Information Geometry (2024) 7:S623–S649
https://doi.org/10.1007/s41884-023-00121-0

SURVEY PAPER

Legendre duality: from thermodynamics to information geometry

Jan Naudts[1] · **Jun Zhang**[2]

Received: 29 March 2023 / Revised: 19 October 2023 / Accepted: 27 October 2023 /
Published online: 8 November 2023

Abstract

This paper reviews the role of convex duality in Information Geometry. It clarifies the notion of bi-orthogonal coordinates associated with Legendre duality by treating its two underlying aspects separately: as a dual coordinate system and as a bi-orthogonal frame. It addresses the deformation of exponential families in a way that stills preserves the dually-flat geometry of 1- and (-1)-connections. The deformation involves a metric which generalizes the Fisher–Rao metric controlled by one degree of freedom and a pair of connections controlled by an additional degree of freedom.

Keywords Convex duality · Dually-flat geometry · Deformed exponential and logarithmic functions · Rho–tau connections

1 Introduction

Several dualities play their roles in Information Geometry (see our companion paper). The dually flat geometry [1, 2] enriches the convex duality based on Legendre transformations. Legendre transformation is a familiar concept in Mechanics, in Thermodynamics and in Statistical Physics. Of particular significance is that in the latter, the family of exponential probability distributions (Boltzmann–Gibbs distributions) is explicitly modeled. While Chentsov's (1972) work on statistical inference also makes

Communicated by Hiroshi Matsuzoe.

✉ Jan Naudts
jan.naudts@uantwerpen.be

Jun Zhang
junz@umich.edu

[1] Department of Physics, Universiteit Antwerpen, Universiteitsplein 1, Wilrijk, Antwerpen 2610, Belgium

[2] Department of Psychology and Department of Statistics, University of Michigan, 530 Church Street, Ann Arbor, Michigan 48109, USA

use of the Legendre duality in exponential families as studied in physics, Amari and Nagaoka's (1982) introduction of a dually flat geometry associated to the Legendre duality is independently motivated by the duality of the exponential and mixture families in statistics. Given the importance of Legendre duality for Information Geometry, the focus of the present work is to revisit the historical roots of convex duality in Thermodynamics and in Information Geometry.

Through mid 19th century the role of Legendre transformations in Classical Mechanics became well-known to the scientific community. In 1869 F Massieu [5] introduces the notion of a thermodynamic potential and shows that quantities such as volume V and pressure P or energy U and temperature T (or rather its inverse $1/T$) are dual parameters linked by a Legendre transformation. The equilibrium state of a substance is characterized by a small number of variables, the choice of which is non-unique. The thermodynamic potential is a function of these variables. By taking derivatives of the thermodynamic potential further properties of the substance are obtained. For instance, the state of a fluid of N particles is determined by its temperature T and volume V. The corresponding thermodynamic potential is $F = U - TS$, where U is the energy and S the entropy. The state of the liquid can as well be described by energy U and volume V. Then the relevant thermodynamic potential is the entropy S, which is obtained by performing the Legendre transform of F/T. The book of H Callen [8] lists the most common changes of thermodynamic variables and the resulting thermodynamic potentials.

The formalism of Statistical Mechanics, as we know it today, is recorded in the book of JW Gibbs [12] dated 1902. The emphasis is on probability distributions now called Gibbs distributions. They depend on a small number of parameters and are used to calculate the expected value of functions defined on the sample space. In more recent mathematical treatises the probability distributions are replaced by probability measures. See for instance Ruelle [27]. The notion of a random variable and its probability distribution does not play an important role, mainly because the models of Statistical Physics are usually studied in the limit of an infinite number of degrees of freedom.

The application of Differential Geometry to Thermodynamics and Statistical Mechanics is picking up pace with the works of Weinhold [31] and of Ruppeiner [28] in the 1970s. The equilibrium states of Thermodynamics form a parameterized manifold. On the tangent planes a metric is defined by taking second derivatives of the entropy function. The description of the geometry of the manifold is completed with the Levi–Civita connection. Curvature of the manifold is interpreted as being due to interactions between particles.

In the 1980s Amari [1–3] introduces the dually flat geometry for models belonging to an exponential family. In the language of Statistical Mechanics these are models with a *parameter-dependent* Gibbs distribution. Dual connections are related to dual parameters by the observation that if a connection allows affine parameters then there exist dual parameters and they are affine parameters for the dual connection (Theorem 3.6 of [3]).

In the Statistical Physics community, Tsallis [29] starts an initiative to study models not belonging to an exponential family. A guiding idea [30] is to modify the exponential function which appears in the definition of an exponential family. Based on this concept

one of us (JN) works out [16] the notion of a deformed exponential family and shows that part of the work of Amari carries over to this larger class of models.

In the same time the other author (JZ) introduces [33] the ρ, τ formalism to deform that exponential function and the information-geometric framework attached to it. The ρ, τ formulation starts from an arbitrary pair ρ, τ of strictly increasing functions, and recognizes an extra degree of freedom in the Legendre transform, which links pairs of convex functions. This makes the deformation of the probability density function separately controlled by the deformation of the entropy, cross-entropy, or relative entropy functions. Adapting the machinery of Information Geometry, including how divergence functions (relative entropies) induce a Riemannian metric and a pair of affine connections, generalizations to the Fisher–Rao metric and to the α-connections are obtained. The functions ρ and τ intermingle two different kinds of duality: *representation duality* and *reference duality* into *reference-representation biduality*. See the companion paper for details.

The goal of this paper is to review the historical steps in application of Legendre duality to modeling the probability families. Section 2 recalls Legendre duality in the context of Riemannian manifolds with a Hessian metric. Section 3 is restricted to manifolds of probability measures and discusses tangent vectors and the Fisher–Rao metric. Section 4 discusses exponential families. Section 5 overviews the deformed exponential models.

2 Many faces of a Hessian metric

In this section, $\mathbb{M}$ is a Riemannian manifold with a global coordinate system. A one-to-one map is given between points μ_θ in $\mathbb{M}$ and points θ in an open convex domain $\mathbb{D} \subset \mathbb{R}^n$. In a differentiable manifold, the latter is usually called the "chart", and θ is treated as the coordinate function on $\mathbb{M}$. However, in Physics and Statistics, one always starts with a coordinate chart and then considers a point on $\mathbb{M}$ as a function on the chart. This is the convention we are going to adopt, i.e., treating μ_θ as the point which is specified by $\theta \in \mathbb{R}^n$. The stipulation that the μ_θ are probability measures is postponed to the next section.

2.1 Dual coordinates

We assume that the manifold is embedded in a Banach space $\mathcal{B}$ and that the partial derivatives $\partial\mu_\theta/\partial\theta^i$ exist as Gâteaux derivatives and form a linearly independent set of vectors in $\mathcal{B}$. This means that the norms

$$\left\| \frac{1}{\epsilon}(\mu_{\theta+\epsilon g_i} - \mu_\theta) - \frac{\partial \mu_\theta}{\partial \theta^i} \right\|$$

tend to 0 as ϵ tends to zero. In this expression g_i is the vector with components g_i^j and ϵ is a real number. Note that the matrix with entries g_i^j is the identity matrix.

The canonical basis vectors in the tangent bundle are denoted ∂_i and satisfy

$$\partial_i \mu_\theta = \frac{\partial \mu_\theta}{\partial \theta^i}.$$

In this way the tangent space $T_\theta \mathbb{M}$ is identified with a linear subspace of the Banach space $\mathcal{B}$. In the case of a manifold of probability measures this identification is worked out in Sect. 3.2 of [4]. The identification is a key element in the unified treatment of manifolds of probability distributions and of manifolds of quantum states [9, 18].

The metric tensor in the tangent plane $T_\theta \mathbb{M}$ is denoted $g(\theta)$ as usual. Its components g_{ij} are given by the inner product $\langle \partial_i, \partial_j \rangle_\theta$ of the tangent vectors ∂_i, ∂_j. The standard conventions of physicists for raising and lowering of indices are followed. In particular, g^{ij} are the components of the inverse g^{-1} of the metric tensor.

Given vector fields X and Y with components X^i, respectively Y^i, i.e.

$$X = X^i \partial_i \quad \text{and} \quad Y = Y^i \partial_i,$$

one has

$$\langle X, Y \rangle_\theta = X^i g_{ij}(\theta) Y^j.$$

The dual basis is a set of vectors $\{\partial^i\}_{i=1}^n$ in the dual space (i.e., the cotangent space $T_\theta^* \mathbb{M}$) such that:

$$\partial^i(\partial_j) = (\partial^i, \partial_j) = g^i_j.$$

Note that $(\cdot, \cdot)$ here denotes the dual pairing of an element of $T_\theta^* \mathbb{M}$ with a element of $T_\theta \mathbb{M}$ while $\langle \cdot, \cdot \rangle_\theta$ refers to the inner-product of two vectors in $T_\theta \mathbb{M}$.

On a Riemannian manifold we can identify the elements of the cotangent space with elements of the tangent space. Therefore, we can construct another frame (i.e., another set of linearly independent vectors *in the tangent space*), denoted $\partial_i^\perp$, $i = 1, \cdots, n$, such that

$$\left\langle \partial_i^\perp, \partial_j \right\rangle_\theta = \delta_{ij}. \tag{1}$$

Here, δ_{ij} is Kronecker's delta function. We call $\{\partial_i^\perp\}_{i=1}^n$ the *bi-orthogonal frame* with respect to the coordinate basis $\{\partial_k\}_{k=1}^n$. It satisfies

$$\partial_i^\perp = \delta_{ij} g^{jk} \partial_k. \tag{2}$$

Note that, given the Riemannian metric g, the bi-orthogonal frame always exists and is uniquely so, the frame in general may not be a coordinate frame. The frame $\{\partial_i^\perp\}_{i=1}^n$ is a coordinate frame if there exists some coordinate system $\{\zeta^i\}_{i=1}^n$ such that

$$\partial_i^\perp = \frac{\partial}{\partial \zeta^i}. \tag{3}$$

Consider another coordinate system $\{\eta_i\}_{i=1}^n$ with base ∂^i

$$\partial^i = \frac{\partial}{\partial \eta_i} = \frac{\partial \theta^j}{\partial \eta_i} \frac{\partial}{\partial \theta^j}.$$

The η coordinate system is said to be the *dual* of the coordinate system $\{\theta^i\}_{i=1}^n$ with respect to g if the Jacobian governing the coordinate transform $\partial \eta / \partial \theta$ coincides with the metric tensor g, that is, when the following is satisfied

$$\partial_i \eta_j = \frac{\partial \eta_j}{\partial \theta^i}(\theta) = g_{ij}(\theta). \tag{4}$$

Because the symmetric and positive-definite requirement of g imposes stringent requirements on the Jacobian of the coordinate transform $\partial \eta_j / \partial \theta_i$, the dual coordinate system may not exist for a generic Riemannian manifold.

In [3] dual coordinates are defined in the context of convex duality, which is discussed in the next section. Duality w.r.t. the metric tensor g is a more general concept since it does not require that the metric is Hessian. This is the topic of the following proposition.

Proposition 1 (Biorthogonal frame, dual coordinate system, bi-orthogonal coordinates) *Let be given a coordinate system θ, and assume that the elements g_{jk} of the metric tensor g are smooth functions of θ. The following are equivalent:*

(a) *There exists a dual coordinate system $(\eta_i)_{i=1}^n$;*
(b) *The bi-orthogonal frame $\{\partial_i^\perp\}_{i=1}^n$ is a coordinate frame, i.e.,*

$$\partial_i^\perp = \frac{\partial}{\partial \zeta^i};$$

for some coordinate system $(\zeta^i)_{i=1}^n$.
(c) *The metric g is Hessian with respect to θ, i.e. there exists a smooth function $\Phi(\theta)$ defined on the convex domain $\mathbb{D}$ such that $g_{ij} = \partial_i \partial_j \Phi$ holds.*

In such case, coordinates ζ^j of the bi-orthogonal frame coincide with the dual coordinates η_i of θ in the sense that $\eta_i = \delta_{ij} \zeta^j$.

Proof **(a) $\Rightarrow$ (b)**
Let ζ^i be determined by $\delta_{ij} \zeta^j = \eta_i$. From

$$g^i_j = \frac{\partial \theta^i}{\partial \theta^j} = \frac{\partial \theta^i}{\partial \eta_k} \frac{\partial \eta_k}{\partial \theta^j} = \frac{\partial \theta^i}{\partial \eta_k} g_{kj}(\theta)$$

one obtains

$$\frac{\partial \theta^i}{\partial \eta_j} = g^{ij}(\theta).$$

Use this to calculate

$$\frac{\partial}{\partial\zeta^i}\mu_\theta = \delta_{ij}\frac{\partial}{\partial\eta_j}\mu_\theta = \delta_{ij}\frac{\partial\theta^k}{\partial\eta_j}\partial_k\mu_\theta = \delta_{ij}g^{jk}(\theta)\partial_k\mu_\theta$$

and

$$\begin{aligned}\left\langle \frac{\partial}{\partial\zeta^i}, \partial_l \right\rangle_\theta &= \left\langle \frac{\partial}{\partial\zeta^i}\mu_\theta, \partial_l\mu_\theta \right\rangle \\ &= \delta_{ij}g^{jk}(\theta)\langle\partial_k\mu_\theta, \partial_l\mu_\theta\rangle \\ &= \delta_{ij}g^{jk}(\theta)g_{kl}(\theta) \\ &= \delta_{ij}g_l^j \\ &= \delta_{il}.\end{aligned}$$

This shows that

$$\partial_j^\perp = \frac{\partial}{\partial\zeta^j}$$

and hence that the bi-orthogonal frame is a coordinate frame.

(b) ⇒ (c)

By assumption there exist coordinates ζ_i such that (3) holds. From

$$\frac{\partial}{\partial\zeta^i}\mu_\theta = \frac{\partial\theta^k}{\partial\zeta^i}\partial_k\mu_\theta$$

and the definition (1) one obtains

$$\delta_{ij} = \langle\partial_i^\perp, \partial_j\rangle = \frac{\partial\theta^k}{\partial\zeta^i}\langle\partial_k\mu_\theta, \partial_j\rangle = \frac{\partial\theta^k}{\partial\zeta^i}g_{kj}.$$

This implies $g_{ij} = \delta_{il}\partial_j\zeta^l$. The condition for a function $\Phi(\theta)$ to exist such that $g_{ij} = \partial_i\partial_j\Phi$ is that $\partial_k g_{ij} = \partial_i g_{kj}$. This is satisfied because

$$\partial_k g_{ij} = \partial_k\delta_{il}\partial_j\zeta^l = \delta_{il}\partial_j\partial_k\zeta^l = \partial_j g_{ki}$$

and the tensor g is symmetric.

(c) ⇒ (a)

Let $\eta_i = \partial_i\Phi$. This is a coordinate system. It satisfies $\partial_i\eta_j = g_{ij}$. Hence, it is the dual coordinate system of $\{\theta_i\}_i$. □

2.2 Convex duality

Assume that the equivalent conditions of Proposition 1 are satisfied.

The metric tensor g is positive-definite by assumption. As a consequence, the potential $\Phi(\theta)$ is strictly convex on its domain of definition $\mathbb{D}$. The convexity of $\Phi(\theta)$ implies convex duality of the pair of coordinate systems θ^i, η_j. The meaning of this statement is explained in what follows.

The Legendre transform/convex conjugate Φ^* of the potential Φ is defined by

$$\Phi^*(\eta) = \sup\{\eta_i\theta^i - \Phi(\theta) : \theta \in \mathbb{D}\}. \tag{5}$$

Note that in the Physics literature a different sign convention is followed.

The domain of definition $\mathbb{D}^*$ is the set of covectors η for which there exists a vector θ in $\mathbb{D}$ such that

$$\eta_i = \frac{\partial\Phi}{\partial\theta^i} \tag{6}$$

holds for all i. If this relation is satisfied then one has

$$\Phi^*(\eta) = \eta_i\theta^i - \Phi(\theta). \tag{7}$$

By assumption the potential $\Phi(\theta)$ is a smooth function. This implies that the domain $\mathbb{D}^*$ is convex and that the dual potential Φ^* is a strictly convex function as well.

The inverse transformation reads

$$\Phi(\theta) = \sup_{\eta}\{\eta_i\theta^i - \Phi^*(\eta) : \eta \in \mathbb{D}^*\}. \tag{8}$$

The maximum is reached for

$$\theta^i = \frac{\partial\Phi^*}{\partial\eta_i}.$$

Take a further derivative to find

$$\frac{\partial\theta^i}{\partial\eta_j} = \frac{\partial^2\Phi^*}{\partial\eta_i\partial\eta_j}.$$

Finally, from the chain rule

$$g^i_k = \frac{\partial\eta_i}{\partial\theta^j}\,\frac{\partial\theta^j}{\partial\eta_k}$$

and the definition of dual coordinates one obtains

$$\frac{\partial\theta^i}{\partial\eta_j} = g^{ij}(\eta) = \frac{\partial^2\Phi^*}{\partial\eta_i\partial\eta_j}.$$

2.3 Dual geometries

The intention of the present section is to clarify the link that exists between convex duality and duality of geometric connections.

If a connection is flat then parallel transport along a curve does not depend on the chosen curve. This property can be verified easily if a coordinate system is available for which all connection coefficients $\Gamma^k_{ij}(\theta)$ vanish identically. If it exists then it is called an *affine coordinate system*. However, such a coordinate system is not always available. The connection can be flat even when the connection coefficients $\Gamma^k_{ij}(\theta)$ do not vanish in the given coordinate system.

Given a connection with coefficients $\Gamma^k_{ij}(\theta)$ the dual connection has connection coefficients $\Gamma^{*k}_{ij}(\theta)$ satisfying

$$\frac{\partial}{\partial\theta^k}g_{ij} = g_{jm}\Gamma^m_{ki} + g_{im}\Gamma^{*m}_{kj}. \tag{9}$$

The dual connection does always exist. Indeed, the above expression can be taken as the defining relation. Given the metric g, the metric connection, also called the Levi–Civita connection, is self-dual. It is the unique solution of the set of equations

$$\frac{\partial}{\partial\theta^k}g_{ij} = g_{jm}\Gamma^m_{ki} + g_{im}\Gamma^m_{kj}.$$

It is straightforward to verify that this solution is given by the well-known expression for the coefficients of the metric connection

$$\Gamma^m_{ki} = \frac{1}{2}g^{mn}\left[\frac{\partial}{\partial\theta^k}g_{ni} + \frac{\partial}{\partial\theta^i}g_{kn} - \frac{\partial}{\partial\theta^n}g_{ki}\right].$$

Information Geometry takes advantage of the freedom of working with other connections than the metric one.

Assume now that the coordinate system θ^i is affine and that it has η_j as a dual coordinate system. Then the $\Gamma^m_{ki}(\theta)$ vanish and (9) simplifies to

$$g_{im}\Gamma^{*m}_{kj}(\theta) = \frac{\partial}{\partial\theta^k}g_{ij} = \frac{\partial}{\partial\theta^k}\frac{\partial}{\partial\theta^j}\frac{\partial}{\partial\theta^i}\Phi(\theta).$$

This can be written as

$$\Gamma^{*i}_{jk}(\theta) = \frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\theta^j}\frac{\partial}{\partial\theta^k}\Phi(\theta) = \partial^i\partial_j\partial_k\Phi(\theta). \tag{10}$$

One concludes that in this case the connection coefficients of the dual geometry can be expressed in terms of the third order derivatives of the potential $\Phi(\theta)$.

The following result covers part of Theorem 3.6 of [3] and is a corner stone of the dually flat geometry of Amari–Nagaoka.

Theorem 2 *Assume that the equivalent conditions of Proposition 1 are satisfied. Assume in addition that the coordinates θ^i are affine for a given connection. Then the bi-orthogonal coordinates ζ^j are affine coordinates for the dual connection.*

Proof Let us first show that the covariant derivatives $\nabla^*_{\partial_i^\perp}\partial_j^\perp$ vanish for the dual connection ∇^*. Next it is noted that the coordinates ζ^j are affine coordinates for ∇^*. This leads to the conclusion that in the bi-orthogonal basis of the ζ^j the connection coefficients do vanish.

One calculates

$$\begin{aligned}\nabla^*_{\partial_i^\perp}\partial_j^\perp &= g^{ik}\nabla^*_{\partial_k} g^{jl}\partial_l \\ &= g^{ik}\left(g^{jl}\nabla^*_{\partial_k}\partial_l + \left[\frac{\partial}{\partial\theta^k}g^{jl}\right]\partial_l\right) \\ &= g^{ik}g^{jl}\Gamma^{*m}_{kl}\partial_m + g^{ik}\left[\frac{\partial}{\partial\theta^k}g^{jl}\right]\partial_l.\end{aligned}$$

With the help of

$$\Gamma^{*m}_{kl} = g^{mn}\partial_k g_{nl}$$

this becomes

$$\begin{aligned}\nabla^*_{\partial_i^\perp}\partial_j^\perp &= g^{ik}g^{jl}g^{mn}(\partial_k g_{nl})\partial_m + g^{ik}(\partial_k g^{jl})\,\partial_l \\ &= 0.\end{aligned}$$

This shows that the connection coefficients of the dual connection vanish in the bi-orthogonal basis.

The ζ^i are the canonical coordinates of the basis vectors $\partial_i^\perp$. Hence, they are affine coordinates. □

3 A manifold of probability measures

3.1 Tangent vectors

The manifold $\mathbb{M}$ considered from now on is a set of probability measures μ_θ labeled with a vector θ of parameters belonging to an open convex domain $\mathbb{D}$ of $\mathbb{R}^n$. The quantum case is mentioned only sporadically. The probability measures μ_θ are the states of the model. They are identified with linear functionals of the random variables. The derivatives $\partial_i\mu_\theta$ are linear functionals as well. Derivatives are the result of a limiting procedure. Hence, they depend on the chosen topology, *in casu* the topology on the space of linear functions defined on the algebra $\mathcal{A}$ of random variables. If the sample space $\mathcal{X}$ is a finite set with n elements then the topology is obvious. The set of probability measures is a simplex in $\mathbb{R}^n$. The topology is that of the Euclidean metric.

In the general case several choices are possible.

Ay et al. [4] consider the manifold $\mathbb{M}$ as a subset of the space $\mathcal{S}(\mathcal{X})$ of finite signed measures on the sample space $\mathcal{X}$ equipped with the total variation norm. Given a non-negative measure ν in $\mathcal{S}(\mathcal{X})$ all measures μ absolutely continuous w.r.t. ν belong to $\mathcal{S}(\mathcal{X})$. They are of the form $\mathrm{d}\mu(x) = f(x)\mathrm{d}\nu(x)$ with Radon-Nikodym derivative f in the space $\mathscr{L}^1(\mathcal{X}, \nu)$ of ν-integrable real-valued functions. The norm of total variation coincides with the $\mathscr{L}^1$-norm of the Radon-Nikodym derivative.

Alternatively, one can require that the points in the manifold $\mathbb{M}$ are normal states on the von Neumann algebra (W^*-algebra) $\mathscr{L}^\infty(\mathcal{X}, \nu, \mathbb{C})$ of essentially bounded complex functions w.r.t. some reference measure ν on the sample space. The space $\mathscr{L}^\infty(\mathcal{X}, \nu, \mathbb{C})$ is the dual Banach space of the space of complex-valued integrable functions $\mathscr{L}^1(\mathcal{X}, \nu, \mathbb{C})$. Any state ω on $\mathscr{L}^\infty(\mathcal{X}, \nu, \mathbb{C})$ belongs to the dual of $\mathscr{L}^\infty(\mathcal{X}, \nu, \mathbb{C})$. It is normal if and only if it belongs to the predual $\mathscr{L}^1(\mathcal{X}, \nu, \mathbb{C})$. See Section 2.4.3. of [6]. The attribute 'normal' for states is the equivalent of σ-additivity for probability measures. The advantage of this approach is that unification with the formalism of quantum statistical manifolds is possible. See [10, 20, 21].

In the approach of Pistone-Sempi [25, 26] the manifold $\mathbb{M}$ is restricted to consist of equivalent probability measures. It is then possible to construct an exponential map, i.e. to obtain locally a one-to-one correspondence between states of the manifold $\mathbb{M}$ of all μ-equivalent measures and vectors in the tangent plane $T_\mu\mathbb{M}$. In this way $\mathbb{M}$ becomes a Banach manifold, i.e. it looks locally like the Banach space $T_\mu\mathbb{M}$, which is in fact a Hilbert space.

The present paper deals only with finitely parameterised manifolds

$$\mathbb{M} = \{\mu_\theta : \theta \in \mathbb{D} \subset \mathbb{R}^n\}.$$

In this context the embedding Banach space $\mathcal{B}$ is the space of signed finite measures with the total variation norm. See Chapter 3 of [4]. The norm is in the first place important for a rigorous definition of the partial derivatives ∂_i.

3.2 Riemannian metric

Let $\mathbb{M}$ be a Riemannian manifold of probability measures μ_θ parameterized with a vector θ that belongs to an open convex domain $\mathbb{D}$ of $\mathbb{R}^n$. The metric tensor in the tangent plane $T_\theta\mathbb{M}$ is denoted $g(\theta)$ as usual.

In the standard situation the probability measure μ_θ is determined by a pdf p^θ and a background measure $\mathrm{d}x$. From the integrability of the pdf it follows that the probability measure μ_θ defines a normal state on the von Neumann algebra $\mathscr{L}^\infty(\mathcal{X}, \mathrm{d}x, \mathbb{C})$. The normal state is denoted μ_θ and is given by

$$\mu_\theta(X) = \mathbb{E}_\theta X$$

for any essentially bounded function $X(x)$.

A short calculation gives

$$\partial_i \mu_\theta(X) = \partial_i \mathbb{E}_\theta X$$

$$= \int_{\mathcal{X}} X(x)\partial_i p^\theta \mathrm{d}x$$
$$= \langle\langle \partial_i \log p^\theta, X\rangle\rangle_\theta \tag{11}$$

with

$$\langle\langle X, Y\rangle\rangle_\theta = \mathbb{E}_\theta(X - \mathbb{E}_\theta X)(Y - \mathbb{E}_\theta Y) = \mathbb{E}_\theta XY - \mathbb{E}_\theta X\mathbb{E}_\theta Y.$$

Note that the score variables $\partial_i \log p^\theta$ have vanishing expectation.

The Fisher–Rao metric at the point μ_θ is given by

$$g_{ij}(\theta) = \mathbb{E}_\theta[\partial_i \log p^\theta][\partial_j \log p^\theta] = \langle\langle \partial_i \log p^\theta, \partial_j \log p^\theta\rangle\rangle_\theta. \tag{12}$$

The evaluation of the score variables is simplified in the case of a model belonging to an exponential family. Mathematical aspects of this case are discussed in the next section. However, before doing so alternatives are discussed in the next subsection.

3.3 Representation of tangent vectors

In Section 2.5 of [3] the e- and m- representations of a tangent vector are discussed. A partially deviating treatment follows now.

Expression (11) writes the tangent vector ∂_i as the covariance of the random variable x with the score, which is the gradient of the logarithm of the probability density. This is called the *logarithmic representation* of the tangent vector. This logarithmic representation is particularly useful in the case of an exponential family in its canonical form because in that case $\log p^\theta$ is a simple expression.

In the case of a statistical model which does not belong to the exponential family other representations may be convenient. They involve so called *escort expectations*. Examples of such models are found in the deformed exponential families, which are discussed in Sect. 5.

4 Exponential families

The bi-orthogonal Hessian geometry is realised by the Boltzmann–Gibbs distribution, known to statisticians as a model belonging to an exponential family.

4.1 Definition and score variables

A parameterized family of probability measures μ_θ belonging to an exponential family is defined by a pdf p^θ of the form

$$p^\theta(x) = \exp(\theta^i F_i(x) - \Phi(\theta)). \tag{13}$$

The probability distribution of the random variables F_i can be used to estimate the parameters θ^i. The function $\Phi(\theta)$ is a subtractive normalization function. It is given by

$$\Phi(\theta) = \log \int \exp(\theta^i F_i(x))\, \mathrm{d}x,$$

where dx is any given reference measure. The integral is assumed to converge for all θ in a convex domain $\mathbb{D} \subset \mathbb{R}^n$ and to result in a smooth strictly convex function. If the variables F_i are bounded functions then the function $\Phi(\theta)$ exists and is convex on any compact convex domain. In the general case a careful choice of domain has to be made.

By taking derivatives of the normalization condition

$$1 = \int \mathrm{d}x\; p^\theta(x)$$

one obtains

$$\begin{aligned} 0 &= \int \mathrm{d}x\; \frac{\partial}{\partial\theta^i} p^\theta(x) \\ &= \int \mathrm{d}x\; p^\theta(x) \left[F_i(x) - \frac{\partial}{\partial\theta^i}\Phi(\theta) \right]. \end{aligned}$$

This can be written as

$$\frac{\partial\Phi}{\partial\theta^i} = \mathbb{E}_\theta F_i. \tag{14}$$

A pdf of this form facilitates the calculation of the score variables $\partial_i \log p^\theta$. One finds

$$\partial_i \log p^\theta = F_i - \partial_i \Phi(\theta).$$

The Fisher information matrix (12) becomes

$$g_{ij}(\theta) = \langle\langle F_i, F_j \rangle\rangle_\theta.$$

On the other hand, a short calculation gives

$$\begin{aligned} \frac{\partial^2\Phi}{\partial\theta^i\partial\theta^j} &= \frac{\partial}{\partial\theta^i}\mathbb{E}_\theta F_j \\ &= \mathbb{E}_\theta F_i F_j - \mathbb{E}_\theta F_i\, \mathbb{E}_\theta F_j \\ &= \langle\langle F_i, F_j \rangle\rangle_\theta. \end{aligned}$$

One concludes that the metric tensor is the Hessian of the potential $\Phi(\theta)$ and the coordinates η_i defined by

$$\eta_i = \frac{\partial \Phi}{\partial \theta^i}$$

are dual coordinates of the θ^i w.r.t. the Fisher metric g. In combination with (14) this gives

$$\eta_i = \mathbb{E}_\theta F_i.$$

It is well-known that the dual potential $\Phi^*(\eta)$ can be expressed in terms of the entropy function $S(p)$ of Boltzmann–Gibbs-Shannon, which is defined for any pdf p by

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

Indeed, one has

$$\begin{aligned} S(p^\theta) &= -\mathbb{E}_\theta \left(\theta^i F_i(x) - \Phi(\theta) \right) \\ &= \Phi(\theta) - \theta^i \eta_i \\ &= -\Phi^*(\eta). \end{aligned}$$

4.2 Gâteaux derivatives

Let us now show that under acceptable conditions the derivation operators ∂_i are Gâteaux derivatives for the $\mathcal{L}^1$-norm on the manifold $\mathbb{M}$.

One calculates

$$\begin{aligned} &|| \ p^{\theta+\epsilon g_i} - p^\theta - \epsilon \partial_i p^\theta \ ||_1 \\ &= || \ p^\theta \left[\exp(\epsilon F_i - \Phi(\theta + \epsilon g_i) + \Phi(\theta)) - 1 - \epsilon(F_i - \partial_i \Phi(\theta) \right] ||_1 \\ &= \mathbb{E}_\theta |e^{\epsilon G_i} - 1 - \epsilon G_i + \Phi(\theta + \epsilon g_i) - \Phi(\theta) - \epsilon \partial_i \Phi(\theta)| \\ &\leq \mathbb{E}_\theta |e^{\epsilon G_i} - 1 - \epsilon G_i| + \mathrm{o}(\epsilon) \end{aligned}$$

with

$$G_i = F_i - \frac{1}{\epsilon}(\Phi(\theta + \epsilon g_i) - \Phi(\theta)).$$

It is clear that if F_i is a bounded variable then the above expression tends to 0 faster than ϵ as ϵ tends to zero. In that case the derivative ∂_i is a Gâteaux derivative in $\mathcal{L}^1$-norm. This result is not completely satisfactory because in many applications the variables F_i are unbounded. The situation is similar for quantum statistical models. Boundedness of the operators replacing the variables F_i facilitates the proof that the basis vectors ∂_i of the tangent bundle are Gâteaux derivatives. See for instance Proposition 14 of [21].

4.3 A variational principle

Statistical models are quite often specified by either minimizing a divergence function/relative entropy or maximizing a linearly constrained entropy function. If this is the case then the model is said to be determined by a variational principle.

The Legendre transform (5) contains already a variational principle: It chooses from the model family $\{p^\theta : \theta \in \mathbb{D}\}$ that probability measure that maximizes the quantity $\eta_i\theta^i - \Phi(\theta)$. It is then obvious to extend this maximization procedure to a larger set of probability distributions. The way to do that is by assuming that there exists a vector of functions F_i such that

$$\eta_i(\theta) = \mathbb{E}_\theta F_i = \int F_i(x)\, p^\theta(x)\mathrm{d}x. \tag{15}$$

Then (6) becomes

$$\eta_i(\theta) = \frac{\partial\Phi}{\partial\theta^i} = \mathbb{E}_\theta F_i.$$

The convex dual of the potential Φ satisfies

$$\Phi(\theta) + \Phi^*(\theta^*) \geq \theta^i \mathbb{E}_\theta F_i, \qquad \theta \in D, \tag{16}$$

with equality when $\theta^*_i = \eta_i$. The second assumption behind the variational principle is that there exists a function $\Psi(p)$, defined on some domain $\mathrm{dom}(\Psi)$ of pdfs that includes all p^θ, such that (16) generalizes to

$$\Phi(\theta) + \Psi(p) \geq \theta^i \mathbb{E}_p F_i, \qquad p \in \mathrm{dom}(\Psi), \theta \in D, \tag{17}$$

with equality if and only if $p = p^\theta$. Note that $\Psi(p^\theta)$ necessarily coincides with $\Phi^*(\eta(\theta))$.

The above statement implies the *variational principle*

$$\Phi(\theta) = \sup\{\theta^k \mathbb{E}_p F_k - \Psi(p) : p \in \mathrm{dom}(\Psi)\}, \qquad \theta \in \mathbb{D}. \tag{18}$$

The maximum in the r.h.s. is reached for $p = p^\theta$ with p_θ given by (13). The parameters θ_i can be interpreted as Lagrange parameters used to enforce given values η_k for the expectations $\mathbb{E}_p F_k$ of the variables F_k.

Note that in the Physics literature different sign conventions are followed for the Legendre transformation. In [17], Section 3.7, it is stated that

$$S(p) - \theta^k \mathbb{E}_p H_k \tag{19}$$

is maximal when the probability distribution p equals the model distribution p^θ. Comparison with (18) shows that minus the function $\Psi(p)$ corresponds with the entropy function $S(p)$. The F_k are replaced by Hamiltonians H_k with an extra minus sign

because Hamiltonians are interpreted as energies and are expected to be bounded from below. For the expectation values the symbol $\mathbb{E}$ is replaced by brackets $\langle\cdot\rangle$. The maximization of expression (19) is referred to as the *maximal entropy principle* [13, 14].

From (18) it follows that the *divergence* $D(p \mid\mid p^\theta)$ defined by

$$D(p \mid\mid p^\theta) = \theta^i \mathbb{E}_\theta F_i - \Psi(p^\theta) - \Big(\theta^i \mathbb{E}_p F_i - \Psi(p)\Big) \tag{20}$$

is non-negative and vanishes if and only if $p = p^\theta$. In the Physics Literature the divergence function is called the *relative entropy*.

Note that the divergence function (20) can be written as

$$\begin{aligned} D(p \mid\mid p^\theta) &= \theta^i \mathbb{E}_\theta F_i + S(p^\theta) - \Big(\theta^i \mathbb{E}_p F_i + S(p)\Big) \\ &= \int \mathrm{d}x \left[f(p(x)) - f(p^\theta(x)) - (p(x) - p^\theta(x)) f'(p^\theta(x)) \right], \end{aligned}$$

with $f(u) = u \log u$. This shows that the divergence is of the Bregman type [7].

4.4 Dual geometries

Consider now the connection, called the e-connection, in which the paths p_t defined by

$$p_t = p^{\theta_t} \quad \text{with} \quad \theta_t = (1-t)\theta_{(1)} + t\theta_{(2)}, \qquad \theta_{(1)}, \theta_{(2)} \in \mathbb{D},$$

are geodesics. The second derivatives w.r.t. t vanish, i.e. $\ddot{\theta}_t = 0$. Hence the θ^i are affine coordinates for this connection and the corresponding connection coefficients $\Gamma^k_{ij}(\theta)$ vanish.

The dual connection of the e-connection is called the m-connection. Geodesics for this dual geometry are the solutions of the set of equations

$$\ddot{\theta}^i + \Gamma^{*i}_{jk} \dot{\theta}^j \dot{\theta}^k = 0. \tag{21}$$

Proposition 3 *Fix two parameter sets $\theta_{(1)}$ and $\theta_{(2)}$ in the domain $\mathbb{D}$ of the model. Then the path*

$$t \in (0,1) \mapsto p_t = (1-t)p_1 + tp_2$$

with $p_\sigma = p^{\theta_{(\sigma)}}$, $\sigma = 1, 2$, is a geodesic for the dual geometry.

Proof The affine coordinates for the dual geometry are the coordinates ζ^i of the bi-orthogonal frame—see Theorem 2. During the proof of Proposition 1 it is shown that

the ζ^i coincide with the dual variables η_i. A short calculation now shows that along the path $t \mapsto p_t$ one has

$$\delta_{ij}\zeta^j = \eta_i = \mathbb{E}_{p_t} F_i = (1-t)\mathbb{E}_{(1)} F_i + t\mathbb{E}_{(2)} F_i .$$

Hence, along this path the coordinates ζ^i satisfy $\ddot{\zeta}^i = 0$, which is the geodesic equation (21) expressed in dual coordinates. One concludes that the path is a geodesic for the dual geometry. □

5 Deformed exponential model

5.1 Conformal Hessian metric

The exponential family of probability models renders the Fisher–Rao metric as a Hessian metric. Treating the natural parameter θ of the exponential family as an affine parameter underlies the duality with its expectation parameterization η. In general a Riemannian metric needs not to be Hessian, but in some applications the Riemannian metric may be conformally equivalent to a Hessian metric. It is called *conformal Hessian geometry*. It is an intermediary between the full-blown Riemannian geometry and the highly specialized Hessian geometry.

In the notation of the current paper, given a Hessian metric tensor $g(\theta)$ and a smooth strictly positive function $z(\theta)$, the conformal Hessian metric $\overline{g}_{ij}(\theta)$ is defined by

$$\overline{g}_{ij}(\theta) = \frac{1}{z(\theta)} g_{ij}(\theta).$$

The coordinate system $\eta_i(\theta)$ which is dual w.r.t. $g_{ij}(\theta)$ satisfies

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}(\theta) = z(\theta)\overline{g}_{ij}(\theta). \tag{22}$$

Hence, the definition (4) of a dual coordinate system is relaxed by allowing a proportionality factor $z(\theta)$. A conformal Hessian metric arises in some scenarios of *deformed* exponential families. It also arises from dual projectively flat manifolds [32].

5.2 Phi-deformation

Fix a strictly positive function ϕ defined on the open interval $(0, +\infty)$. It is used to define a deformed logarithm $\log_\phi$ by integration

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

The inverse function is denoted $\exp_\phi(u)$ and is referred to as the ϕ-deformed exponential function. The standard logarithmic and exponential functions are recovered when the function ϕ is the identity function, i.e. $\phi(u) = u$ for all u.

It is obvious to define the ϕ-deformed exponential family by replacing the exponential function in (13) by the deformed function $\exp_\phi$. This gives a pdf of the form

$$p^\theta(x) = \exp_\phi(\theta^i F_i(x) - \alpha(\theta)) \tag{23}$$

with subtractive normalization $\alpha(\theta)$. The domain $\mathbb{D}$ of allowed θ-values is now restricted by the requirement that a smooth function $\alpha(\theta)$ exists such that the normalization integral converges

$$\int \mathrm{d}x\, p^\theta(x) = 1.$$

By taking the derivative of the normalization condition one obtains

$$\begin{aligned} 0 &= \int \mathrm{d}x\, \frac{\partial}{\partial\theta^i} p^\theta(x) \\ &= \int \mathrm{d}x\, \phi(p^\theta(x)) \left[F_i(x) - \frac{\partial\alpha}{\partial\theta^k} \right]. \end{aligned}$$

This can be written as

$$\frac{\partial\alpha}{\partial\theta^k} = \tilde{\mathbb{E}}_\theta F_k \tag{24}$$

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

$$\tilde{\mathbb{E}}_\theta X = \frac{1}{z(\theta)} \int \mathrm{d}x\, \phi(p^\theta(x))\, X(x),$$

and with multiplicative normalization $z(\theta)$ given by

$$z(\theta) = \int \mathrm{d}x\, \phi(p^\theta(x)),$$

provided that these integrals converge.

Expression (11) generalizes to

$$\partial_i \mathbb{E}_\theta X = z(\theta) \langle\langle F_i, X \rangle\rangle_\theta^{\mathrm{esc}} \tag{25}$$

with the covariance $\langle\langle \cdot, \cdot \rangle\rangle_\theta^{\mathrm{esc}}$ defined by the escort probabilities

$$\langle\langle Y, X \rangle\rangle_\theta^{\mathrm{esc}} = \tilde{\mathbb{E}}_\theta (Y - \tilde{\mathbb{E}}_\theta Y)\, (X - \tilde{\mathbb{E}}_\theta X).$$

In the *linear growth case* [15, 19, 24] one chooses

$$\phi(u) = \frac{u}{1+u}.$$

In this specific case the existence of the subtractive normalization $\alpha(\theta)$ and of the escort probabilities has been studied by careful estimation of the growth rate of the deformed exponential function.

5.3 Deformed metric

In this class of models one can make several choices for the metric tensor $g(\theta)$, each of which can be considered to be a straightforward generalization of the Fisher information matrix. Based on (24) it is tempting to make the choice

$$\overline{g}_{ij}(\theta) = \frac{\partial}{\partial\theta^j}\tilde{\mathbb{E}}_\theta F_i = \frac{\partial^2\alpha}{\partial\theta^j\partial\theta^i}.$$

However, in general additional restrictions must be imposed to prove that the matrix $\overline{g}_{ij}(\theta)$ is positive-definite.

An alternative is given by

$$\tilde{g}_{ij}(\theta) = \int \mathrm{d}x\, \frac{1}{\phi(p^\theta(x))}\left[\frac{\partial}{\partial\theta^i}p^\theta(x)\right]\left[\frac{\partial}{\partial\theta^j}p^\theta(x)\right]. \tag{26}$$

It is straightforward to show that it is positive-definite. In combination with (23) this expression yields

$$\begin{aligned}\tilde{g}_{ij}(\theta) &= \int \mathrm{d}x\, \phi(p^\theta(x))\left[F_i(x) - \frac{\partial\alpha}{\partial\theta^i}\right]\left[F_j(x) - \frac{\partial\alpha}{\partial\theta^j}\right] \\ &= z(\theta)\langle\langle F_i, F_j\rangle\rangle_\theta^{\mathrm{esc}}.\end{aligned}$$

Note that

$$\frac{\partial}{\partial\theta^j}\mathbb{E}_\theta F_i = z(\theta)\langle\langle F_j, F_i\rangle\rangle_\theta^{\mathrm{esc}} = \tilde{g}_{ij}(\theta).$$

This shows that the coordinates η_i, defined by $\eta_i = \mathbb{E}_\theta F_i$, are dual coordinates w.r.t. $\tilde{g}_{ij}$. On the other hand, the metric $\hat{g}_{ij}(\theta)$ defined by

$$\hat{g}_{ij}(\theta) = \langle\langle F_i, F_j\rangle\rangle_\theta^{\mathrm{esc}}$$

is conformally equivalent with the Hessian metric $\tilde{g}_{ij}(\theta)$. The dual coordinate system $\eta(\theta)$ satisfies the relaxed condition (22) w.r.t. $\hat{g}_{ij}(\theta)$.

5.4 Deformed entropy

In the non-deformed case, corresponding with $\phi(u) = u$, the normalization function $\alpha(\theta)$ coincides with the potential $\Phi(\theta)$ the Hessian of which is the metric tensor g. In the general case the two functions differ. This raises the question whether one can derive an explicit expression for $\Phi(\theta)$. In principle the potential $\Phi(\theta)$ can be obtained by integrating the dual coordinates η_j. Instead, a proposal is made for the entropy function S, which is minus the Legendre transform of Φ.

Introduce a deformed entropy $S_\phi(p)$ of the pdf $p(x)$. The expression introduced in [16] can be written as

$$S_\phi(p) = -\int \mathrm{d}x \int^{p(x)} \mathrm{d}u \ \log_\phi(u) + \text{ constant} = -\int \mathrm{d}x \ U_\phi^*(p(x)) + \text{ constant}$$

where we use the notation U_ϕ, U_ϕ^* to denote the indefinite integral functions of $\exp_\phi$ and $\log_\phi$, respectively,

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

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

$$\begin{aligned} U_\phi &\xrightarrow{\prime} \exp_\phi \xrightarrow{\prime} \phi(\exp_\phi); \\ U_\phi^* &\xrightarrow{\prime} \log_\phi \xrightarrow{\prime} \frac{1}{\phi}. \end{aligned} \tag{27}$$

and U_ϕ and U_ϕ^* are a pair of strictly convex functions that are conjugate to one another, so $*$ is the convex "conjugate" operation. The use of the U, U^* notation here is in recognition of Eguchi's (independently discovered) U-model [11], which turns out to be identical to Naudts' ϕ-model.

This expression of ϕ-entropy clearly reduces to the Boltzmann–Gibbs-Shannon entropy (4.1) in the case that $\phi(u) = u$. Next, let

$$\Phi(\theta) = \theta^i \eta_i + S_\phi(p^\theta).$$

where $\eta_i = \mathbb{E}_\theta \, F_i$. Then one has

$$\begin{aligned} \frac{\partial \Phi}{\partial \theta^k} &= \eta_k + \theta^i \frac{\partial \eta_i}{\partial \theta^k} - \int \mathrm{d}x \ \log_\phi(p^\theta(x)) \, \frac{\partial}{\partial \theta^k} p^\theta(x) \\ &= \eta_k + \theta^i \frac{\partial \eta_i}{\partial \theta^k} - \int \mathrm{d}x \ \left[\theta^i F_i(x) - \alpha(\theta)\right] \frac{\partial}{\partial \theta^k} p^\theta(x) \\ &= \eta_k + \theta^i \frac{\partial \eta_i}{\partial \theta^k} - \theta^i \frac{\partial}{\partial \theta^k} \mathbb{E}_\theta F_i \\ &= \eta_k. \end{aligned}$$

This shows that $\Phi(\theta)$ as defined above is the potential we are looking for.

The Legendre transform $S_\phi(p^\theta)$ of $\Phi(\theta)$ can be extended to a variational principle for probability distributions. From

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\bigg|_{\lambda=0} S_\phi((1-\lambda)p^{(1)}+\lambda p^{(2)}) = -\int \mathrm{d}x\, \frac{(p^{(2)}(x)-p^{(1)}(x))^2}{\phi(p^{(1)}(x))}$$

and positivity of the function $\phi(u)$ one concludes that the entropy function $S_\phi(p)$ is strictly concave. It is then straightforward to verify that the variational principle

$$\Phi(\theta) = \sup_p\{\theta^k \mathbb{E}_p F_k - \Psi(p)\}.$$

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

5.5 Rho-tau deformation and gauge freedom in Legendre duality

The deformed logarithmic and exponential functions of the previous section are a pair of strictly increasing functions that are inverses of each other. Such pairs of functions occur naturally in the context of convex duality.

Start from a pair of strictly increasing functions $\rho(u)$ and $\tau(u)$ of a real variable u. Then there exists a strictly convex function $f(u)$ such that almost everywhere its derivative $f'(u)$ is given by $f'(u) = \tau(\rho^{-1}(u))$. This function $f(u)$ is given by the Stieltjes integral

$$f(u) = \int^u \tau(s)\,\mathrm{d}\rho(s).$$

One verifies that the second derivative $f''(u)$ is given by

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

It is strictly positive because both τ and ρ are strictly increasing functions. Hence $f(u)$ is a strictly convex function.

Consider now the convex dual $f^*(v)$ of the function $f(u)$. It is given by

$$f^*(v) = \sup\{vu - f(u) : u \in \mathrm{dom}(f)\}.$$

If $v = f'(u)$ for some point $u \in \mathrm{dom}(f)$ then one has the equality $f^*(v) = vu - f(u)$. The derivative of the latter expression reads

$$f^{*\prime}(v) = u + \big(v - f'(u)\big)\frac{\mathrm{d}u}{\mathrm{d}v}.$$

Replace in this expression v by $f'(u)$ to find that

$$f^{*\prime}(f'(u)) = u.$$

This shows that the derivatives $f'(u)$ and $f^{*\prime}(v)$ are each others inverse.

Let us verify this duality for the pair $\log_\phi(u), \exp_\phi(v)$ of deformed exponential / logarithmic functions. The condition $f'(u) = \exp_\phi(u)$ is realized on a subinterval of the domain of $\log_\phi(u)$ for instance with $\tau(u) = u$ and $\rho(u) = \log_\phi(u)$. The function $f(u)$ is then given by

$$\begin{aligned} f(u) &= \int^u v \ \mathrm{d}\log_\phi(v) \\ &= \int^{\exp_\phi u} \frac{v}{\phi(v)} \mathrm{d}v. \end{aligned}$$

Its convex dual is

$$f^*(v) = \sup_{u>0}\{vu - f(u)\}.$$

The maximum is reached for every value of $v > 1$. One finds

$$\begin{aligned} f^*(v) &= v\log_\phi(v) - f(\log_\phi(v)) \\ &= v\log_\phi(v) - \int^v s \,\mathrm{d}\log_\phi(s). \end{aligned}$$

By partial integration this becomes

$$f^*(v) = \int^v \log_\phi(s)\mathrm{d}s.$$

The latter implies $f^{*\prime}(v) = \log_\phi(v)$, which is indeed the inverse of $f'(u) = \exp_\phi(u)$. So f, f^* are nothing but U, U^* functions as defined earlier.

Conversely, given a pair ρ, τ of strictly increasing functions the second derivative of the strictly convex dual function f^* can be used to define almost everywhere in the domain $\mathrm{dom}(f^*)$ a strictly positive function $\phi(v)$ by

$$\phi(v) = \frac{1}{f^{*\prime\prime}(v)}.$$

Integration then yields

$$\log_\phi(v) = f^{*\prime}(v) - f^{*\prime}(1).$$

This implies

$$f'(u) = \exp_\phi(u - f^{*\prime}).$$

The convex duality $f \leftrightarrow f^*$ and the $\rho \leftrightarrow \tau$ duality coincide. Indeed, because $f'(u) = \tau(\rho^{-1}(u))$ is the inverse function of $f^{*\prime}(v)$ and the inverse function of

$u \mapsto \tau(\rho^{-1}(u))$ is $v \mapsto \rho(\tau^{-1}(v))$ the interchange of ρ and τ corresponds with the interchange of f and f^* provided that integration constants are carefully chosen. The results of Sect. 2.3 then show that in the case of a dual pair of affine coordinates θ^i and η_j also the dually flat geometries are interchanged. This can be made explicit by postulating connection coefficients which depend on the functions ρ and τ. See Sect. 5.9 below.

5.6 Extended metric and gauge freedom

Just as the metric tensor (26), proposed by Naudts [16], Zhang [33] proposed the following form of the metric tensor

$$\tilde{g}_{ij}(\theta) = \int \mathrm{d}x \left[\frac{\partial}{\partial\theta^i}\rho(p^\theta(x))\right]\left[\frac{\partial}{\partial\theta^j}\tau(p^\theta(x))\right] \tag{28}$$

and of the divergence that induces it

$$D_{\rho,\tau}(p \,||\, q) = \int_{\mathcal{X}} \mathrm{d}x \left[f \circ \rho(p(x)) + f^* \circ \tau(q(x)) - \rho(p(x))\tau(q(x))\right].$$

It is easy to see that the tensor $\tilde{g}_{ij}(\theta)$ is symmetric and positive-definite, and that

$$\frac{\partial}{\partial\theta^i_{(1)}}\frac{\partial}{\partial\theta^j_{(2)}} D_{\rho,\tau}(\theta_{(1)} \,||\, \theta_{(2)})\Big|_{\theta_{(1)}=\theta_{(2)}=\theta} = -\tilde{g}_{ij}(\theta)$$

The two expressions (26) and (28) for the metric tensor $\tilde{g}_{ij}$ coincide if one makes the identification

$$\phi(u) = \frac{1}{\rho'(u)\tau'(u)} \tag{29}$$

Note that

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

so exchanging ρ, τ is equivalent to exchanging the role of p, q. The symmetry between ρ and τ can be broken for instance by taking $\rho(u) = u$ and $\tau(u) = \log_\phi(u)$, a choice which guarantees that (29) is satisfied.

5.7 Rho-tau connections

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.

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. Fixing the gauge means choosing a deformed logarithm τ, keeping $\rho'\tau'$ fixed. This is what we call [23] the gauge freedom of the rho-tau formalism.

Given a pair of strictly increasing functions ρ and τ and a model p^θ, Zhang [33] introduced the following connections, which are induced from the divergence function (5.6)

$$\Gamma^{(\alpha)}_{ij,k} = \frac{1+\alpha}{2}\int_{\mathcal{X}} \mathrm{d}x\, \left[\partial_i\partial_j\rho(p^\theta(x))\right]\left[\partial_k\tau(p^\theta(x))\right] + \frac{1-\alpha}{2}\int_{\mathcal{X}} \mathrm{d}x\, \left[\partial_i\partial_j\tau(p^\theta(x))\right]\left[\partial_k\rho(p^\theta(x))\right]. \tag{30}$$

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

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

This shows that $\Gamma^{(-\alpha)}$ is the dual connection of $\Gamma^{(\alpha)}$.

Take for instance $\rho(u) = \tau(u) = 2\sqrt{u}$. Then the metric is undeformed, i.e. it is the Fisher information metric, because $\rho'(u)\tau'(u) = 1/u$. Due to the rho-tau symmetry, the connection coefficient $\Gamma^{(\alpha)}$ is independent of α and equals half the derivative of the Fisher information $g(\theta)$, which is the value of the Levi–Civita connection coefficient. On the other hand, take $\rho(u) = u$ and $\tau(u) = \log u$, then $\rho'(u)\tau'(u) = 1/u$ is still valid. However, the connection coefficients $\Gamma^{(\alpha)}$ now depend in general on α.

5.8 Affine coordinates

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

$$\int_{\mathcal{X}} \mathrm{d}x\, \left[\partial_i\partial_j\tau(p^\theta(x))\right]\left[\partial_k\rho(p^\theta(x))\right] = 0. \tag{31}$$

This is the case if the model pdf p^θ satisfies

$$\partial_i\partial_j\tau(p^\theta) = 0,$$

which translates to a set of non-linear equations

$$\partial_i\partial_j p^\theta + \frac{\tau''(p^\theta)}{\tau'(p^\theta)}[\partial_i p^\theta][\partial_j p^\theta] = 0.$$

So, when (31) holds, then the θ^i are affine coordinates for the connection $\Gamma^{(-1)}$ and the bi-orthogonal coordinates ζ^i are affine coordinates for the connection $\Gamma^{(1)}$. See Proposition 1 in Sect. 5 of [22].

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

$$\int_{\mathcal{X}} \mathrm{d}x \left[\partial_i \partial_j \rho(p^\theta(x))\right]\left[\partial_k \tau(p^\theta(x))\right] = 0, \tag{32}$$

in which case the dual coordinates $\eta_i \equiv \zeta^i$ are affine for the connection $\Gamma^{(-1)}$ and the θ^i are affine for the connection $\Gamma^{(1)}$.

5.9 Fixing the gauge

Note that the exchange of ρ, τ leads to a switch between the dual connections $\Gamma^{(1)}$, $\Gamma^{(-1)}$ though the metric remains unaffected. In [34], $\rho \leftrightarrow \tau$ is called "representation duality" since this encodes the duality of Legendre conjugation, while swapping the arguments of the (ρ, τ) divergence function $D^{(\alpha)}(p, p')$ is called "reference duality" since it encodes their non-symmetric status, the latter leading to $\alpha \leftrightarrow -\alpha$ in affine connections (see our companion paper for details). In the special case when ρ, τ are conjugated power functions (with exponent β and $-\beta$), the two dualities are intermingled as "reference-representation biduality" [33], concisely reflected by the alpha-connection with $\alpha\beta$ being the parameter to index the family of connections.

From the infinitely possible combinations of ρ and τ, producing the same function ϕ as defined by (26), two stand out:

(i) $\rho(u) = \log_\phi(u)$; This implies $\rho'(u) = 1/\phi(u)$ and $\tau'(u) = 1$. We call this *Type I gauge*.
(ii) $\tau(u) = \exp(\log_\phi(u))$; This implies $\tau'(u) = \tau(u)/\phi(u)$ and $\rho'(u) = 1/\tau(u)$. We call this *Type II gauge*.

In the type I gauge the rho-tau divergence $D_{\rho,\tau}(p \;||\; q)$ coincides with the phi-deformed divergence introduced in [16]. This choice of gauge is appropriate when the model p^θ belongs to a phi-deformed exponential family. In the type II gauge the rho-tau entropy

$$S_{\rho,\tau}(p) = -\int \mathrm{d}x \int^{p(x)} \tau(u)\mathrm{d}\rho(u)$$

is constant.

Given a ϕ-exponential family a pair of strictly increasing functions ρ, τ satisfying (29) can be chosen in many ways. With each choice corresponds a dual pair of connections. From the definition (30) of the connection coefficients $\Gamma^{(\alpha)}$ it is clear that if the function $\rho(u)$ is linear then the coordinates θ_i are affine for $\Gamma^{(1)}$ and the pair of connections $\Gamma^{(1)}$, $\Gamma^{(-1)}$ is dually flat. The same conclusion is reached when the function $\tau(u)$ is linear because then the dual coefficients η_i are affine.

Table 1 Examples of ρ, τ combinations

$\rho(u)$	$\tau(u)$	$(\rho'\tau')(u)$	$f(u)$	$f^*(u)$
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	$\log_\phi(u)$	$\frac{1}{\phi(u)}$	$u\log_\phi(u) - \int_1^u \frac{v}{\phi(v)}\mathrm{d}v$	$\int_1^{\exp_\phi(u)} \frac{v}{\phi(v)}\mathrm{d}v$

From another perspective, if a pair of strictly increasing functions ρ, τ is given then the relation (29) can be used to find a function $\phi(u)$ that one can use to construct a ϕ-exponential family equipped with a dual geometry.

6 Summary

This paper reviews the role of convex duality in Information Geometry. The notion of dual coordinates is clarified in the context of a Riemannian manifold. The transition to dual coordinates is necessarily a Legendre transformation. The connection between dual coordinates and Amari's dually flat geometry is recalled.

The notion of an exponential family is reviewed. The original theory deals with a finite space of events, the more recent generalizations to general event spaces are shortly mentioned. Under mild conditions the partial derivatives ∂_i, basis vectors of the tangent bundle, are Gâteaux derivatives in the $\mathscr{L}_1$-norm. The probability distributions of an exponential family satisfy a variational principle. The dually-flat geometry consists of the e- and m-connections.

Deformed exponential families are briefly reviewed. They involve a metric which generalizes the Fisher–Rao metric and still exhibit a dually flat pair of connections. The rho-tau formalism has an additional degree of freedom which can be used to modify the connections in a controlled way.

Data Availability Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

Declarations

Conflict of interest Jun Zhang is a Co-Editor of the journal. Jan Naudts is a board member of the journal. Both were not involved in the peer review or handling of the manuscript. On behalf of all authors, the corresponding author states that there is no other potential conflict of interest to declare.

References

1. Amari, S.: Differential geometry of curved exponential families–curvatures and information loss. Ann Stat. **10**(2), 357–385 (1982)
2. Amari, S.: Differential-Geometrical Methods in Statistics, Lecture Notes in Statistics, vol. 28. Springer, New York, Berlin (1985)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191. Oxford University Press, Oxford (2000)
4. Ay, N., Jost, J., Vân Lê, H., Schwachhöfer, L.: Inform. Geom. Springer, Berlin (2017)
5. Ballian, R.: François Massieu et les potentiels thermodynamiques, Inst. France Acad. Sci. (2015)
6. Bratteli, O., Robinson, D.W.: Operator algebras and Quantum Statistical Mechanics I. Springer, Berlin (1979)
7. Bregman, L.M.: The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. USSR Comp. Math. Math. Phys. **7**, 200–217 (1967)
8. Callen, H.B.: Thermodynamics and an introduction to thermostatistics, 2nd edn. Wiley, Hoboken (1985)
9. Ciaglia, F.M., Di Cosmo, F., González-Bravo, L.: Can Čencov meet Petz, In: Nielsen, F., Barbaresco, F. (eds.), Geometric Science of Information, LNCS 14072, Springer, pp. 363–371 (2023)
10. Ciaglia, F.M., Di Nocera, F., Jost, J., Schwachhöfer, L.: Parametric models and information geometry on W^*-algebras. Info. Geo. (2023). https://doi.org/10.1007/s41884-022-00094-6
11. Eguchi, S.: Information geometry and statistical pattern recognition, Sugaku Expositions. Am. Math. Soc. **19**, 197–216 (2006)
12. Gibbs, J.W.: Elementary principles in statistical mechanics. Dover, New York (1960). (**Reprint**)
13. Jaynes, E.T.: Information theory and statistical mechanics. II. Phys. Rev. **108**, 171–190 (1957)
14. Jaynes, E.T.: Papers on probability, statistics and statistical physics, ed. R.D. Rosenkrantz, Kluwer (1989)
15. 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, Springer, pp. 239–246 (2017)
16. Naudts, J.: Estimators, escort probabilities, and phi-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
17. Naudts, J.: Generalised Thermostatistics. Springer, Berlin (2011)
18. Naudts, J.: Quantum Statistical Manifolds. Entropy **20**, 472 (2018). (**correction Entropy 20, 796 (2018)**)
19. Naudts, J.: Quantum statistical manifold: the linear growth case. Rep. Math. Phys. **84**, 151–169 (2019)
20. Naudts, J.: Exponential arcs in the manifold of vector states on a σ-finite von Neumann algebra. Inf. Geom. **5**, 1–30 (2022)
21. Naudts, J.: Exponential arcs in manifolds of quantum states. Front. Phys. **11**, 1042257 (2023). https://doi.org/10.3389/fphy.2023.1042257
22. Naudts, J., Zhang, J.: Information geometry under monotone embedding. Part II: Geometry. In: Nielsen, F., Barbaresco, F. (eds.), Geometric Science of Information, GSI 2017 LNCS proceedings, Springer, pp. 215–222 (2017)
23. Naudts, J., Zhang, J.: Rho-tau embedding and gauge freedom in information geometry. Inform. Geom. **1**(1), 79–115 (2018)
24. Newton, N.J.: An infinite-dimensional statistical manifold modeled on Hilbert space. J. Funct. Anal. **263**, 1661–1681 (2012)
25. Pistone, G.: Nonparametric Information Geometry. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information, pp. 5–36. Springer, Berlin (2013)
26. Pistone, G., Sempi, C.: An infinite-dimensional structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**, 1543–1561 (1995)
27. Ruelle, D.: Statistical Mechanics, Rigorous Results. W.A. Benjamin Inc, New York (1969)
28. Ruppeiner, G.: Thermodynamics: a Riemannian geometric model. Phys. Rev. A **20**, 1608–1613 (1979)
29. Tsallis, C.: Possible generalization of Boltzmann–Gibbs statistics. J. Stat. Phys. **52**, 479–487 (1988)
30. Tsallis, C.: What are the numbers that experiments provide? Quimica Nova **17**, 468 (1994)
31. Weinhold, F.: Metric geometry of equilibrium thermodynamics. J. Chem. Phys. **63**, 2479–2483 (1975)
32. Wong, T.K.L., Yang, J.: Logarithmic divergences: geometry and interpretation of curvature. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information, pp. 413–422. Springer, Berlin (2019)

33. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
34. Zhang, J.: Nonparametric information geometry: from divergence function to referential-representational biduality on statistical manifolds. Entropy **15**, 5384–5418 (2013)