Rho-Tau Embedding of Statistical Models

Jan Naudts and Jun Zhang

Abstract Two strictly increasing functions ρ and τ determine the rho-tau embedding of a statistical model. The Riemannian metric tensor is derived from the rho-tau divergence. It depends only on the product $\rho'\tau'$ of the derivatives of ρ and τ. Hence, once the metric tensor is fixed still some freedom is left to manipulate the geometry. We call this the *gauge freedom*. A sufficient condition for the existence of a dually flat geometry is established. It is shown that, if the coordinates of a parametric model are affine then the rho-tau metric tensor is Hessian and the dual coordinates are affine as well. We illustrate our approach using models belonging to deformed exponential families, and give a simple and precise characterization for the rho-tau metric to become Hessian.

1 Introduction

A *statistical manifold* [7, 1, 2] is an abstract manifold $\mathbb{M}$ equipped with a Riemannian metric g and an Amari-Chentsov tensor T. If the manifold is a smooth differentiable manifold then it can be realized [8] as a *statistical model*.

Most studies of statistical models are based on the widely used logarithmic embedding of probability density functions. Here, more generally embeddings are considered. Recent work [23, 11, 12] unifies the formalism of rho-tau embeddings [19] with statistical models belonging to deformed exponential families [10]. The present exposition continues this investigation.

The notion of a statistical manifold has been generalized in the non-parametric setting [14, 15] to include Banach manifolds. The corresponding terminology is

Jan Naudts
Universiteit Antwerpen, Antwerpen Belgium, e-mail: jan.naudts@uantwerpen.be

Jun Zhang
University of Michigan, Ann Arbor, MI U.S.A., e-mail: junz@umich.edu

used here, although up to now only a few papers have combined non-parametric manifolds with deformed exponential families [16, 13, 18, 9].

The rho-tau divergence is discussed in the next section. Eguchi [4, 5] proved under rather general conditions that, given a differentiable manifold, a divergence function defines a metric tensor and a pair of connections. These are derived in Section 4, respectively Section 6. Parametric statistical models are discussed in Section 7, which discusses Hessian geometry, and Section 8, which deals with deformed exponential families.

2 The statistical manifold

The points of ~~the~~ a given statistical manifold $\mathbb{M}$ are assumed to be random variables over some measure space $(\mathscr{X}, \mu)$. A random variable X is defined as any measurable real function. The expectation, if it exists, is denoted $\mathbb{E}_\mu X$. Throughout the text it is assumed that the manifold is differentiable and that for each X in $\mathbb{M}$ the tangent plane $T_X\mathbb{M}$ is well-defined.

The ~~Fréchet~~ derivative of a random variable is again a random variable. Therefore one can expect that the tangent vectors at a point X of $\mathbb{M}$ are random variables with vanishing expectation value. Let us assume that these tangent vectors can be used as a local chart in the vicinity of the point X and that they belong to some Banach space $\mathscr{B}$. Then $\mathbb{M}$ is a Banach manifold, provided a number of technical conditions are satisfied.

In the simplest case the manifold $\mathbb{M}$ consists of all strictly positive probability distributions on a discrete set $\mathscr{X}$. These probability distributions can be considered as positive-valued random variables with expectation equal to 1. The space $\mathscr{B}$ of all random variables is a Banach space for instance for the L^1 norm. The manifold $\mathbb{M}$ is a Banach manifold. Our approach here is the same as that adopted in [21], where random variables are called χ-functions, and functions of random variables are called χ-functionals.

In the more general situation the choice of an appropriate norm for the tangent vectors is not so simple. See the work of Pistone et all [14, 15, 16].

3 Rho-tau divergence

Given a strictly convex differentiable function h and a pair of real-valued random variables P and Q the Bregman divergence [3] is given by

$$\mathscr{D}(P,Q) = \mathbb{E}_\mu \left[h(P) - h(Q) - (P-Q)h'(Q) \right], \tag{1}$$

where h' denotes the derivative of h. A generalization involving two strictly increasing real functions $\rho(u)$ and $\tau(u)$ is proposed in [19]. For the sake of completeness

the definition is repeated here. Throughout the text these functions ρ and τ are assumed to be at least once, sometimes twice differentiable.

There exists a strictly convex function f with the property that $f' \circ \rho = \tau$. It is given by

$$f(u) = \int^{\rho^{-1}(u)} \tau(v) \mathrm{d}\rho(v). \tag{2}$$

The convex conjugate function f^* is therefore given by

$$f^*(u) = \int^{\tau^{-1}(u)} \rho(v) \mathrm{d}\tau(v), \tag{3}$$

provided the lower boundary of the integrals is chosen appropriately.

The original definition [19] of the rho-tau divergence can be written as

$$\mathscr{D}_{\rho,\tau}(P,Q) = \mathbb{E}_\mu \left[f(\rho(P)) + f^*(\tau(Q)) - \rho(P)\tau(Q) \right] \tag{4}$$

which is assumed to be $\leq +\infty$. The reformulation given below simplifies the proof of some of its properties.

Definition 1. Let be given two strictly increasing differentiable functions ρ and τ, defined on a common open ~~convex domain~~ interval D in $\mathbb{R}$. The rho-tau divergence of two random variables P and Q with values in D is given by

$$\mathscr{D}_{\rho,\tau}(P,Q) = \mathbb{E}_\mu \left(\int_Q^P \left[\tau(v) - \tau(Q) \right] \mathrm{d}\rho(v) \right). \tag{5}$$

This definition is equivalent to (4). To see this, split (5) into two parts. Use (2) to write the former contribution as $\mathbb{E}_\mu f \circ \rho(P) - \mathbb{E}_\mu f \circ \rho(Q)$ and the latter as $-\mathbb{E}_\mu \tau(Q)[\rho(P) - \rho(Q)]$. Use partial integration to prove that $f \circ \rho + f^* \circ \tau = \rho\tau$. This definition also generalizes (1). To see this take $I = f$, $\rho = \mathrm{id}$, and $\tau = I'$.

Note that the integral in (5) is a Stieltjes integral, which is well-defined because ρ and τ are strictly increasing functions. The result is non-negative. Hence, the μ-expectation is either convergent or it diverges to $+\infty$.

~~Let $p(\zeta,\eta)$ be the joint probability distribution that P has value ζ and Q has value η.~~ Let P and Q be two random variables with joint probability distribution $p(\zeta,\eta)$. Then (5) can be written as

$$\begin{aligned} \mathscr{D}_{\rho,\tau}(P,Q) &= \int p(\zeta,\eta) \mathrm{d}\zeta \mathrm{d}\eta \left(\int_\eta^\zeta \left[\tau(v) - \tau(\eta) \right] \mathrm{d}\rho(v) \right) \\ &\leq \int p(\zeta,\eta) \mathrm{d}\zeta \mathrm{d}\eta \, |\tau(\zeta) - \tau(\eta)| \, |\rho(\zeta) - \rho(\eta)| \\ &\leq \left\{ \mathbb{E}_\mu |\tau(P) - \tau(Q)|^2 \mathbb{E}_\mu |\rho(P) - \rho(Q)|^2 \right\}^{1/2}. \end{aligned} \tag{6}$$

To obtain the latter the Cauchy-Schwarz inequality is used.

Theorem 1. *$\mathscr{D}_{\rho,\tau}(P,Q) \geq 0$ with equality if $P = Q$. If μ is faithful, i.e. $\mathbb{E}_\mu P = 0$ implies $P = 0$ for any non-negative P, then $\mathscr{D}_{\rho,\tau}(P,Q) = 0$ implies $P = Q$.*

Proof

From (5) it is immediately clear that $\mathscr{D}_{\rho,\tau}(P,Q) \geq 0$ and $\mathscr{D}_{\rho,\tau}(P,P) = 0$. Assume now that $\mathscr{D}_{\rho,\tau}(P,Q) = 0$. By assumption this implies that

$$\int_Q^P [\tau(v) - \tau(Q)]\, \mathrm{d}\rho(v) = 0 \quad \mu\text{-almost everywhere.}$$

However, because τ and ρ are strictly increasing the integral is strictly positive unless $P = Q$, μ-almost everywhere.

□

It can be easily verified that the rho-tau divergence satisfies the following generalized ~~Pythagorian~~ Pythagorean equality for any three points P, Q, R

$$\mathscr{D}_{\rho,\tau}(P,Q) + \mathscr{D}_{\rho,\tau}(Q,R) - \mathscr{D}_{\rho,\tau}(P,R) = \mathbb{E}_\mu \left\{ [\rho(P) - \rho(Q)][\tau(R) - \tau(Q)] \right\}.$$

The general expression for the rho-tau entropy is

$$S_{\rho,\tau}(P) = -\mathbb{E}_\mu f(\rho(P)) + \text{ constant} = -\mathbb{E}_\mu \int^P \tau(u) \mathrm{d}\rho(u). \tag{7}$$

See for instance Section 2.6 of [23]. The function f is a strictly convex function which, given ρ, can still be chosen arbitrarily and then determines τ. ~~Hence, the~~ The following identity holds

$$\mathscr{D}_{\rho,\tau}(P,Q) = -S_{\rho,\tau}(P) + S_{\rho,\tau}(Q) - \mathbb{E}_\mu \left[\rho(P) - \rho(Q)\right] \tau(Q). \tag{8}$$

In [23, 12], we also discuss rho-tau cross-entropy, as well as the notion of "dual entropy" arising out of rho-tau embedding.

Rho-tau divergence $\mathscr{D}_{\rho,\tau}(P,Q)$ is a special form of the more general divergence function $\mathscr{D}^{(\alpha)}_{f,\rho}(P,Q)$ arising out of convex analysis, see [19, 20]:

$$\begin{aligned} \mathscr{D}^{(\alpha)}_{f,\rho}(P,Q) &= \frac{4}{1-\alpha^2} \\ &\times \mathbb{E}_\mu \left\{ \frac{1-\alpha}{2} f(\rho(P)) + \frac{1+\alpha}{2} f(\rho(Q)) - f\left(\frac{1-\alpha}{2}\rho(P) + \frac{1+\alpha}{2}\rho(Q) \right) \right\}. \end{aligned} \tag{9}$$

Clearly

$$\lim_{\alpha\to 1} \mathscr{D}^{(\alpha)}_{f,\rho}(P,Q) = \mathscr{D}_{\rho,\tau}(P,Q) = \mathscr{D}_{\tau,\rho}(Q,P);$$
$$\lim_{\alpha\to -1} \mathscr{D}^{(\alpha)}_{f,\rho}(P,Q) = \mathscr{D}_{\rho,\tau}(Q,P) = \mathscr{D}_{\tau,\rho}(P,Q);$$

with $f' \circ \rho = \tau$ (and equivalent $(f^*)' \circ \tau = \rho$, with f^* denoting convex conjugate of f). Though in $\mathscr{D}^{(\alpha)}_{f,\rho}(P,Q)$ 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 [19, 22].

4 Tangent vectors

The rho-tau divergence introduced above can be used to fix a Riemannian metric on the tangent planes of the statistical manifold $\mathbb{M}$. ~~The Fréchet derivative of a random variable is again a random variable. Therefore one can expect that the tangent vectors at the point P of $\mathbb{M}$ are random variables with vanishing expectation value. The metric tensor is then used to turn the tangent plane $T_P\mathbb{M}$ into a (pre-) Hilbert space.~~

In the standard situation of the Fisher-Rao metric the point P is a probability density function p^θ, parametric with $\theta \in \mathbb{R}^n$. A short calculation gives

$$\partial_j \mathbb{E}_\mu p^\theta Y = \left\langle \partial_j \log p^\theta, Y \right\rangle_\theta, \tag{10}$$

with $\langle X, Y\rangle_\theta = \mathbb{E}_\mu p^\theta XY$, and where ∂_j is an abbreviation for $\partial/\partial\theta^j$. The metric tensor is then given by

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

The score variables $\partial_j \log p^\theta$ have vanishing expectation and span the tangent plane at the point p^θ.

These expressions are now generalized. ~~using the Fréchet derivative in non-parametric setting.~~ Fix P in $\mathbb{M}$. Make the assumption that there exists some open neighborhood U of P in $\mathbb{M}$ and a one-to-one correspondence χ_P between elements Q of U and tangent vectors $X = \chi_P(Q)$ of $T_P\mathbb{M}$, satisfying $\chi_P(P) = 0$. This map χ_P is used as a local chart centered at the point P. The directional derivative d_X is then defined as

$$d_X P := \lim_{\varepsilon \to 0} \frac{\chi_P^{-1}(\varepsilon X) - \chi^{-1}(0)}{\varepsilon},$$

and is assumed to exist for all $X \in T_P\mathbb{M}$. Here, we leave the topology unspecified.

~~Let X be random variable. First, its Fréchet derivative (directional derivative) d_X is defined as (for a smooth function h)~~

$$d_X h(P) := \lim_{\varepsilon \to 0} \frac{h(P + \varepsilon X) - h(P)}{\varepsilon}, \quad REMOVE$$

~~which is a random function linearly dependent on X.~~

Now we take one of the two increasing functions ρ and τ, say ρ, to define a two-point correlation function $\mathbb{E}_\mu \rho(P) Y$, and the other function, τ, to ~~deform~~ act as a deformed logarithmic function replacing the logarithmic function which appears in the definition of the standard scores. The expression analogue to (10) now involves

~~Fréchet~~ derivatives of $\mathbb{E}_\mu \rho(P)Y$ and of $\tau(P)$. It becomes

$$d_X \mathbb{E}_\mu \rho(P)Y = \left\langle d_X \tau(P), Y \right\rangle_P, \tag{11}$$

with

$$\langle X, Y \rangle_P = \mathbb{E}_\mu \frac{\rho'(P)}{\tau'(P)} XY.$$

This relation should hold for any P in $\mathbb{M}$ and X in $T_P\mathbb{M}$, and for any random variable Y. The metric tensor $g_{XY} \equiv g(X,Y)$ becomes

$$\begin{aligned} g_{XY}(P) &= \left\langle d_X \tau(P), d_Y \tau(P) \right\rangle_P \\ &= \mathbb{E}_\mu \rho'(P)\tau'(P) d_X P d_Y P. \end{aligned} \tag{12}$$

This metric tensor is related to the divergence function introduced in the previous section by

$$d_Y^{\mathrm{P}} d_X^{\mathrm{Q}} \mathscr{D}_{\rho,\tau}(P,Q)\Big|_{P=Q} = -g_{XY}(P),$$

where d^{P} is the ~~Fréchet~~ derivative acting only on P and d^{Q} acts only on Q. See [21] for the derivation of the metric tensor in the form of (12) for the non-parametric setting.

In the case of a model p^θ which belongs to the exponential family the tangent plane can be identified with the coordinate space. The chart becomes $\chi_{p^\theta}(p^\zeta) = \zeta - \theta$ so that

$$d_\zeta p^\theta := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(p^{\theta + \varepsilon(\zeta - \theta)} - p^\theta \right).$$

If $(\zeta - \theta)_i = \delta_{i,j}$ then $d_\zeta p^\theta = \partial_j p^\theta$ follows and (11) reduces to (10).

5 Gauge freedom

From (12) it is clear that the metric tensor depends only on the product $\rho'\tau'$ and not on ρ and τ separately. This implies that once the metric tensor is fixed there remains one function to be chosen freely, either the embedding ρ or the deformed logarithm τ, keeping $\rho'\tau'$ fixed. This is what we call the gauge freedom of the rho-tau formalism.

The notion of gauge freedom is common in Physics to mark the introduction of additional degrees of freedom which do not modify the model but control some of its appearances. Here, the Riemannian metric of the manifold is considered to be an essential feature while the different geometries such as the Riemannian geometry or Amari's dually flat geometries are attributes which give a further characterization.

It is known for long that distinct choices of the divergence function can lead to the same metric tensor. The present formalism offers the opportunity to profit from this freedom. Quantities such as the divergence function, the entropy or the

alpha-family of connections depend on the specific choice of both ρ and τ. This is illustrated further on.

$\rho(u)$	$\tau(u)$	$(\rho'\tau')(u)$	$f(u)$	$f^*(u)$
u	$\log u$	$\frac{1}{u}$	$u[\log u - 1]$	e^u
$2\sqrt{u}$	$2\sqrt{u}$	$\frac{1}{u}$	$\frac{1}{2}u^2$	$\frac{1}{2}u^2$
u	$\log_q(u)$	$\frac{1}{u^q}$	$\frac{u}{2-q}\left[\log_q(u)-1\right]$	$\frac{1}{2-q}\left[\exp_q(u)\right]^{2-q}$
$\rho(u)$	$\log_\rho(u)$	$\frac{\rho'}{\rho}(u)$	$u[\log u - 1]$	e^u
u	$\log_\phi(u)$	$\frac{1}{\phi(u)}$	$u\log_\phi(u) - \int_1^u \frac{v}{\phi(v)}\mathrm{d}v$	$\int_1^{\exp_\phi(u)} \frac{v}{\phi(v)}\mathrm{d}v$

Table 1 Examples of ρ, τ combinations

~~The rho-affine gauge, is characterized by $\rho = \mathrm{id}$.~~ The simplest choice to fix the gauge is $\rho = \mathrm{id}$. Several ~~of the generalized~~ classes generalizing Bregman divergences found in the literature , e.g. [10, 6], belong to this case. The phi-divergence of [10] is obtained by choosing τ equal to the deformed logarithm $\log_\phi$ (see Section 8), the derivative of which is $1/\phi$. This implies $\rho'\tau' = 1/\phi$, which is also the condition for the deformed metric tensor of [10] to be conformally equivalent with (12). The U-divergence of [6] is obtained by taking τ equal to the inverse function of U'. These were discussed in detail in [23, 11, 12].

Also of interest is the gauge defined by $\rho(u) = 1/\tau'(u)$. Let $\log_\rho$ be the corresponding deformed logarithm (see (21) below). It satisfies $\log_\rho(u) = \tau(u) - \tau(1)$. Hence, the entropy becomes

$$S_{\rho,\tau}(P) = -\mathbb{E}_\mu \rho(P)\tau(P) + \mathbb{E}_\mu P + \text{ constant.}$$

The divergence becomes

$$\mathscr{D}_{\rho,\tau}(P,Q) = \mathbb{E}_\mu \rho(P)\left[\log_\rho(P) - \log_\rho(Q)\right] - \mathbb{E}_\mu\left[P - Q\right].$$

This expression is an obvious generalization of the Kullback-Leibler divergence.

6 Induced geometry

A divergence function not only fixes a metric tensor by taking two derivatives, it also fixes a pair of torsion-free connections by taking an extra derivative w.r.t. the first argument [4] [5]. In particular, the rho-tau-divergence (5) determines an alpha-family of connections [19, 21, 11].

A covariant derivative ∇_Z with respect to a vector field Z is defined by

$$\langle \nabla_Z d_X \tau(P), d_Y \tau(P)\rangle_P = -d_Z^{\mathrm{P}} d_Y^{\mathrm{P}} d_X^{\mathrm{Q}} \mathscr{D}_{\rho,\tau}(P,Q)\Big|_{Q=P}.$$

A short calculation of the righthand side, with $\mathscr{D}_{\rho,\tau}$ defined by (4), gives

$$\langle \nabla_Z d_X \tau(P), d_Y \tau(P)\rangle_P = \mathbb{E}_\mu\left[d_X \tau(P)\right] d_Z d_Y \rho(P).$$

Let $\nabla_Z^{(1)} = \nabla_Z$ and let $\nabla_Z^{(-1)}$ be the operator obtained by interchanging ρ and τ. This is

$$\begin{aligned}\langle \nabla_Z^{(-1)} d_X \tau(P), d_Y \tau(P)\rangle_P &= \mathbb{E}_\mu\left[d_X \rho(P)\right] d_Z d_Y \tau(P)\\ &= \langle d_X \tau(P), d_Z d_Y \tau(P)\rangle_P. \end{aligned} \tag{13}$$

This shows that $\nabla_Z^{(-1)}$ is the ~~hermitian conjugate~~ adjoint of d_Z with respect to g. In addition one has

$$\langle \nabla_Z^{(1)} d_X \tau(P), d_Y \tau(P)\rangle_P + \langle d_X \tau(P), \nabla_Z^{(-1)} d_Y \tau(P)\rangle_P = d_Z g_{XY}(P). \tag{14}$$

The latter expression shows that the connections $\nabla^{(1)}$ and $\nabla^{(-1)}$ are the dual of each other with respect to g. The alpha-family of connections is then obtained by linear interpolation with $\alpha \in [-1,1]$

$$\nabla_Z^{(\alpha)} = \frac{1+\alpha}{2}\nabla_Z^{(1)} + \frac{1-\alpha}{2}\nabla_Z^{(-1)}, \tag{15}$$

such that the covariant derivatives $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are mutually dual. In particular, $\nabla^{(0)}$ is self-dual and therefore coincides with the Levi-Civita connection. The family of α-connections (15) is induced by the divergence function $\mathscr{D}_{f,\rho}^{(\alpha)}(P,Q)$ given by (9), with corresponding α-values. Furthermore, upon switching $\rho \leftrightarrow \tau$ in the divergence function, the designation of 1-connection vs (-1)-connection also switches.

From (13) it is clear that the covariant derivative $\nabla_Z^{(-1)}$ vanishes on the tangent plane when

$$\langle d_X \tau(P), d_Z d_Y \tau(P)\rangle_P = 0 \quad \text{for all } X,Y \in T_P\mathbb{M}. \tag{16}$$

If this holds for all P in $\mathbb{M}$ then the ~~$\nabla_Z^{(-1)}$~~ $\nabla^{(-1)}$-geometry is flat. This implies that the dual geometry ~~$\nabla_Z^{(1)}$~~ $\nabla^{(1)}$ is also flat — see Theorem 3.3 of [1]. The interpretation of (16) is that all second derivatives $d_Z d_Y \tau(P)$ are orthogonal to the tangent plane.

7 Parametric models

The previous sections deal with the geometry of arbitrary manifolds consisting of random variables, without caring whether they possess special properties. Now parametric models with a Hessian metric g are considered.

From here on the random variables of the manifold $\mathbb{M}$ are probability distribution functions p^θ, labeled with coordinates θ belonging to some open convex subset U of $\mathbb{R}^n$. The manifold is assumed to be differentiable. In particular, the θ^i are covariant coordinates and the assumption holds that the derivatives $\partial_i p^\theta \equiv \partial p^\theta / \partial \theta^i$ form a basis for the tangent plane $T_\theta \mathbb{M} \equiv T_{p^\theta}\mathbb{M}$. The simplifications induced by this setting are that the tangent planes are finite-dimensional and that the dual coordinates belong again to $\mathbb{R}^n$. For general Banach manifolds both properties need not to hold. ~~The tangent plane $T_\theta \mathbb{M} \equiv T_{p^\theta}\mathbb{M}$ now has a finite basis of vectors $\partial_i \tau(p^\theta) = \tau'(p^\theta)\partial_i$, which is isomorphic to (in fact iso-direction with) ∂_i, and hence linearly independent.~~ The assumptions imply that the metric tensor

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

is a strictly positive-definite matrix.

The metric g of the manifold $\mathbb{M}$ is said to be Hessian if there exists a strictly convex function $\Phi(\theta)$ with the property that $g_{ij}(\theta) = \partial_i \partial_j \Phi(\theta)$. See for instance [17]. Let $\Psi(\eta)$ denote the convex dual of $\Phi(\eta)$. This is

$$\Psi(\eta) = \sup_\theta \{ \langle \eta, \theta \rangle - \Phi(\theta) : \theta \in U \}.$$

Let U^* denote the subset of $\mathbb{R}^n$ of η for which the maximum is reached at some θ in U. This θ is unique and defines a bijection $\theta \mapsto \eta$ between U and U^*. These η are dual coordinates for the manifold $\mathbb{M}$. Conversely [11], if there exist coordinates η_i for which $g_{ij}(\theta) = \partial_j \eta_i$ then the rho-tau metric tensor g is Hessian.

The condition (16) for $\nabla^{(-1)}$ to vanish can now be written as

$$\langle \partial_i \tau(p^\theta), \partial_k \partial_j \tau(p^\theta) \rangle_\theta = 0, \qquad \text{for all } \theta \in U \text{ and for all } i, j, k. \tag{17}$$

Theorem 2. *Assume that the θ^i are affine coordinates such that ~~$\Gamma^{(-1)}(\theta) = 0$~~ $\nabla^{(-1)} = 0$. Then*

1) the metric tensor g is Hessian;
2) the η_i are affine coordinates for the ~~connection $\Gamma^{(1)}$~~ $\nabla^{(1)}$-geometry.

Proof

1) ~~Use ([17]) and the duality relation ([14]) to obtain~~ The metric tensor (12) becomes

$$g_{ij}(p^\theta) = \langle \partial_i \tau(p^\theta), \partial_j \tau(p^\theta) \rangle_\theta = \mathbb{E}_\mu \left(\partial_i \tau(p^\theta) \right) \partial_j \rho(p^\theta) = \mathbb{E}_\mu \left(\partial_j \tau(p^\theta) \right) \partial_i \rho(p^\theta).$$

This implies

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu\left(\partial_k\partial_i\tau(p^\theta)\right)\partial_j\rho(p^\theta) + \mathbb{E}_\mu\left(\partial_i\tau(p^\theta)\right)\partial_k\partial_j\rho(p^\theta),$$

but also

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu\left(\partial_k\partial_j\tau(p^\theta)\right)\partial_i\rho(p^\theta) + \mathbb{E}_\mu\left(\partial_j\tau(p^\theta)\right)\partial_k\partial_i\rho(p^\theta).$$

These equations simplify by means of (17). The result is

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu\left(\partial_i\tau(p^\theta)\right)\partial_k\partial_j\rho(p^\theta) = \mathbb{E}_\mu\left(\partial_j\tau(p^\theta)\right)\partial_k\partial_i\rho(p^\theta).$$

This implies that $\partial_k g_{ij}(\theta) = \partial_i g_{kj}(\theta)$. Hence there exist functions $\eta_j(\theta)$ such that $g_{ij}(\theta) = \partial_i\eta_j(\theta)$. As remarked above, it is proved in [11] that this suffices to conclude that the metric g is Hessian.

2) Let us show that

$$\eta(t) = (1-t)\eta^{(1)} + t\eta^{(2)}. \tag{18}$$

is a solution of the Euler-Lagrange equations

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta^i + \Gamma^i_{km}\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right) = 0. \tag{19}$$

Here, the Γ^i_{km} are the coefficients of the connection $\Gamma^{(1)}$ induced by the $\nabla^{(1)}$-geometry. They follow from

$$\Gamma_{ij,k} = \partial_i g_{jk}(\theta). \tag{20}$$

One has

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^i = \frac{\partial\theta^i}{\partial\eta_j}\frac{\mathrm{d}\eta_j}{\mathrm{d}t} = g^{ij}(\theta)\left[\eta_j^{(2)} - \eta_j^{(1)}\right]$$

and

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta^i &= \frac{\mathrm{d}}{\mathrm{d}t}g^{ij}(\theta)\left[\eta_j^{(2)} - \eta_j^{(1)}\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]\frac{\mathrm{d}\theta^k}{\mathrm{d}t}\left[\eta_j^{(2)} - \eta_j^{(1)}\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]g^{kl}(\theta)\left[\eta_l^{(2)} - \eta_l^{(1)}\right]\left[\eta_j^{(2)} - \eta_j^{(1)}\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]g_{jm}(\theta)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right).\end{aligned}$$

The l.h.s. of (19) becomes

$$\text{l.h.s.} = \left\{\left[\partial_k g^{ij}(\theta)\right]g_{jm}(\theta) + \Gamma^i_{km}\right\}\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right).$$

This vanishes because (20) implies

$$\Gamma^i_{km} = -\left[\partial_k g^{ij}(\theta)\right] g_{jm}(\theta).$$

□

It is important to realize that the discussion in this section is generic for parametric models, without assuming particular parametric families.

8 The deformed exponential family

A repeated measurement of n independent random variables $F_1, \cdots, F_n$ results in a joint probability distribution $\pi(\zeta_1, \cdots, \zeta_n)$, which describes the probability that the true value of the measured data equals ζ. ~~A model belonging to the exponential family can be used to approximate the empirical data.~~ More generally, the model can be taken ~~in~~ to be a deformed exponential family (Generalized Linear Model), obtained by using a deformed exponential function $\exp_\phi$. Following [10], a deformed logarithm $\log_\phi$ is defined by

$$\log_\phi(u) = \int_1^u \mathrm{d}v \, \frac{1}{\phi(v)}, \tag{21}$$

where $\phi(v)$ is strictly positive and integrable on the open interval $(0, +\infty)$. The deformed exponential function $\exp_\phi(u)$ is the inverse function of $\log_\phi(u)$. It is defined on the range $\mathscr{R}$ of $\log_\phi(u)$, but is eventually extended with the value 0 if $u < \mathscr{R}$ and with the value $+\infty$ if $u > \mathscr{R}$.

The expression for the probability density function then becomes

$$p^\theta(x) = \exp_\phi\left(\sum_{k=1}^n \theta^k F_k(x) - \alpha(\theta)\right). \tag{22}$$

The function $\alpha(\theta)$ serves to normalize p^θ and is assumed to exist~~s~~ within the open convex domain $U \subset \mathbb{R}^n$ in which the model is defined. One can show [10] that it is a convex function. However, in general it does not coincide with the potential $\Phi(\theta)$ of the previous section. The explanation is that *escort probabilities* come into play. Indeed, from

$$0 = \partial_i \mathbb{E}_\mu p^\theta = \mathbb{E}_\mu \phi(p^\theta)\left[F_i - \partial_i \alpha\right]$$

follows that

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

with the escort expectation $\tilde{\mathbb{E}}_\theta$ defined by

$$\tilde{\mathbb{E}}_\theta Y = \frac{\mathbb{E}_\mu \phi(p^\theta) Y}{\mathbb{E}_\mu \phi(p^\theta)}.$$

Only in the non-deformed case, when $\phi(u) = u$, the escort $\tilde{\mathbb{E}}_\theta$ coincides with the model expectation $\mathbb{E}_\theta$. Then the dual coordinates η_i satisfy $\eta_i = \mathbb{E}_\theta F_i = \partial_i \alpha(\theta)$.

In general, the rho-tau metric tensor g of the deformed exponential model is *not* Hessian.We have the following Theorem (see [12])

Theorem 3. *With respect to the (deformed) ϕ-exponential family p^θ obeying (22), the rho-tau metric tensor g is Hessian if and only if*

$$\rho' \tau' \phi = id.$$

In such a situation, the rho-tau metric tensor is conformally equivalent with the metric tensor obtained by taking the Hessian of the normalization function α; for the latter the potential $\Phi(\theta)$ is constructed in [10]. However, there still leaves a gauge freedom. The question is then whether one can choose ρ and τ so that condition (16) for the dually flat geometry is satisfied. A sufficient condition is that $\rho = \mathrm{id}$ and $\tau = \log_\phi$. This is the rho-affine gauge. In this gauge both the θ^i and the η_i coordinates are affine and the model has a dually flat structure.

Acknowledgement
The research reported here is supported by DARPA/ARO Grant W911NF-16-1-0383 (PI: Jun Zhang).

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry. AMS Monograph, Oxford University Press, 2000. (Originally published in Japanese by Iwanami Shoten, Tokyo, Japan, 1993.)
2. Ay N., Jost, J., LÊ, H.V., Schwachhöfer, L.: Information Geometry. (Springer, 2017)
3. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.* **7**0, 200–217 (1967).
4. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **11**, 793–803 (1983).
5. Eguchi, S.: A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **15**, 341–391 (1985).
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. Lauritzen, S.: Statistical manifolds. In *Differential Geometry in Statistical Inference*; Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R., Eds.; IMS: Hayward, CA, USA, Lecture Notes **10**, 163–216 (1987).
8. Lê, H.V.: Statistical manifolds are statistical models. *J. Geom.* **84** 83 – 93 (2005).
9. Montrucchio, L., Pistone, G.: Deformed exponential bundle: The linear growth case. In: *Geometric Science of Information,* GSI 2017 LNCS proceedings, F. Nielsen and F. Barbaresco eds., (Springer, 2017), p. 239–246.
10. Naudts, J.: Estimators, escort probabilities, and phi-exponential families in statistical physics. *J. Ineq. Pure Appl. Math.* **5**, 102 (2004).

11. Naudts, J., Zhang, J.: Information geometry under monotone embedding. Part II: Geometry. in: *Geometric Science of Information,* GSI 2017 LNCS proceedings, F. Nielsen and F. Barbaresco eds., (Springer, 2017), p.215–222.
12. Naudts, J., Zhang J.:Information Geometry Under Monotone Embedding. *Information Geometry,* (under review).
13. Newton, N. J.: An infinite-dimensional statistical manifold modeled on Hilbert space. *J. Funct. Anal.* **263**, 1661–1681 (2012).
14. Pistone, G., Sempi, C.: An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **33**, 1543–1561 (1995).
15. Pistone G., Rogantin M.P.: The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli* **5**, 721–760 (1999).
16. Pistone, G.: κ-exponential models from the geometrical viewpoint. *Eur. Phys. J. B* **70**, 29–37 (2009).
17. Shima, H.: The Geometry of Hessian Structures. (World Scientific, 2007)
18. Vigelis, R.F., Cavalcante, C.C.: On ϕ-families of probability distributions. *J. Theor. Probab.* **26**, 870–884 (2013).
19. Zhang, J.: Divergence function, duality, and convex analysis. *Neural Comput.* **16**, 159–195 (2004).
20. Zhang, J.: Referential duality and representational duality on statistical manifolds. In *Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo*, Japan, 2005, pp. 58–67.
21. Zhang, J.: Nonparametric Information Geometry: From Divergence Function to Referential-Representational Biduality on Statistical Manifolds. *Entropy* **15** 1 (2013).
22. Zhang, J.: On monotone embedding in information geometry. *Entropy* **17**, 4485–4499 (2015).
23. Zhang, J., Naudts, J.: Information geometry under monotone embedding. Part I: Divergence functions. in: *Geometric Science of Information,* GSI 2017 LNCS proceedings, F. Nielsen and F. Barbaresco eds., (Springer, 2017), p.205–214.