an alpha-family of rho-tau connections, with the former controlled by a single function and the latter by both embedding functions ρ and τ in general. We identify conditions under which the rho-tau metric becomes Hessian and hence the ± 1 rho-tau connections are dually flat. For any choice of rho and tau there exist models belonging to the phi-deformed exponential family for which the rho-tau metric is Hessian. In other cases the rho–tau metric may be only conformally equivalent with a Hessian metric. Finally, we show a formulation of the maximum entropy framework which yields the phi-exponential family as the solution.

Keywords Phi-embedding · U-embedding · Rho–tau embedding · Rho–tau metric · Rho–tau divergence · Rho–tau cross-entropy · U cross-entropy · Phi-exponential model · Escort distribution · Hessian metric · Gauge freedom

✉ Jun Zhang
junz@umich.edu

Jan Naudts
Jan.Naudts@uantwerpen.be

[1] Universiteit Antwerpen, Antwerpen, Belgium

[2] University of Michigan, Ann Arbor, MI, USA

1 Introduction

In classical information geometry [1,3] the Fisher–Rao metric, as a Riemannian metric on the manifold of parametric probability models, is accompanied by a family of α-connections $\Gamma^{(\alpha)}$ with a dualistic structure such that $\Gamma^{(\alpha)}$ and $\Gamma^{(-\alpha)}$ jointly preserve the metric. This so-called "α-geometry" is induced by the family of α-divergence functions which include the Kullback-Leibler divergence as a special case ($\alpha = \pm 1$). Furthermore, when the statistical model belongs to the exponential family, then the connections $\Gamma^{(\pm 1)}$ are dually flat.

Zhang [20,23] carefully delineated the different roles played by the interpolation parameter α in information geometry:

1. it parametrizes the divergence function (as in α-divergence);
2. it parametrizes the monotone embedding of probabilities (as in α-embedding);
3. it parametrizes the convex combination of connections (as in α-connection).

A thorough understanding of the subtleties of these various roles of α in α-geometry leads not only to the class of two-parameter (α, β)-divergence (generalizing α-divergence in different ways), which nevertheless results in the parametric family of $\alpha\beta$-connections, with $\alpha \cdot \beta$ as a single parameter [20], but also to the more profound notion of reference-representation biduality uniquely embodied in information geometry [20,22].

There has been considerable interest in generalizing the "standard model" and the corresponding exponential (and its dual, mixture) family of probability functions. By generalizing, we mean that the dualistic α-geometry is still preserved while one relaxes from the restrictive exponential (or mixture) family. The generalizations are often achieved in the context of various monotone embedding functions, from α-embedding (power function) to arbitrary deformed exponential embedding function, such as *phi-embedding* [11] and *U-embedding* [6]. Zhang [20,22,23] uses two arbitrary functions, referred to as *conjugate rho–tau embedding*. Our paper surveys these approaches and their links, with the goal of providing a unifying account in generalizing Amari's α-geometry with its characteristic biduality (reference duality and representation duality [21]). A particular outcome is the demonstration of the dually flat nature of $\Gamma^{(\pm 1)}$, despite of considerable relaxation both in terms of deforming the exponential family and the canonical divergence function.

1.1 The standard model

1.1.1 Fisher–Rao metric and α-connections

Let be given a measure space $(\mathcal{X}, \mathrm{d}x)$. Let $\mathcal{M}$ denote the space of probability density functions defined on the sample space $\mathcal{X}$. A parametric family of density functions, $p^\theta \equiv p(\cdot|\theta)$, called a parametric statistical model, is the association $\theta \mapsto p(\cdot|\theta)$ of a point $\theta = [\theta^1, \ldots, \theta^n]$ in a connected open subset $\mathcal{D}$ of $\mathbb{R}^n$ with a probability density function p^θ in $\mathcal{M}$. The elements of the parametric statistical model form a Riemannian manifold $\mathbb{M}$. For simplicity we assume that a single chart $p^\theta \mapsto \theta$ covers all of $\mathbb{M}$, so that

$$\mathbb{M} = \left\{ p^{\theta} \in \mathcal{M} : \theta \in \mathcal{D} \subset \mathbb{R}^n \right\} \subset \mathcal{M}.$$

The Fisher–Rao metric and the α-connections are given by

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \left\{ p(x|\theta) \frac{\partial \log p(x|\theta)}{\partial \theta^i} \frac{\partial \log p(x|\theta)}{\partial \theta^j} \right\}; \tag{1}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \frac{\partial p(x|\theta)}{\partial \theta^k} \left(\frac{1-\alpha}{2} \frac{\partial \log p(x|\theta)}{\partial \theta^i} \frac{\partial \log p(x|\theta)}{\partial \theta^j} + \frac{\partial^2 \log p(x|\theta)}{\partial \theta^i \partial \theta^j} \right). \tag{2}$$

The α-connections satisfy the dualistic relation

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = \Gamma^{(-\alpha)}_{ij,k}(\theta). \tag{3}$$

Here, $*$, denotes conjugate (dual) connection. The pair of conjugate connections preserves the dual pairing of vectors in the tangent space with co-vectors in the cotangent space when the tangent and cotangent spaces are mapped to each other by the Riemannian metric. Any Riemannian manifold with its metric, g, and conjugate connections, Γ, Γ^*, given in the form of Eqs. (1)–(3), is called a *statistical manifold* (in the narrower sense) and is denoted as $\{\mathbb{M}, g, \Gamma^{(\pm\alpha)}\}$. In the broader sense, a statistical manifold $\{\mathbb{M}, g, \Gamma, \Gamma^*\}$ is a differentiable manifold equipped with a Riemannian metric g and a pair of torsion-free conjugate connections $\Gamma \equiv \Gamma^{(1)}$, $\Gamma^* \equiv \Gamma^{(-1)}$ which jointly preserve the metric g, without necessarily requiring g and Γ, Γ^* to take the forms of Eqs. (1)–(3).

1.1.2 Exponential and mixture families

An exponential family of probability density functions is defined as

$$p^{(e)}(x|\theta) = \exp\left(\sum_i \theta^i \, F_i(x) - \Phi(\theta) \right) \tag{4}$$

where θ is its canonical parameter and $F_i(x)$ ($i = 1, \cdots, n$) is a set of linearly independent functions with the same support in $\mathcal{X}$, and the cumulant generating function ("potential function") $\Phi(\theta)$ is:

$$\Phi(\theta) = \log \int_{\mathcal{X}} \mathrm{d}x \left\{ \exp\left(\sum_i \theta^i \, F_i(x) \right) \right\}. \tag{5}$$

Substitution of (4) into (1) and (2) results in the Fisher metric

$$g_{ij}(\theta) = \int_{\mathcal{X}} p^{(e)}(x|\theta) \left(F_i(x) - \int_{\mathcal{X}} p^{(e)} F_i(x) \mathrm{d}x \right) \left(F_j(x) - \int_{\mathcal{X}} p^{(e)} F_j(x) \mathrm{d}x \right) \mathrm{d}x,$$

which can be written as

$$g_{ij}(\theta) = \frac{\partial^2 \Phi(\theta)}{\partial \theta^i \partial \theta^j}. \tag{6}$$

The α-connections can be written as

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \frac{1-\alpha}{2} \frac{\partial^3 \Phi(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}.$$

The α-connection for the exponential family is dually flat when $\alpha = \pm 1$. In particular, all components of $\Gamma^{(1)}_{ij,k}$ vanish on the manifold of an exponential family.

On the other hand, the mixture family:

$$p^{(m)}(x|\theta) = \sum_i \theta^i F_i(x), \tag{7}$$

when viewed as a manifold charted by its mixture parameter θ, with the constraints $\sum_i \theta^i = 1$ and $\int_X F_i(x) d\mu = 1$, turns out to have identically zero $\Gamma^{(-1)}_{ij,k}$.

The connections, $\Gamma^{(1)}$ and $\Gamma^{(-1)}$, are also called the exponential and mixture connections, or the e- and m-connection, respectively.

1.2 α-Embedding function

Amari [1,3] considered a one-parameter family of denormalized probability density functions $p^{(\alpha)}(\cdot|\theta)$ defined by $p^{(\alpha)}(x|\theta) = p(x)$ with

$$l^{(\alpha)}\left(p(x)\right) = F_0\left(x\right) + \sum_i \theta^i F_i\left(x\right). \tag{8}$$

The *α-embedding* function $l^{(\alpha)} : \mathbb{R}^+ \to \mathbb{R}$, is defined as

$$l^{(\alpha)}(t) = \left\{ \begin{array}{cc} \log t & \alpha = 1 \\ \frac{2}{1-\alpha} t^{(1-\alpha)/2} & \alpha \neq 1 \end{array} \right\}. \tag{9}$$

Under α-embedding, the denormalized density functions form the so-called *α-affine manifold*, see, [3], p. 46. It is remarkable that the Fisher–Rao metric and the α-connections, under such α-representation, have the following expressions:

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \left\{ \frac{\partial l^{(\alpha)}(p(x|\theta))}{\partial \theta^i} \frac{\partial l^{(-\alpha)}(p(x|\theta))}{\partial \theta^j} \right\}, \tag{10}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \left\{ \frac{\partial^2 l^{(\alpha)}(p(x|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial l^{(-\alpha)}(p(x|\theta))}{\partial \theta^k} \right\}. \tag{11}$$

Clearly, for any given α value, the components of $\Gamma^{(\alpha)}$ are all identically zero on the (unnormalized) α-affine manifold, by virtue of the definition (8) of the α-family. Hence, the $\pm\alpha$-connections are dually flat.

1.3 A plethora of probability embeddings

There have been various attempts at generalizing the standard exponential model of normalized probability density functions. Efforts considered here have been centered on "deforming" the exponential function to other functional forms, where the embedding functions of the α-family are treated as deformed logarithm functions, whose inverse functions are then deformed exponentials. Better known examples are q-exponential functions [17] and the κ-functions [7]. More general deformations were introduced in [11]. The corresponding deformed exponential families coincide with the models of U-statistics [6]. From the point of view of [20,22,23] these models involve generalized embeddings which fit under a universal framework of conjugate rho-tau embeddings.

(i) *q-logarithmic embedding* Tsallis [17] investigates the equilibrium distribution of statistical physics which is obtained by maximization of the Boltzmann–Gibbs–Shannon entropy under constraints. He replaces the entropy function by a q-dependent entropy, $q \in \mathbb{R}$. This results in a deformed version of statistical physics. The q-logarithmic/exponential functions were introduced in [18]:

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

Note that q-embedding and α-embedding functions are different: $\log_q(u) = l^{(\alpha)}(u) - 2/(1-\alpha)$ with $\alpha = 2q - 1$. Like α-embedding, q-embedding reduces to the standard logarithm as q tends to 1.

(ii) *κ-logarithmic embedding* An alternative to the q-deformed exponential model for statistical physics is Kaniadakis' κ-model [7], 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.

The ϕ-, U- and (ρ, τ)-embedding are monotone embeddings which rely on one or two free functions. So instead of using a one-parameter family of functions which include the logarithm/exponential function for a particular parameter value, arbitrary functions are used which replace the logarithm/exponential function. They are the main focus of this paper. The phi-model [11], U-model [6], and rho-tau model [20] were independently conceived around 2004 under different motivations.

1.4 Goals, organization, and notations

Our goal is to provide a unified theory of monotone embedding which generalizes that of logarithmic embedding and the classic α-geometry. Specifically, we revisit the divergence function, cross-entropy and entropy determined by rho-tau embedding [20], their induced α-geometry with respect to the phi-deformed exponential families [11], and show a unification of the "deformed" approach of [11] and conjugate embedding approach of [20]. It is also shown that the independently proposed U-embedding [6] is identical with phi-embedding in terms of divergence function and entropy functions, both being subsumed by the rho–tau embedding. The duality in rho-tau entropy is shown to be important in the formulation of the generalized maximum entropy principle, the solution of which is the phi-exponential family.

In Sect. 2, we first review the deformed logarithm, $\log_\phi$, and the deformed exponential, $\exp_\phi$. Then we point out that $\log_\phi$ and $\exp_\phi$ are nothing but an arbitrary pair of mutually inverse monotone functions, and can be represented 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 deformed entropy and cross-entropy is reviewed, as well as their construction starting from the U-embedding. Then, we review the rho-tau embedding, which provides two independently chosen embedding functions. We explicitly identify its entropy and cross-entropy. Theorem 1 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.

In Sect. 3 we explore the freedom of choosing two functions ρ and τ, such that they lead to the same weighting function associated with the Riemannian metric. We call it the *gauge freedom*. Two prominent gauges, plus their duals, are studied. They lead to the entropy and cross-entropy functions given by phi/U-embedding and to those given by Tsallis.

In Sect. 4, we study the Riemannian metric induced from the rho-tau divergence (and equivalently, rho-tau cross-entropy), as well as induced from the entropy and dual entropy functions. Emphasis is put on the gauge freedom which is left once the metric is fixed. The metric tensor absorbs only one of the two degrees of freedom offered by the independent choice of two strictly increasing functions rho and tau. We then provide a characterization of conditions under which rho–tau metric is Hessian. The rho–tau connections were also investigated.

In Sect. 5, we study deformed exponential family of probability models, and show the Riemannian geometry they induce. We show that each phi-exponential family is associated with two special rho–tau metrics, (i) a Hessian one related to its entropy and (ii) a non-Hessian one that is conformally equivalent to the Hessian of a normalization function. We shown how these models are related to the maximum entropy principle.

In the final section we provide a summary and discusssions.

Throughout the paper it is assumed that two strictly increasing differentiable functions ρ and τ are given. The rho–tau divergence induces a metric tensor g on finite-dimensional manifolds of probability distributions and turns them into Riemannian manifolds. Here we assume regularity conditions such that the relevant integrals all exist, Cf. [10]. A preliminary version of this report appeared in [15,24].

2 Divergence, entropy, and cross-entropy

2.1 "Deforming" exponential and logarithmic functions

Naudts [11,13] 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 such that $1/\phi(v)$ is integrable. 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:

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

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

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

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

$$\begin{aligned} \psi\,(u) &= \phi\left(\exp_\phi(u)\right), \quad u \in \text{range}\left(\log_\phi\right), \\ \phi\,(u) &= \psi\left(\log_\phi(u)\right), \quad u > 0. \end{aligned}$$

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

As examples, $\phi(u) = u$ gives rise to the classic natural logarithm and exponential. The choice $\phi(u) = u^q$, $q \neq 1$ reproduces the q-deformed logarithm and exponential, as introduced by Tsallis [18] and mentioned in the introduction. Taking $\phi(u) = u/(1+u)$ leads to (see, [10,16]) $\log_\phi(u) = u - 1 + \log(u)$. Taking $\phi(u) = u(1 + \epsilon u)$ leads to (see, [25])

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

2.2 Deformed entropy and deformed divergence functions

The phi-entropy of the probability distribution p is defined by (see [11,13])

$$S_\phi(p) = -\mathbb{E}_p \log_\phi p + \int_{\mathcal{X}} \mathrm{d}x \int_0^{p(x)} \mathrm{d}u \, \frac{u}{\phi(u)} + \text{ constant.} \tag{12}$$

By partial integration one obtains an equivalent expression

$$S_\phi(p) = -\int_{\mathcal{X}} \mathrm{d}x \int_1^{p(x)} \mathrm{d}u \, \log_\phi(u) + \text{ constant.} \tag{13}$$

For the standard logarithm is $\phi(u) = u$. Then the above expression coincides with the well-known entropy of Boltzmann–Gibbs–Shannon

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

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

$$D_\phi(p, q) = \int_{\mathcal{X}} \mathrm{d}x \int_{q(x)}^{p(x)} \mathrm{d}v \left[\log_\phi(v) - \log_\phi(q(x))\right]. \tag{14}$$

An equivalent expression is

$$D_\phi(p, q) = S_\phi(q) - S_\phi(p) - \int_{\mathcal{X}} \mathrm{d}x \, [p(x) - q(x)] \log_\phi(q(x)). \tag{15}$$

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

$$S_\phi(p) = -\int_{\mathcal{X}} \mathrm{d}x \, f(p(x)) + \text{ constant,} \tag{16}$$

$$D_\phi(p, q) = \int_{\mathcal{X}} \mathrm{d}x \left\{f(p(x)) - f(q(x)) - [p(x) - q(x)] f'(q(x))\right\}. \tag{17}$$

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, U entropy, and U cross-entropy

Eguchi [6] introduces the U-divergence, which is essentially the Bregman divergence under a strictly convex function U coupled with an embedding using $\psi_{\mathrm{U}} \equiv (U')^{-1}$.

The U cross-entropy $C_U(p, q)$ is defined as:

$$C_U\,(p, q) = \int_{\mathcal{X}} \mathrm{d}x\ \{U\,(\psi_{\mathrm{U}}\,(q\,(x))) - p\,(x) \cdot \psi_{\mathrm{U}}\,(q\,(x))\}\,, \tag{18}$$

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

$$\begin{aligned} D_U(p, q) &= C_U(p, q) - H_U(p, p) \\ &= \int_{\mathcal{X}} \mathrm{d}x \left\{ U\,(\psi_{\mathrm{U}}(q(x))) - U\,(\psi_{\mathrm{U}}(p(x))) - p(x)\,[(\psi_{\mathrm{U}}(q(x)) - \psi_{\mathrm{U}}(p(x))] \right\}. \end{aligned} \tag{19}$$

Note that the U-embedding only has one arbitrarily chosen function, as does the phi-embedding. In fact, it was noted in [14] that the U-divergence and the phi-divergence of the previous section map onto each other when the derivative U' of U is considered as a deformed exponential function.

2.4 Conjugate rho-tau embedding

In contrast with the "single function" embedding of the phi-model and the U-model, Zhang's [20] rho–tau framework uses *two* arbitrarily and independently chosen monotone functions (see also [23]). 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

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

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

$$f''((f^*)'(u)) \cdot (f^*)''(u) = (f^*)''(f'(u)) \cdot f''(u) = 1. \tag{20}$$

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 the group operation, an observation made in [20,23]. 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

$$f'(\rho(u)) = \tau(u) \quad \text{and} \quad (f^*)'(\tau(u)) = \rho(u). \tag{21}$$

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 (ρ, f), (ρ, f^*), (τ, 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:

$$\begin{aligned} f''(\rho(u))\,\rho'(u) &= \tau'(u), \qquad (f^*)''(\tau(u))\,\tau'(u) = \rho'(u), \\ f''(\rho(u))\,(\rho'(u))^2 &= (f^*)''(\tau(u))\,(\tau'(u))^2, \\ f''(\rho(u))\,(f^*)''(\tau(u)) &= 1. \end{aligned} \tag{22}$$

The (ρ, τ)-embedding mechanism can have another equivalent representation. Denote $f \circ \rho = F$, $f^* \circ \tau = G$. We seek to use F, G as independently chosen functions from which ρ and τ are derived. From

$$\tau \cdot \rho' = F', \qquad \rho \cdot \tau' = G',$$

and

$$F(u) + G(u) = \rho(u)\tau(u),$$

we obtain

$$\frac{\rho'}{\rho} \cdot (F + G) = F',$$

or

$$\frac{d \log \rho}{du} = \frac{F'}{F + G}.$$

Thus, we obtain that

$$\log \rho(u) = \int^u \frac{F'(s)\mathrm{d}s}{F(s) + G(s)} = \int^u \frac{\mathrm{d}F(s)}{F(s) + G(s)}.$$

and similarly

$$\log \tau(u) = \int^u \frac{G'(s)\mathrm{d}s}{F(s) + G(s)} = \int^u \frac{\mathrm{d}G(s)}{F(s) + G(s)}.$$

So this gives ρ, τ in terms of F, G.

2.5 Divergence of the rho-tau embedding

Zhang [20] introduces[1] the rho-tau divergence (see Proposition 6 of [20])

$$D_{\rho,\tau}(p,q) = \int_{\mathcal{X}} \mathrm{d}x \, \left\{ f(\rho(p(x))) + f^*(\tau(q(x))) - \rho(p(x))\tau(q(x)) \right\}, \quad (23)$$

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

Proposition 1 *Expression* (23) *can be written as*

$$\begin{aligned} D_{\rho,\tau}(p,q) &= \int_{\mathcal{X}} \mathrm{d}x \left\{ f(\rho(p(x))) - f(\rho(q(x))) - [\rho(p(x)) - \rho(q(x))]\tau(q(x)) \right\} \\ &= \int_{\mathcal{X}} \mathrm{d}x \int_{q(x)}^{p(x)} [\tau(v) - \tau(q(x))] \, \mathrm{d}\rho(v) \\ &= \int_{\mathcal{X}} \mathrm{d}x \int_{\rho(q(x))}^{\rho(p(x))} \mathrm{d}u \, \left[f'(u) - f'(\rho(q(x))) \right]. \end{aligned} \quad (24)$$

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

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

The "reference-representation biduality" [20,22,23] reveals as

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

It can be easily verified that the rho-tau divergence satisfies the following generalized Pythagorean equality for any three probability functions p, q, r

$$\begin{aligned} &D_{\rho,\tau}(p,q) + D_{\rho,\tau}(q,r) - D_{\rho,\tau}(p,r) \\ &\quad = \int_{\mathcal{X}} \mathrm{d}x \, \{[\rho(p(x)) - \rho(q(x))][\tau(r(x)) - \tau(q(x))]\}. \end{aligned} \quad (25)$$

2.6 Entropy and cross-entropy of the rho–tau embedding

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

$$S_{\rho,\tau}(p) = -\int_{\mathcal{X}} \mathrm{d}x \; f(\rho(p(x))) + \text{constant}, \quad (26)$$

[1] The original definition as found in [20,23] uses the notation $D_{f,\rho}(p,q)$ and treats f and ρ as independent. Under the present notation $D_{\rho,\tau}(p,q)$ the function f is taken to depend on ρ, τ. The difference is only notational and inconsequential.

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

$$\begin{aligned} S_{\rho,\tau}(p) &= -\int_{\mathcal{X}} \mathrm{d}x \int^{\rho(p(x))} f'(v)\mathrm{d}v + \text{ constant} \\ &= -\int_{\mathcal{X}} \mathrm{d}x \int^{p(x)} \tau(u)\mathrm{d}\rho(u) + \text{ constant}. \end{aligned} \tag{27}$$

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

$$C_{\rho,\tau}(p,q) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x))\tau(q(x)). \tag{28}$$

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

The rho–tau divergence can then be given by

$$\begin{aligned} D_{\rho,\tau}(p,q) &= S_{\rho,\tau}(q) - S_{\rho,\tau}(p) - \int_{\mathcal{X}} \mathrm{d}x\, \left[\rho(p(x)) - \rho(q(x))\right]\tau(q(x)) \\ &= \left[S_{\rho,\tau}(q) - C_{\rho,\tau}(q,q)\right] - \left[S_{\rho,\tau}(p) - C_{\rho,\tau}(p,q)\right]. \end{aligned} \tag{29}$$

Note that, unlike the standard case, 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

$$S^*_{\rho,\tau}(p) = -\int_{\mathcal{X}} \mathrm{d}x\, f^*(\tau(p(x))). \tag{30}$$

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

$$S_{\rho,\tau}(p) - C_{\rho,\tau}(p,p) + S^*_{\rho,\tau}(p) = 0 \tag{31}$$

which is, after integrating $\int_{\mathcal{X}} \mathrm{d}x$, a re-write of (25). Therefore,

$$D_{\rho,\tau}(p,q) = C_{\rho,\tau}(p,q) - S_{\rho,\tau}(p) - S^*_{\rho,\tau}(q). \tag{32}$$

Because rho-tau cross-entropy does not degenerate to rho-tau entropy in general:

$$C_{\rho,\tau}(p,p) \neq S_{\rho,\tau}(p),$$

we can also define the modified cross-entropy:

$$\overline{C}_{\rho,\tau}(p,q) = C_{\rho,\tau}(p,q) - S^*_{\rho,\tau}(q). \tag{33}$$

The main properties of this modified version of cross-entropy $\overline{C}_{\rho,\tau}(p,q)$ are

1. $\overline{C}_{\rho,\tau}(p,p) = S_{\rho,\tau}(p)$. Indeed, (32) implies that

$$\begin{aligned}\overline{C}_{\rho,\tau}(p,p) &= C_{\rho,\tau}(p,p) - S^*_{\rho,\tau}(p)\\ &= S_{\rho,\tau}(p) - D_{\rho,\tau}(p,p)\\ &= S_{\rho,\tau}(p).\end{aligned}$$

2. From (32), the previous result and the definition (33) follows that

$$D_{\rho,\tau}(p,q) = \overline{C}_{\rho,\tau}(p,q) - \overline{C}_{\rho,\tau}(p,p). \tag{34}$$

2.7 Rho–tau divergence from convex $D^{(\alpha)}_{f,\rho}(p,q)$-divergence

Refs. [20–23] studied the following general divergence function $D^{(\alpha)}_{f,\rho}(p,q)$ from the perspective of convex analysis (with $\alpha \in \mathbb{R}$)

$$\begin{aligned}D^{(\alpha)}_{f,\rho}(p,q) = \frac{4}{1-\alpha^2} \times \int_{\mathcal{X}} \mathrm{d}x \; \Bigg\{ &\frac{1-\alpha}{2} f(\rho(p)) + \frac{1+\alpha}{2} f(\rho(q))\\ &- f\left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q)\right)\Bigg\}.\end{aligned} \tag{35}$$

Clearly, the role of α is to effect an exchange of the position of p, q

$$D^{(-\alpha)}_{f,\rho}(p,q) = D^{(\alpha)}_{f,\rho}(q,p).$$

Rho–tau divergence $D_{\rho,\tau}(p,q)$ arises as a special form of the above convex $D^{(\alpha)}_{f,\rho}$-divergence function:

$$\begin{aligned}\lim_{\alpha\to 1} D^{(\alpha)}_{f,\rho}(p,q) &= D_{\rho,\tau}(p,q) = D_{\tau,\rho}(q,p);\\ \lim_{\alpha\to -1} D^{(\alpha)}_{f,\rho}(p,q) &= D_{\rho,\tau}(q,p) = D_{\tau,\rho}(p,q);\end{aligned}$$

with $f' \circ \rho = \tau$ (and equivalent $(f^*)' \circ \tau = \rho$, with f^* denoting convex conjugate of f).

Though in $D^{(\alpha)}_{f,\rho}(p,q)$ the two free functions are f (a strictly convex function) and ρ (a strictly monotone increasing function), as reflected in its subscripts, there is only notational difference from the ρ, τ specification of two function's choice. This is because for f, f^*, ρ, τ, a choice of any two functions (one of which would have to be either ρ or τ) would specify the remaining two; see footnote 1 and Sect. 2.4.

3 Fixing the gauge

Zhang's conjugate rho–tau embedding involves two freely chosen functions. However, the induced Riemannian metric, called rho-tau metric tensor (to be introduced in the next section) depends on a single function ψ which is the following combination of the functions ρ and τ

$$\psi(u) = \frac{1}{\rho'(u)\tau'(u)}. \tag{36}$$

Many choices of ρ and τ give rise to the same function ψ. We call this the "gauge freedom". In the physics literature gauge theories cope with redundant degrees of freedom, either by fixing the gauge or by introduction of equivalence classes. In the present theory, the symmetry between the functions ρ and τ implies that their exchange leads to equivalent theories. The simplest way to deal with gauge freedom is by breaking the symmetry, in the present context by assigning a different role to ρ, respectively τ. For instance, ρ can be used for embedding the probability distribution while τ is then used to fix the entropy or score variables of the corresponding statistical theory. Two specific types of gauges are now considered in more detail: Type I gauge where either ρ or τ is identity, and Type II gauge where ρ and τ are linked through the deformed logarithm/exponential transformation.

3.1 Rho-id gauge (ρ = id, τ = $\log_\psi$)

This gauge is characterized by $\rho = \text{id}$, the identity function, and $\tau = \log_\psi$. In this case, $\rho' = 1$ and $\tau' = 1/\psi$, satisfying (36).

Compare expression (27) of the rho–tau entropy with that of the phi-deformed entropy as given by (13). They coincide up to an additive constant when the choice $\rho = \text{id}$ and $\tau(u) = \log_\phi(u)$ are made. This means that the function ψ, defined by (36), can be identified with the function ϕ of the phi-deformation formalism. With these choices one has

$$\rho = \text{id}, \quad f' = \tau = \log_\phi, \quad (f^*)' = \exp_\phi .$$

In the notations of Eguchi this becomes

$$\rho = \text{id}, \quad f' = \tau = \psi_{\text{U}}, \quad f^* = U, \quad (f^*)' = U',$$

where ψ_{U} is the inverse function of U', see Sect. 2.3.

Divergence Expression (24) of $D_{\rho,\tau}(p, q)$ reduces to the phi-divergence $D_\phi(p, q)$, as given by (14), and to the U-divergence (19). Phi-divergence and U-divergence coincide with $U' = \exp_\phi$, as noted in [14].

Entropy As mentioned earlier, in the present gauge the rho–tau entropy coincides with the phi-deformed entropy (12). This suggests that the rho–tau entropy is more general than the phi-deformed entropy. However, although the rho-tau entropy (26) has two free functions in appearance, it is the result of their function composition which matters. So any rho–tau entropy is also a phi-entropy for a well-chosen function ϕ.

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

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

Cross-entropy In the present gauge the rho–tau cross-entropy (28) becomes

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

The cross-entropy $C_U(p,q)$ introduced by Eguchi (see (18)) contains an additional term

$$C_U(p,q) = C_{\rho,\tau}(p,q) + \int_{\mathcal{X}} \mathrm{d}x\ U(\psi(q(x))). \tag{37}$$

Since in the present gauge $f^* = U$ with $(f^*)' = U'$, this additional term is nothing but the negative of dual entropy $S^*_{\rho,\tau}$

$$S^*_{\rho,\tau}(q) = -\int_{\mathcal{X}} \mathrm{d}x\ U\left(\psi(q(x))\right).$$

Therefore,

$$C_U(p,q) = \overline{C}_{\rho,\tau}(p,q),$$

which satisfies $C_U(p,p) = H_U(p)$.

3.2 Tau-id gauge ($\tau = \mathrm{id}$, $\rho = \log_\psi$)

This gauge is characterized by $\tau = \mathrm{id}$ and $\rho = \log_\psi$. This gauge is checked to satisfy (36). It is needed a number of times in what follows. Because of the rho–tau duality much of the previous section can be repeated with obvious modifications.

The rho-id and tau-id gauges are collectively called Type I gauges.

3.3 Constant entropy gauge ($\rho = \log_\tau$, $\log\tau = \log_\psi$)

This gauge is characterized by $f \circ \rho = \mathrm{id}$, or $f = \rho^{-1}$.

From (26) then follows that the rho–tau entropy $S_{\rho,\tau}(p)$ is a constant independent of the probability distribution p:

$$\begin{aligned} S_{\rho,\tau}(p) &= -\int_{\mathcal{X}} f(\rho(p(x)))\,\mathrm{d}x + \text{constant} \\ &= -\int_{\mathcal{X}} p(x)\,\mathrm{d}x + \text{constant} = -1 + \text{constant}. \end{aligned}$$

In this case, $\rho'\tau = 1$, so $\rho' = 1/\tau = (\log_\tau)'$, so

$$\rho = \log_\tau, \quad \tau = \exp_\rho .$$

From $\rho'\tau' = 1/\psi$, we obtain $(\log\tau)' = 1/\psi = (\log_\psi)'$. Therefore

$$\log\tau = \log_\psi .$$

Cross-entropy Integration of $\rho' = (\log_\tau)'$ gives $\rho = \log_\tau +$ constant. This constant may be omitted. Using $\rho = \log_\tau$ one can write

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

Divergence and dual entropy The rho–tau divergence (23) takes on the simplified form

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

which reminds of (34).

3.4 Constant-S^* gauge ($\tau = \log_\rho$, $\log\rho = \log_\psi$)

Dual to the constant-S gauge, this gauge is characterized by $f^* \circ \tau = \text{id}$, or $f^* = \tau^{-1}$. This implies $\rho\tau' = 1$, so $\tau' = 1/\rho$. Therefore, this gauge is same as taking $\tau = \log_\rho$.

Because of the rho–tau duality the conclusions of the previous gauge can be adapted. In particular, $(\log\rho)' = 1/\psi$.

Entropy and cross-entropy In the present gauge the rho-tau cross-entropy (28) becomes

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

Because in this gauge $S^*(p)$ is a constant independent of p, the same expression holds for the modified cross-entropy $\overline{C}_{\rho,\tau}(p,q)$. Therefore the rho–tau entropy can be written as

$$S_{\rho,\tau}(p) = \overline{C}_{\rho,\tau}(p,p) = -\int_{\mathcal{X}} \mathrm{d}x\,\rho(p(x))\log_\rho(p(x)) + \text{ constant.} \tag{38}$$

Divergence In this case,

$$D_{\rho,\tau}(p,q) = C_{\rho,\tau}(p,q) - C_{\rho,\tau}(p,p). \tag{39}$$

Because $S^* =$ constant, this also gives

$$D_{\rho,\tau}(p,q) = \overline{C}_{\rho,\tau}(p,q) - \overline{C}_{\rho,\tau}(p,p),$$

in agreement with (34). Now write (39) as

$$D_{\rho,\tau}(p,q) = \int_{\mathcal{X}} \mathrm{d}x\, \rho(p(x)) \left[\log_\rho(p(x)) - \log_\rho(q(x))\right]. \tag{40}$$

Expressions (38) and (40) look very similar to the standard expressions for the Boltzmann–Gibbs–Shannon entropy and the Kullback–Leibler divergence, respectively.

The constant-S or constant-S^* gauge is called a Type II gauge.

4 Riemannian geometry under rho–tau embedding

We now investigate the Riemannian geometry related to the rho-tau embedding, and expect a full generalization to Amari's α-geometry as reviewed in Sect. 1. Throughout this section we consider a parametrized statistical model $\theta \mapsto p^\theta$.

4.1 The metric tensor

The rho–tau divergence $D_{\rho,\tau}(p,q)$ can be used (see [20,22,23]) to define a metric tensor $g(\theta)$ by

$$g_{ij}(\theta) = -\partial_i \partial'_j D_{\rho,\tau}(p^\theta, p^{\theta'})\Big|_{\theta'=\theta}, \tag{41}$$

with $\partial_i = \partial/\partial\theta^i$ and $\partial'_j = \partial/\partial\theta'^j$. One has

$$g_{ij}(\theta) = -\partial_i \partial'_j C_{\rho,\tau}(p^\theta, p^{\theta'})\Big|_{\theta'=\theta},$$

and also

$$g_{ij}(\theta) = \partial_i \partial_j D_{\rho,\tau}(p, p^\theta)\Big|_{p=p^\theta}.$$

A short calculation gives

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x\, \left[\partial_i \rho(p^\theta(x))\right] \left[\partial_j \tau(p^\theta(x))\right]. \tag{42}$$

Because $\tau = f' \circ \rho$, the rho–tau metric $g(\theta)$ also takes the form

$$\begin{aligned} g_{ij}(\theta) &= \int_{\mathcal{X}} \mathrm{d}x \, \left[\partial_i \rho(p^\theta(x))\right] \left[\partial_j f'(\rho(p^\theta(x)))\right] \\ &= \int_{\mathcal{X}} \mathrm{d}x \, f''(\rho(p^\theta(x))) \left[\partial_i \rho(p^\theta(x))\right] \left[\partial_j \rho(p^\theta(x))\right]. \end{aligned} \tag{43}$$

This shows that the matrix $g_{ij}(\theta)$ is symmetric. Moreover, it is positive-definite, because the derivatives ρ' and f'' are strictly positive and the matrix with entries $\left[\partial_j \rho(p^\theta(x))\right] \left[\partial_i \rho(p^\theta(x))\right]$, when pre- and post-multiplied with any vector, gives rise to a positive real number. Finally, $g(\theta)$ is covariant, so g is indeed a metric tensor on the Riemannian manifold p^θ. From (42) follows that it is invariant under the exchange of ρ and τ.

4.2 Freedom of choice of the rho–tau metric

Remarkably, (22) can be used to write (43) as

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, (f^*)''(\tau(p^\theta(x))) \left[\partial_i \tau(p^\theta(x))\right] \left[\partial_j \tau(p^\theta(x))\right].$$

Hence, the gauge freedom of choosing the metric g_{ij} under the rho embedding, by choosing an arbitrary function f, also exists when under the tau embedding.

Write the rho–tau metric g_{ij} as

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \frac{1}{\psi(p^\theta)} \left[\partial_i p^\theta(x)\right] \left[\partial_j p^\theta(x)\right], \tag{44}$$

where $\psi = 1/(\rho'\tau')$ is the function as in (36). So despite of the two independent choices of embedding functions ρ and τ, the metric tensor g_{ij} is determined by one function ψ only.

There is another way of looking at the functional freedom in the g_{ij} metric tensor. Taking a look at (43) reveals that we can choose to specify the function f given any embedding function ρ. So specifying ψ or specifying f achieves the same purpose.

Although the metric tensor g_{ij} is invariant under changes of rho and tau which leave ψ unchanged, other quantities such as the entropy, cross- entropy and divergence function are not. This gauge freedom, which is left once the function ψ and hence the metric tensor is fixed, explains why specific choices of ρ and τ simplify the relation between rho–tau expressions and expressions found in the literature.

4.3 Tangent vectors

Let us now introduce the plane tangent to the rho embedding of the statistical model p^θ. A similar construction can be done for the tau embedding.

From the form of rho–tau metric

$$g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \frac{\rho'(p^\theta(x))}{\tau'(p^\theta(x))} \left[\partial_i \tau(p^\theta(x))\right] \left[\partial_j \tau(p^\theta(x))\right],$$

we introduce a bilinear form $\langle \cdot, \cdot \rangle$ defined on pairs of random variables $u(x)$, $v(x)$

$$\langle u, v \rangle_\theta = \int_{\mathcal{X}} \mathrm{d}x \, \frac{\rho'(p^\theta(x))}{\tau'(p^\theta(x))} \, u(x) \, v(x). \tag{45}$$

Introduce the notation $X^\theta(x) = \tau(p^\theta(x))$, so

$$g_{ij}(\theta) = \langle \partial_i X^\theta, \partial_j X^\theta \rangle_\theta.$$

For any random variable u, it holds that

$$\partial_j \int_{\mathcal{X}} \mathrm{d}x \, \rho(p^\theta(x)) u(x) = \int_{\mathcal{X}} \mathrm{d}x \, \frac{\rho'(p^\theta(x))}{\tau'(p^\theta(x))} \partial_j \tau(p^\theta(x)) u(x) = \langle \partial_j X^\theta, u \rangle_\theta.$$

Because of this relation one says that, by definition, $\partial_j X^\theta$ is tangent to the rho representation $\rho(p^\theta)$ of the model p^θ.

Next decompose X^θ into a component Y^θ in the tangent plane

$$Y^\theta = \sum_i y^i(\theta) \partial_i X^\theta$$

plus a component $X^\theta - Y^\theta$ *orthogonal* to the tangent plane, i.e., satisfying

$$\left\langle X^\theta - Y^\theta, \partial_i X^\theta \right\rangle_\theta = 0 \quad \text{for all } i.$$

A short calculation gives

$$y^i(\theta) = \sum_j g^{ij}(\theta) \langle X^\theta, \partial_j X^\theta \rangle,$$

where $g^{ij}(\theta)$ is the matrix inverse of $g_{ij}(\theta)$. On the other hand, from

$$-\partial_i S_{\rho,\tau}(p^\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \tau(p^\theta) \partial_i \rho(p^\theta),$$

we have

$$-\partial_i S_{\rho,\tau}(p^\theta) = \langle X^\theta, \partial_i X^\theta \rangle_\theta. \tag{46}$$

Hence, the orthogonal projection of $X(\theta)$ onto the tangent plane equals

$$Y^\theta = \sum_{i,j} [-\partial_i S_{\rho,\tau}(p^\theta)] g^{ij}(\theta) \partial_j X^\theta.$$

A special case, of interest later on, occurs when $y^i(\theta) = \theta^i$ so that

$$\left\langle X^\theta - \sum_j \theta^j \partial_j X^\theta, \partial_i X^\theta \right\rangle_\theta = 0 \tag{47}$$

and

$$-\partial_i S_{\rho,\tau}(p^\theta) = \sum_j g_{ij}(\theta)\theta^j. \tag{48}$$

Written out explicitly in terms of ρ and τ, condition (47) is

$$\int_{\mathcal{X}} \mathrm{d}x \, \frac{\partial \rho(p^\theta(x))}{\partial \theta^i} \left(\tau(p^\theta(x)) - \sum_j \theta^j \frac{\partial \tau(p^\theta(x))}{\partial \theta^j} \right) = 0.$$

Its importance follows from the possibility to use the entropy $S_{\rho,\tau}$ as a potential function generating coordinates $\theta_i = \sum_j g_{ij}(\theta)\theta^j$.

We point out that the above analysis yields identical conclusions if we adopt $X^\theta(x) = \rho(p^\theta(x))$ and

$$\langle u, v \rangle_\theta = \int_{\mathcal{X}} \mathrm{d}x \, \frac{\tau'(p^\theta(x))}{\rho'(p^\theta(x))} u(x)\, v(x). \tag{49}$$

4.4 Difference between rho–tau metric and entropic metric

Starting from the rho–tau entropy $S_{\rho,\tau}$ of the parametric family p^θ

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

we take the second derivative to obtain

$$h_{ij}(\theta) = -\partial_i \partial_j S_{\rho,\tau}(p^\theta). \tag{50}$$

Likewise, define

$$h^*_{ij}(\theta) = -\partial_i \partial_j S^*_{\rho,\tau}(p^\theta)$$

using the dual entropy function $S^*_{\rho,\tau}(p^\theta)$. So h_{ij} (and its dual h^*_{ij}) is symmetric in i, j. When positive-definite, $h(\theta)$ can also serve as a metric tensor, as is found sometimes in the physics literature. We may call it the "entropic metric".

Recall that the rho–tau metric (44) is induced by the rho–tau divergence (14) by differentiating twice, see, (41). Though the entropic metric $h(\theta)$ (induced from rho–tau entropy) differs in general from the rho–tau metric $g(\theta)$ (induced from rho–tau divergence or equivalently rho–tau cross-entropy), the first-order derivatives of $C_{\rho,\tau}$, when evaluated at $p = p^\theta$, is equal to that of $S_{\rho,\tau}(p^\theta)$ or of $S^*_{\rho,\tau}(p^\theta)$

$$\partial_i S_{\rho,\tau}(p^\theta) = \partial_i C_{\rho,\tau}(p^\theta, p^{\theta'})\Big|_{\theta'=\theta} = -\int_{\mathcal{X}} \mathrm{d}x\, \tau(p^\theta)\partial_i \rho(p^\theta), \tag{51}$$

$$\partial_i S^*_{\rho,\tau}(p^\theta) = \partial'_i C_{\rho,\tau}(p^\theta, p^{\theta'})\Big|_{\theta=\theta'} = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p^\theta)\partial_i \tau(p^\theta). \tag{52}$$

They reflect, respectively, the vanishing of $\partial_i D_{\rho,\tau}(p^\theta, p^{\theta'})$ and of $\partial'_i D_{\rho,\tau}(p^\theta, p^{\theta'})$ at $p^\theta = p^{\theta'}$.

Making use of (42), one obtains, respectively

$$h_{ij}(\theta) = g_{ij}(\theta) + A_{ij}(\theta), \tag{53}$$

$$h^*_{ij}(\theta) = g_{ij}(\theta) + B_{ij}(\theta), \tag{54}$$

where $A_{ij}(\theta)$ and $B_{ij}(\theta)$ are functions symmetric in i, j, given by

$$A_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x\, \tau(p^\theta(x))\partial_i\partial_j \rho(p^\theta(x)),$$
$$B_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x\, \rho(p^\theta(x))\partial_i\partial_j \tau(p^\theta(x)).$$

When they are non-zero, they reflect the difference of the rho–tau metric $g(\theta)$ induced from cross-entropy C, from $h(\theta)$ or $h^*(\theta)$ induced from entropy S or dual entropy S^*. From (53) or (54), it can be seen that if either A_{ij} or B_{ij} can be written as the Hessian of a function, then so can g_{ij}—the rho–tau metric becomes Hessian.

4.5 Hessian geometry

We now consider the conditions under which the rho-tau metric g becomes Hessian.

Theorem 1 (Conditions for the rho–tau metric to be Hessian) *Let be given a C^∞-manifold of probability distributions p^θ. For fixed strictly increasing functions ρ and τ, let the metric tensor $g(\theta)$ be given by (42). Then the following statements are equivalent:*

1. *g is Hessian, i.e., there exists $\Phi(\theta)$ such that*

$$g_{ij}(\theta) = \partial_i\partial_j \Phi(\theta).$$

2. *There exists a function* $U(\theta)$ *such that*

$$\partial_i\partial_j U(\theta) = -\int_{\mathcal{X}} \mathrm{d}x\, \tau(p^\theta(x))\partial_i\partial_j\rho(p^\theta(x)). \tag{55}$$

3. *There exists a function* $V(\theta)$ *such that*

$$\partial_i\partial_j V(\theta) = -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p^\theta(x))\partial_i\partial_j\tau(p^\theta(x)). \tag{56}$$

Proof *(i)* $\longleftrightarrow$ *(ii)* From the identity (53), the existence of $\Phi(\theta)$ to represent g_{ij} as its second derivatives allows us to choose the function U as $U = \Phi + S$. So from (i) we obtain (ii). Conversely when the integral term can be represented by the second derivative of $U(\theta)$, we can choose $\Phi = U - S$, which satisfies (53). This yields (i) from (ii).
(i) $\longleftrightarrow$ *(iii)* The proof is similar to the previous paragraph, except that we invoke (54). □

The case when g is Hessian is very special, because of the existence of various bi-orthogonal coordinates. From

$$U = \Phi + S,$$
$$V = \Phi + S^*,$$

there are *three* "potential functions": Φ which generates g, S which generates h, and U which measures the discrepancy between g and h. Because of the $\rho \longleftrightarrow \tau$ duality there are two additional potentials S^* and V. Each of these potential functions can define conjugate coordinates with respect to θ. In particular, one defines

$$\eta_i = \partial_i\Phi, \qquad \xi_i = \partial_i U, \qquad \zeta_i = \partial_i V.$$

They are linked via

$$\begin{aligned}\xi_i(\theta) &= -\int_{\mathcal{X}} \mathrm{d}x\, \tau(p^\theta(x))\partial_i\rho(p^\theta(x)) + \eta_i = \partial_i S_{\rho,\tau}(p^\theta) + \eta_i\\ \zeta_i(\theta) &= -\int_{\mathcal{X}} \mathrm{d}x\, \rho(p^\theta(x))\partial_i\tau(p^\theta(x)) + \eta_i = \partial_i S^*_{\rho,\tau}(p^\theta) + \eta_i.\end{aligned} \tag{57}$$

We call η_i the dual coordinates of the θ^i. The meaning of ξ_i and ζ_i will be explained in Sect. 5.1.

This multitude of potentials is well-known in thermodynamics, where they are interpreted in the context of the theory of ensembles. See for instance [4].

4.6 Rho–tau connections and dually flat geometry

Under Hessian geometry, there exists a pair of dually-flat connections. In the case of conjugate rho-tau embedding of a parametric model p^θ, Zhang introduced the following connections [20]

$$\Gamma^{(\alpha)}_{ij,k} = \frac{1+\alpha}{2}\int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_i\partial_j\rho(p^\theta(x))\right]\left[\partial_k\tau(p^\theta(x))\right] + \frac{1-\alpha}{2}\int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_i\partial_j\tau(p^\theta(x))\right]\left[\partial_k\rho(p^\theta(x))\right], \tag{58}$$

where $\Gamma^{(\alpha)}_{ij,k} \equiv (\Gamma^{(\alpha)})^l_{ij} g_{lk}$. One readily verifies

$$\Gamma^{(\alpha)}_{ij,k} + \Gamma^{(-\alpha)}_{jk,i} = \partial_i g_{jk}(\theta). \tag{59}$$

This shows that, by definition, $\Gamma^{(-\alpha)}$ is the dual connection of $\Gamma^{(\alpha)}$. In particular, $\Gamma^{(0)}$ is self-dual and therefore coincides with the Levi-Civita connection. The family of α-connections (58) is induced by the divergence function $D^{(\alpha)}_{f,\rho}(p,q)$ given by (35), with corresponding α-values. Furthermore, upon switching $\rho \longleftrightarrow \tau$ in the divergence function, the designation of 1-connection versus (-1)-connection also switches.

The coefficients of the connection $\Gamma^{(-1)}$ vanish identically if

$$\int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_i\partial_j\tau(p^\theta(x))\right]\left[\partial_k\rho(p^\theta(x))\right] = 0. \tag{60}$$

This condition can be written as

$$\langle \partial_i\partial_j X^\theta, \partial_k X^\theta\rangle_\theta = 0. \tag{61}$$

It expresses that the second derivatives $\partial_i\partial_j X^\theta$ are orthogonal to the tangent plane of the statistical manifold. If satisfied, then the dual of $\Gamma^{(-1)}$ satisfies

$$\Gamma^{(1)}_{ij,k} = \partial_i g_{jk}(\theta). \tag{62}$$

Likewise, the coefficients of the connection $\Gamma^{(1)}$ vanish identically if

$$\int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_i\partial_j\rho(p^\theta(x))\right]\left[\partial_k\tau(p^\theta(x))\right] = 0. \tag{63}$$

In the case of a ϕ-deformed exponential family (see the next section) condition (60) is satisfied in the $\rho = \mathrm{id}$ gauge while (63) is satisfied in the $\tau = \mathrm{id}$ gauge.

Proposition 2 *With respect to conditions* (60) *and (63),*

1. *When (60) holds, the coordinates θ^i are affine coordinates for $\Gamma^{(-1)}$; the dual coordinates η_i are affine coordinates for $\Gamma^{(1)}$;*
2. *When (63) holds, the coordinates θ^i are affine coordinates for $\Gamma^{(1)}$; the dual coordinates η_i are affine coordinates for $\Gamma^{(-1)}$;*
3. *In either case above, $g(\theta)$ is Hessian.*

Proof Recall that when $\Gamma = 0$ under a coordinate system θ, then θ^i's are affine coordinates—the geodesics are straight lines:

$$\theta(t) = (1-t)\,\theta(0) + t\,\theta(1).$$

The geodesics of the dual connection Γ^* satisfy the Euler-Lagrange equations

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta^i + \Gamma^{*i}_{km}\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right) = 0. \tag{64}$$

Its solution is such that the dual coordinates η are affine coordinates:

$$\eta(t) = (1-t)\,\eta(0) + t\,\eta(1).$$

For Statement 1, apply the above knowledge, taking $\Gamma = \Gamma^{(-1)}$ and $\Gamma^* = \Gamma^{(1)}$. For Statement 2, take $\Gamma = \Gamma^{(1)}$ and $\Gamma^* = \Gamma^{(-1)}$.

To prove Statement 3 observe that

$$\partial_k g_{ij}(\theta) = \int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_i \tau(p^\theta(x))\right] \partial_j \partial_k \rho(p^\theta(x)) + \int_{\mathcal{X}} \mathrm{d}x\ \left[\partial_j \rho(p^\theta(x))\right] \partial_i \partial_k \tau(p^\theta(x)).$$

So the vanishing of either term, i.e., either (60) or (63) holding, will yield $\partial_k g_{ij}(\theta)$ to be symmetric in j, k or in i, k, respectively. This, in conjunction with the fact that g_{ij} is symmetric in i, j, leads to the conclusion that $\partial_k g_{ij}(\theta)$ is totally symmetric in an exchange of any two of the three indices i, j, k. This implies that η_i exist for which $g_{ij}(\theta) = \partial_j \eta_i = \partial_i \eta_j$. The symmetry of g implies now that it equals the Hessian of a potential Φ. □

5 Deformed exponential models

In the previous Section, we show how the rho–tau geometry fully generalizes the α-geometry of Amari's. The approach considered was largely based on generalization of entropy, cross-entropy, divergence functions, and the geometry induced by those quantities. Here we consider the generalization of exponential family to deformed exponential (phi-exponential) family, and show how they give rise to Hessian geometry. In this way, the α-geometry is fully generalized with conjugate rho-tau embedding in terms of (i) entropy, cross-entropy, divergence function; (ii) Riemannian metric and affine connections; and (iii) parametric probability families.

5.1 Phi-exponential model

Fix an arbitrary monotone function ϕ along with real random variables $F_1, F_2, \cdots, F_n$. These functions determine a model $\theta \to p^\theta$ belonging to the phi-exponential family by the relation (see, [11–13])

$$p^\theta(x) = \exp_\phi \left[\sum_k \theta^k F_k(x) - \alpha(\theta) \right], \tag{65}$$

provided that one can prove that the normalization function $\alpha(\theta)$ exists. Normalization of p^θ leads to

$$\partial_i \alpha(\theta) = \tilde{\mathbb{E}}_\theta F_i,$$

where the so-called *escort expectation*, denoted $\tilde{\mathbb{E}}_\theta$,

$$\tilde{\mathbb{E}}_\theta\{\cdot\} = \int_{\mathcal{X}} \mathrm{d}x \; \tilde{p}^\theta\{\cdot\}$$

is with respect to the the *escort family* of probability distributions $\tilde{p}^\theta$ as defined by

$$\tilde{p}^\theta(x) = \frac{1}{z(\theta)} \phi(p^\theta(x)).$$

Here the integral

$$z(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \phi(p^\theta(x)) \tag{66}$$

is assumed to converge. Using the properties of the deformed exponential function one obtains

$$\partial_i p^\theta(x) = z(\theta)\tilde{p}^\theta(x)\left[F_i - \partial_i\alpha(\theta)\right] = \phi(p^\theta(x))\left[F_i(x) - \partial_i\alpha(\theta)\right]. \tag{67}$$

For later convenience, we also derive the first derivative of the $z(\theta)$ function

$$\begin{aligned}\partial_j z(\theta) &= \int_{\mathcal{X}} \mathrm{d}x \, \phi'(p^\theta(x))\partial_j p^\theta(x) \\ &= \int_{\mathcal{X}} \mathrm{d}x \, \phi'(p^\theta(x))\phi(p^\theta(x)) \left[F_j(x) - \partial_j\alpha(\theta)\right]\end{aligned}$$

and the second derivatives of the α function

$$
\begin{aligned}
\partial_i \partial_j \alpha(\theta) &= \partial_j \left(\frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) F_i(x) \right) \\
&= \frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \partial_j \phi(p^\theta(x)) F_i(x) - [\partial_i \alpha(\theta)] \frac{1}{z(\theta)} \partial_j z(\theta) \\
&= \frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \phi'(p^\theta(x)) \phi(p^\theta(x)) (F_j(x) - \partial_j \alpha(\theta)) F_i(x) \\
&\quad - [\partial_i \alpha(\theta)] \frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \phi'(p^\theta(x)) \phi(p^\theta(x)) \left[F_j(x) - \partial_j \alpha(\theta) \right] \\
&= \frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \phi'(p^\theta(x)) \phi(p^\theta(x)) [F_j(x) - \partial_j \alpha(\theta)] [F_i(x) - \partial_i \alpha(\theta)]. \quad (68)
\end{aligned}
$$

Proposition 3 *Denote*

$$\eta_i = \mathbb{E}_\theta F_i = \int_{\mathcal{X}} \mathrm{d}x\, p^\theta(x) F_i(x).$$

Then, there exists a function Φ such that

$$\eta_i = \partial_i \Phi(\theta).$$

Proof We compute

$$
\begin{aligned}
\partial_j \eta_i &= \int_{\mathcal{X}} \mathrm{d}x\, \partial_j p^\theta F_i(x) \\
&= \int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) \left[F_j(x) - \partial_j \alpha(\theta) \right] F_i(x) \\
&= \int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) [F_i(x) - \partial_i \alpha(\theta)] \left[F_j(x) - \partial_j \alpha(\theta) \right] \quad (69)
\end{aligned}
$$

which is symmetric in i, j. Therefore, there exists a function Φ such that $\eta_i = \partial_i \Phi$, and that

$$\partial_i \partial_j \Phi(\theta) = \int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) [F_i(x) - \partial_i \alpha(\theta)] \left[F_j(x) - \partial_j \alpha(\theta) \right]. \quad (70)$$

□

We remark that with respect to any deformed-exponential model $p^\theta(x)$, we have two sets of coordinates *dual* with respect to θ:

1. $\eta_i = \mathbb{E}_\theta F_i$, which is given by $\eta_i = \partial_i \Phi$ for some function $\Phi(\theta)$;
2. $\zeta_i = \tilde{\mathbb{E}}_\theta F_i$, which is given by $\zeta_i = \partial_i \alpha$ for the α function associated with the deformed-exponential $\log_\phi$.

In the literature, η is called the *expectation coordinates* while ζ is called the *escort (expectation) coordinates*.

Simple calculations show that the first and second derivatives of the η coordinates with respect to θ can be expressed as the second and third derivative of Φ:

$$\partial_i \eta_j = \int_{\mathcal{X}} \mathrm{d}x\, \partial_i p^\theta(x) F_j(x) = \partial_i \partial_j \Phi;$$
$$\partial_i \partial_j \eta_k = \int_{\mathcal{X}} \mathrm{d}x\, \partial_i \partial_j p^\theta(x) F_k(x) = \partial_i \partial_j \partial_k \Phi.$$

Now, let us consider the rho–tau metric (42) or (44) applied to the ϕ-exponential model (65),

$$\begin{aligned} g_{ij}(\theta) &= \int_{\mathcal{X}} \mathrm{d}x\, \frac{1}{\psi(p^\theta(x))} \left[\partial_i p^\theta(x)\right] \left[\partial_j p^\theta(x)\right] \\ &= \int_{\mathcal{X}} \mathrm{d}x\, \frac{(\phi(p^\theta(x)))^2}{\psi(p^\theta(x))} \left[F_i(x) - \partial_i \alpha(\theta)\right] \left[F_j(x) - \partial_j \alpha(\theta)\right]. \end{aligned} \tag{71}$$

Below, we will consider two subcases, with $\psi = \phi$ or $\psi = \phi/\phi'$, both resulting in interesting geometries for the ϕ-exponential family. Since ψ is controlled by two embedding functions ρ and τ, for simplicity we choose $\tau = \log_\phi$, which leaves only the ρ function to be specified.

5.2 The case of $\psi = \phi$

Upon choosing $\psi = \phi$, the expression of the rho-tau metric g in (71) takes the form of the right-hand side of (70). Therefore,

Theorem 2 *With the choice of the weighting funcion $\psi = \phi$, the rho-tau metric tensor* (71) *of the phi-exponential family obeying* (65) *is Hessian.*

In this case, the rho–tau metric coincides with the Hessian metric g^Φ defined as second derivative of the potential Φ given by (70)

$$g^\Phi_{ij}(\theta) = \partial_i \partial_j \Phi.$$

In the meanwhile, the rho–tau metric tensor (71) also takes the form

$$g_{ij}(\theta) = z(\theta) \left[\tilde{\mathbb{E}}_\theta F_i F_j - \tilde{\mathbb{E}}_\theta F_i \tilde{\mathbb{E}}_\theta F_j\right], \tag{72}$$

as originally derived in [11]. This expression implies that the rho–tau metric tensor in this case is conformally equivalent to the metric tensor $\tilde{g}$ derived from the escort expectation of the random variables F_i:

$$\tilde{g}_{ij}(\theta) = \tilde{\mathbb{E}}_\theta F_i F_j - \tilde{\mathbb{E}}_\theta F_i \tilde{\mathbb{E}}_\theta F_j = \tilde{\mathbb{E}}_\theta \left[(F_i - \tilde{\mathbb{E}}_\theta F_i)(F_j - \tilde{\mathbb{E}}_\theta F_j)\right]. \tag{73}$$

For later convenience, we refer to $\tilde{g}$ as given by (73) as the “escort metric”.

Note that when $\psi = \phi$ and $\tau = \log_\phi$, then $\rho = \mathrm{id}$. That is, we have adopted the rho-id gauge. that is, Type I gauge. In such case, $S_{\rho,\tau}$ reduces to the phi-entropy S_ϕ.

Proposition 4 *Under the rho-id gauge, the Hessian potential Φ of the ϕ-exponential model*

1. *is given by*

$$\Phi(\theta) = S_\phi(p^\theta) + \sum_k \theta^k \mathbb{E}_\theta F_k; \tag{74}$$

2. *equals the convex dual $S_\phi^{\mathrm{cd}}(\theta)$ of the ϕ-entropy S_ϕ;*

where $S_\phi(p^\theta) = -\int_{\mathcal{X}} f(p^\theta)\mathrm{d}x,\ f' = \log_\phi$.

Proof To prove Statement 1, note that from the definitions (12) and (65) follows

$$\begin{aligned}
-\partial_i S_\phi(p^\theta) &= \int_{\mathcal{X}} \mathrm{d}x\ \log_\phi(p^\theta(x))\,\partial_i(p^\theta(x)) \\
&= \int_{\mathcal{X}} \mathrm{d}x \left[\sum_k \theta^k F_k(x) - \alpha(\theta)\right] \partial_i(p^\theta(x)) \\
&= \sum_k \theta^k \partial_i \mathbb{E}_\theta F_k.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\partial_i \Phi(\theta) &= \partial_i S_\phi(p^\theta) + \mathbb{E}_\theta F_i + \sum_k \theta^k \partial_i \mathbb{E}_\theta F_k \\
&= \mathbb{E}_\theta F_i = \eta_i.
\end{aligned}$$

The convex function Φ defined by (74) is hence the potential function generating the Hessian metric g_{ij}.

To prove Statement 2, that is, the potential Φ can be seen as the convex dual S^{cd} of the phi-entropy S_ϕ, recall the definition of convex duality

$$S_\phi^{\mathrm{cd}}(\theta) = \sup_p \{S_\phi(p) + \sum_k \theta^k \mathbb{E}_p F_k\}.$$

From (15) follows that for any probability distribution p is

$$D_\phi(p, p^\theta) = S_\phi(p^\theta) - S_\phi(p) - \sum_k \theta^k \mathbb{E}_p F_k + \sum_k \theta^k \mathbb{E}_\theta F_k, \tag{75}$$

with equality if and only if $p = p^\theta$. This implies

$$S_\phi(p) + \sum_k \theta^k \mathbb{E}_p F_k \le S_\phi(p^\theta) + \sum_k \theta^k \mathbb{E}_\theta F_k,$$

with equality if and only if $p = p^\theta$. One concludes that

$$S^{\mathrm{cd}}_\phi(\theta) = S_\phi(p^\theta) + \sum_k \theta^k \mathbb{E}_\theta F_k.$$

From Statement 1 then follows $\Phi = S^{\mathrm{cd}}$. □

It is important to note that the duality between S and S^* is not convex duality, $S^{\mathrm{cd}} \neq S^*$, but rather a duality arising from $\rho \longleftrightarrow \tau$.

Under the rho-id gauge, we have $S_{\rho,\tau} = S_\phi$ so that

$$\begin{aligned}\partial_i S_\phi(p^\theta) &= -\sum_k \theta^k \partial_i \partial_k \Phi,\\ \partial_i \partial_j S_\phi(p^\theta) &= -\partial_i \partial_j \Phi - \sum_k \theta^k \partial_i \partial_j \partial_k \Phi.\end{aligned}$$

In this case (i.e., rho-id gauge for phi-exponential family)

$$S^*_\phi(p^\theta) = -f^* \left(\sum_k \theta^k F_k(x) - \alpha(\theta) \right),$$

so that

$$\begin{aligned}\partial_i S^*_\phi(p^\theta) &= -\partial_i \alpha + \mathbb{E}_\theta F_i = \eta_i - \zeta_i,\\ \partial_i \partial_j S^*(p_\theta) &= -\partial_i \partial_j \alpha + \partial_i \partial_j \Phi.\end{aligned}$$

That selecting the rho-id gauge causes the rho–tau metric of the ϕ-exponential family to become a Hessian metric can also be seen via (see Theorem 1)

$$\begin{aligned}\partial_i \partial_j V(\theta) &= -\int_{\mathcal{X}} \mathrm{d}x\, p^\theta(x) \partial_i \partial_j \log_\phi(p^\theta(x)) = -\int_{\mathcal{X}} \mathrm{d}x\, p^\theta(x)(-\partial_i \partial_j \alpha(\theta))\\ &= \partial_i \partial_j \alpha(\theta).\end{aligned}$$

So we can take $V(\theta) = \alpha(\theta)$. The convex potential Φ function can have an equivalent expression

$$\Phi(\theta) = \alpha(\theta) + \int_{\mathcal{X}} \mathrm{d}x\, f^* \left(\sum_k \theta^k F_k(x) - \alpha(\theta) \right),$$

where $(f^*)' = \exp_\phi$. In this case,

$$\begin{aligned}
U(\theta) &= \alpha(\theta) + S_\phi(p^\theta) - S_\phi^*(p^\theta) \\
&= \alpha(\theta) - \int_{\mathcal{X}} \mathrm{d}x\, f\left(\exp_\phi\left(\sum_k \theta^k F_k(x) - \alpha(\theta)\right)\right) \\
&\quad + \int_{\mathcal{X}} \mathrm{d}x\, f^*\left(\sum_k \theta^k F_k(x) - \alpha(\theta)\right),
\end{aligned}$$

with

$$\partial_i U = \eta_i - \sum_k \theta^k \partial_k \eta_i .$$

The cross-entropies for deformed-exponential model (under rho-id gauge) are:

$$\begin{aligned}
C_{\rho,\tau}(p^\theta, q^\theta) &= -\sum_k (\theta_p)^k \left(\int \mathrm{d}x\, p^\theta(x) F_i(x)\right) + \alpha(\theta_q), \\
\overline{C}_{\rho,\tau}(p^\theta, q^\theta) &= C_{\rho,\tau}(p^\theta, q^\theta) - S^*(q^\theta) = -\sum_k (\theta_p)^k \left(\int \mathrm{d}x\, p^\theta(x) F_i(x)\right) + \Phi(\theta_q).
\end{aligned}$$

Finally, the Pythagorean Theorem 3.8 of [3] can be easily generalized to the ϕ-exponential models. Let $t \in \mathbb{R} \mapsto p_t$ be a differentiable map, defined on a neighborhood of $t = 0$, taking values in the manifold of the p^θ. A random variable P is said to be tangent to p_t at $t = 0$ in the rho-embedding if

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int \mathrm{d}x\, \rho(p_t(x)) u(x) = \langle P, u \rangle_\theta$$

for any random variable u, with the inner product $\langle \cdot, \cdot \rangle_\theta$ defined by (45) with $p^\theta = p_{t=0}$.

Theorem 3 *Let p^θ obey (65). Let $t \in \mathbb{R} \mapsto p_t$ and $s \in \mathbb{R} \mapsto r_s$ be two differentiable maps with values in the manifold of the p^θ. Let P and R be the corresponding tangent vectors at $s = t = 0$ and assume they are orthogonal in the sense that $\langle P, R \rangle_\theta = 0$. Assume $t \in \mathbb{R} \mapsto p_t$ is a geodesic for $\Gamma^{(-1)}$ and $s \in \mathbb{R} \mapsto r_s$ is a geodesic for $\Gamma^{(1)}$. Assume the two geodesics intersect at $s = t = 0$ in a common point $p_0 = r_0 \equiv q$. If $\psi = \phi$ then the following Pythagorean relation holds*

$$D_{\rho,\tau}(p_t, q) + D_{\rho,\tau}(q, r_s) = D_{\rho,\tau}(p_t, r_s).$$

Proof Let the θ-coordinates of p_t be denoted $\theta_t = (1-t)\theta_0+t\theta_1$ and the η-coordinates of r_s be denoted $\eta_s = (1-s)\eta_0 + s\eta_1$. A short calculation gives

$$\begin{aligned} P &= \sum_k [\theta_1 - \theta_0]^k \partial_k X^\theta, \\ R &= \sum_{i,j} [\eta_1 - \eta_0]_i g^{ij}(\theta) \partial_j X^\theta. \end{aligned} \tag{76}$$

Orthogonality of P and R yields

$$\begin{aligned} 0 &= \langle P, R\rangle_\theta \\ &= \sum_{i,j,k} [\theta_1 - \theta_0]^k [\eta_1 - \eta_0]_i g^{ij}(\theta) \langle \partial_k X^\theta, \partial_j X^\theta\rangle_\theta \\ &= \sum_{i,j,k} [\theta_1 - \theta_0]^k [\eta_1 - \eta_0]_i g^{ij}(\theta) g_{kj}(\theta) \\ &= \sum_k [\theta_1 - \theta_0]^k [\eta_1 - \eta_0]_k. \end{aligned} \tag{77}$$

This is used in the following calculation. From (75) follows

$$\begin{aligned} & D_{\rho,\tau}(p^t, q) + D_{\rho,\tau}(q, r^s) - D_{\rho,\tau}(p^t, r^s) \\ &\quad = \sum_k [\theta_t - \theta_0]^k \ [\eta_t - \eta_0]_k \\ &\quad = (1-t)(1-s) \sum_k [\theta_1 - \theta_0]^k \ [\eta_1 - \eta_0]_k \,. \end{aligned} \tag{78}$$

As shown above the summation term of the r.h.s. of this expression vanishes. Hence the desired result follows. □

5.3 The case of $\psi = \phi/\phi'$

The phi-deformed exponential family, considered in the previous section, is the generalization of the q-exponential model historically introduced by Tsallis [17]. The second [5] and third [19] version of the Tsallis formalism can be characterized by the observation that the role of expectations $\mathbb{E}_\theta$ and escort $\tilde{\mathbb{E}}_\theta$ is exchanged.

For convenience the model discussed below is called the Tsallis model. Consider a phi-deformed exponential family p^θ defined by (65). Assume now that the α function is strictly convex. Then its Hessian can be used to define a metric $g^\alpha(\theta)$

$$g^\alpha_{ij}(\theta) = \partial_i \partial_j \alpha(\theta).$$

For convenience, let us call this the *Tsallis metric*. Below, we show that this metric is conformally equivalent to the rho-tau metric (44), the latter being non-Hessian upon choosing $\psi = \phi/\phi'$.

Theorem 4 *For a phi-deformed exponential family p^θ of the form (65), when the weighting function ψ of the rho–tau metric in the form of (71) satisfies*

$$\frac{1}{\psi} = (\log \phi)' = \frac{\phi'}{\phi}, \tag{79}$$

then the rho–tau metric g, while itself non-Hessian, is conformally equivalent to a Hessian metric g^α (Tsallis metric).

Proof From the expression of $\partial_i \partial_j \alpha$ given by (68), we see that if we set

$$\phi'(u) = \phi(u)/\psi(u)$$

in the rho–tau metric as given by (71), then we have

$$g_{ij}(\theta) = z(\theta)\, g^\alpha_{ij}(\theta).$$

This says that the rho–tau metric g in this case is conformally related to the Tsallis metric g^α, which is the Hessian of $\alpha(\theta)$ function. □

Note that with this choice of weighting function $\psi = \phi/\phi'$ and embedding function $\tau = \log_\phi$, then $\rho = f^{-1}$ so $f(\rho(p)) = p$. This corresponds to the constant-S gauge. Therefore in the Tsallis model, we can either choose $\rho = \phi$, $\tau = \log_\phi$ (which is what we assumed in the above calculations) or dually $\rho = \log_\phi$, $\tau = \phi$. Under these Type II gauges, the rho-tau metric g_{ij} is non-Hessian while the Tsallis metric $g^\alpha_{ij}(\theta)$, if it exists, is (always) Hessian by construction.

It is known that the Tsallis metric g^α_{ij} induces a dually flat structure associated with escort expectations $\tilde{\mathbb{E}}_\theta$, see, [2,8]. The dual coodinates with respect to the α function are $\zeta_i = \tilde{\mathbb{E}}_\theta F_i$. The dual potential of $\alpha(\theta)$ is $\tilde{\mathbb{E}}_\theta \log_\phi(p^\theta)$, as shown below:

$$\begin{aligned}
\sum_k \theta^k \partial_k \alpha(\theta) - \alpha(\theta) &= \sum_k \theta^k \frac{1}{z(\theta)} \left(\int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) F_k(x) \right) - \alpha(\theta) \\
&= \frac{1}{z(\theta)} \left(\int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) \sum_k \theta^k F_k(x) \right) - \alpha(\theta) \\
&= \frac{1}{z(\theta)} \left(\int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) [\log_\phi(p^\theta(x)) + \alpha(\theta)] \right) - \alpha(\theta) \\
&= \frac{1}{z(\theta)} \int_{\mathcal{X}} \mathrm{d}x\, \phi(p^\theta(x)) \log_\phi(p^\theta(x)) \\
&= \tilde{\mathbb{E}}_\theta \log_\phi(p^\theta).
\end{aligned}$$

This expression is nothing but $S^*(p^\theta)$, apart from the conformal factor of $z(\theta)$. This shows the duality (modulo a conformal factor) between α and S^* under Type II gauge, just as Φ and S_ϕ are convex dual under Type I gauge.

5.4 Maximum entropy models

The derivation of the phi-exponential family by means of the maximum entropy method is found in [11]. The treatment here is further generalized so as to cover the approach of [5] as well.

Consider the problem of maximizing the rho–tau entropy $S_{\rho,\tau}(p)$ under the constraint that for a finite number of random variables $F_1, \cdots, F_n$ the functions

$$\epsilon_k(p) = \int_{\mathcal{X}} \mathrm{d}x\, \sigma(p(x)) F_k(x), \qquad k = 1, 2, \cdots, n,$$

have some given values. Here, σ is a given strictly increasing twice differentiable function. We are interested in two specific cases. In the rho-id gauge ($\rho = \mathrm{id}$) the given values are the expectation values of the variables F_k. Then $\sigma = \mathrm{id}$. In the case of the Tsallis model the given values are the unnormalized escort expectations of the variables F_k. This requires $\sigma = \phi$.

Introduce now Lagrange multipliers θ^k. Because of the requirement that the maximizing probability distribution is normalized, an extra multiplier α is needed. The function of Lagrange can then be chosen equal to

$$\mathcal{L}(p) = S_{\rho,\tau}(p) + \theta^k \epsilon_k(p) - \alpha \int_{\mathcal{X}} \mathrm{d}x\, p(x).$$

Stationarity implies that the optimizing probability distribution $p = p^\theta$ must satisfy an expression of the form

$$\tau(p^\theta(x))\rho'(p^\theta(x)) = \theta^k F_k(x)\sigma'(p^\theta(x)) - \alpha(\theta). \tag{80}$$

Two cases exist in which the resulting model belongs to a deformed exponential family. First take $\sigma = \mathrm{id}$. Then (80) becomes

$$\tau(p^\theta(x))\rho'(p^\theta(x)) = \theta^k F_k(x) - \alpha(\theta).$$

This can be written as (65) with ϕ such that $\tau\rho' = \log_\phi$. In the rho-id gauge this condition is satisfied with $\rho = \mathrm{id}$ and $\tau = \log_\phi$.

The other case occurs when $\tau\rho'$ is proportional to σ', say $\tau\rho' = \sigma'$. Then (80) becomes

$$\sigma'(p^\theta(x)) = \frac{\alpha(\theta)}{\theta^k F_k(x) - 1}.$$

This is of the form (65) provided ϕ is such that

$$\log_\phi(u) = B\left(\frac{1}{\sigma'(u)} - \frac{1}{\sigma'(1)}\right) \tag{81}$$

for some constant B. The result is

$$p^{\theta}(x) = \exp_{\phi}\left[\sum_k \tilde{\theta}^k F_k(x) - \tilde{\alpha}(\theta)\right]$$

with

$$\tilde{\theta}^k = B\frac{\theta^k}{\alpha(\theta)} \quad \text{and} \quad \tilde{\alpha}(\theta) = B\left(\frac{1}{\sigma'(1)} + \frac{1}{\alpha(\theta)}\right).$$

In the Tsallis context $\sigma(u) = \phi(u) = u^q$ for some $q \neq 1$. The condition (81) is then satisfied with $B = q/(1-q)$. The choice $\sigma = \phi$ works because if σ is a power law then $\sigma\sigma''/(\sigma')^2$ is a constant.

6 Summary and discussions

The classic information geometry (Amari's α-geometry) contains three inter-related parts: (i) the Fisher–Rao metric g with the family of α-connections; (ii) the divergence functions inducing g and α-connections; (iii) the exponential family corresponding to the dually flat $\alpha = 1$ connection. Over the years, various aspects of classic information geometry were generalized by relaxing from the logarithm/exponential embedding functions, predominantly in the deformed exponential approach of Naudts [11], the U model of Eguchi [6], and conjugate rho-tau embedding of Zhang [20]. In this paper, these approaches are all synthesized to give a full generalization of classical information geometry with arbitrary monotone embedding.

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. This is through adopting the rho-id gauge (or dually, tau-id gauge). 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.

On the other hand, fixing the gauge $f^* = \tau^{-1}$ (constant S^* gauge) renders the rho-tau cross-entropy to be the U cross-entropy, where the dual-entropy is constant. In this case, $\tau \longleftrightarrow \rho$ is akin to the $\log_\phi \longleftrightarrow \phi$ transformation encountered in studying the phi-exponential family.

With respect to the induced geometry, we first show that the rho-tau metric tensor $g(\theta)$ depends on a single function ψ which is defined by $\psi(u) = 1/(\rho'(u)\tau'(u))$. Theorem 1 gives equivalent conditions for the rho-tau metric to be Hessian. If the probability model is phi-exponential with the same function $\phi = \psi$, then the rho–tau metric is Hessian. The potential function is the convex conjugate of the rho–tau entropy. However, in general the rho–tau metric is not Hessian. A non-Hessian special case is to choose $\psi = \phi/\phi'$ for the phi-exponential family; the resulting metric is conformally equivalent to the metric given by the second-derivative of the normalizing function $\alpha(\theta)$.

In our generalization of Amari's α-geometry, there is a variety of (semi-) Riemannian metrics:

(i) rho–tau metric, induced from (ρ, τ)-divergence or (ρ, τ) cross-entropy; it contains one free function ψ given by $\psi \rho' \tau' = 1$;
(ii) entropic metric, induced from the (ρ, τ)-entropy—it is a Hessian metric.

When the probability p^θ is the ϕ-exponential family, with $\alpha(\theta)$ representing the normalization function, then it is shown that there always exists another potential function Φ, which is usually different from α (unless $\phi = \text{id}$, the case of vanilla exponential family). Assuming convexity, both α and Φ can be used to induce dual or expectation coordinates, respectively ζ and η, with respect to the θ parameter (natural coordinate) indexing the ϕ-exponential family. The rho–tau metric g of the ϕ-exponential family, being dependent on the weighting function ψ, may or may not be Hessian. After fixing one embedding function τ to be $\log_\phi$, it turns out that

(i) $g = g^\Phi$ upon choosing $\psi = \phi$ (which forces $\rho = \text{id}$, and hence adopting the Type I gauge). That is, the rho–tau metric g coincides with the Hessian metric g^Φ as induced from Φ; it is conformally equivalent to the (non-Hessian) escort metric (associated with the escort expectation);
(ii) $g = z(\theta) g^\alpha$ upon choosing $\psi = \phi/\phi'$ (which forces $\rho = \log_\tau$, and hence adopting Type II or constant-S gauge). That is, the rho–tau metric g, though not Hessian, becomes conformally equivalent to the Tsallis metric g^α, a Hessian metric induced from the normalization function α.

Therefore, one should carefully distinguish the various metrics: rho-tau metric (which may become Hessian) and entropic metric, which is always Hessian, and in the case of phi-exponential family, Tsallis metric (which is always Hessian), and the escort metric (which is generally non-Hessian).

Note that conformal equivalence for the case of $\psi = \phi$ were previously studied e.g., [2,8]. For the case of $\psi = \phi/\phi'$, we were brought to the awareness (by an anonymous reviewer) that a recent report [9] derived identical conclusions using a different approach—there the weighting function is viewed as arising as the second derivative of $\exp_\phi$:

$$(\exp_\phi)' = \phi \circ \exp_\phi \ ,$$
$$(\exp_\phi)'' = (\phi' \circ \exp_\phi) \cdot (\phi \circ \exp_\phi) = (\phi' \cdot \phi) \circ \exp_\phi \ .$$

Interestingly, the first derivative of $\exp_\phi$ corresponds to the $\psi = \phi$ selection. Future research will elucidate whether the result obtained by this "sequential derivative" approach of [9] and by our current rho–tau embedding approach to specify the weighting function of the Riemannian metric is merely a coincidence or reflects a deep cause.

Our current analysis 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:

1. In general, the divergence function (as a two-variable function) is the difference of cross-entropy (as a two-variable function) and *a pair of* dual entropies (as one-

variable functions); one can always define a “modified” cross-entropy to absorb one of the entropies (as in the U cross-entropy case);

2. In general, the deformed-exponential family always has *two* potentials, Φ and α, which are not equal unless there is no deformation. Therefore, there are always two expectation coordinates (standard expectation and escort expectation) with respect to the same natural parameter of the deformed-exponential family. This is regardless of the rho-tau metric of the Riemannian manifold (which is induced from the divergence function);
3. When the rho–tau metric is Hessian (i.e., under Type I gauge), there are actually multiple potentials, including Φ and phi-entropy, as well as α and dual entropy;
4. When the rho–tau metric is conformally equivalent to a Hessian metric (i.e., under Type II gauge), the α and S^* form convex dual (after a conformal scaling factor);
5. The U model and the phi-model are identical models under different notations.

The conjugate rho–tau embedding mechanism and phi-exponential model together provide the necessary ingredients for generalizing the α-geometry while preserving its elegant geometric structure. This greatly expands the reach of information geometric analysis to a much larger applied setting. In particular, the principle of Maximum Entropy inference can be generalized to the case of a generalized linear model. Future research will show how this generalized formulation of maxent duality and calculations may lead to practical impact in statistics, information science, and machine learning.

Acknowledgements The project is supported in part by DARPA/ARO Grant W911NF-16-1-0383 (“Information Geometry: Geometrization of Science of Information”, PI: Zhang). Both authors contributed equally to this paper.

References

1. Amari, S.: Differential-geometric methods in statistics. Lecture notes in statistics, vol. 28. Springer, Berlin (1985)
2. Amari, S., Ohara, A., Matsuzoe, H.: Geometry of deformed exponential families: invariant, dually-flat and conformal geometries. Physica A Stat. Mech. Appl. **391**(18), 4308–4319 (2012)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry. Translations of mathematical monographs 191 (Am. Math. Soc., 2000; Oxford University Press, 2000); Originally in Japanese (Iwanami Shoten, Tokyo, 1993)
4. Callen, H.B.: Thermodynamics and an Introduction to Thermostatistics, 2nd edn. Wiley, New York (1985)
5. Curado, E.M.F., Tsallis, C.: Generalized statistical mechanics: connection with thermodynamics. J. Phys. A24, L69 (1991); Corrigenda 24, 3187 (1991) and 25, 1019 (1992)
6. Eguchi, S.: Information geometry and statistical pattern recognition. Sugaku Expositions (Amer. Math. Soc.) **19**, 197–216 (2006). (originally Sūgaku 56 (2004) 380 in Japanese)
7. Kaniadakis, G.: Non-linear kinetics underlying generalized statistics. Physica A Stat. Mech. Appl. **296**, 405–425 (2001)
8. Matsuzoe, H.: Hessian structures on deformed exponential families and their conformal structures. Diff. Geom. Appl. **35**, 323–333 (2014)
9. Matsuzoe, H., Scarfone, A.M., Wada, T.: A sequential structure of statistical manifolds on deformed exponential family. LNCS **10589**, 223–230 (2017)
10. Montrucchio, L., Pistone, G.: Deformed exponential bundle: the linear growth case. In: Nielsen, F., Barbaresco, F. (eds.) Geometric science of information, GSI 2017 LNCS proceedings, pp. 239–246. Springer, Berlin (2017)

11. Naudts, J.: Estimators, escort probabilities, and phi-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
12. Naudts, J.: Generalised exponential families and associated entropy functions. Entropy **10**, 131–149 (2008)
13. Naudts, J.: Generalised Thermostatistics. Springer, Berlin (2011)
14. Naudts, J., Anthonis, B.: The exponential family in abstract information theory. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2013 LNCS Proceedings, pp. 265–272. Springer, Berlin (2013)
15. Naudts, J., Zhang, J.: Information geometry under monotone embedding. Part II: geometry. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2017 Proceedings. LNCS, pp. 215–222. Springer, Berlin (2017)
16. Newton, N.J.: Information geometric nonlinear filtering. Inf. Dim. Anal., Quantum Prob. Rel. Topics **18**, 1550014 (2015)
17. Tsallis, C.: Possible generalization of Boltzmann–Gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
18. Tsallis, C.: What are the numbers that experiments provide? Quim. Nova **17**, 468 (1994)
19. Tsallis, C., Mendes, R.S., Plastino, A.R.: The role of constraints within generalized nonextensive statistics. Physica A **261**, 543–554 (1998)
20. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
21. Zhang, J.: Referential duality and representational duality on statistical manifolds. Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, pp. 58-67 (2005)
22. Zhang, J.: Nonparametric information geometry: from divergence function to referential-representational biduality on statistical manifolds. Entropy **15**, 5384–5418 (2013)
23. Zhang, J.: On monotone embedding in information geometry. Entropy **17**, 4485–4499 (2015)
24. Zhang, J., Naudts, J.: Information geometry under monotone embedding. Part I: divergence functions. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2017 Proceedings LNCS, pp. 205–214. Springer, Berlin (2017)
25. Zhou, J.: Information theory and statistical mechanics revisited. arXiv:1604.08739