entropy

MDPI

Article

Lagrangian Submanifolds of Symplectic Structures Induced by Divergence Functions

Marco Favretti

Dipartimento di Matematica Tullio Levi-Civita, Università degli Studi di Padova, 35121 Padova, Italy; favretti@math.unipd.it

Received: 29 July 2020; Accepted: 2 September 2020; Published: 3 September 2020

Abstract: Divergence functions play a relevant role in Information Geometry as they allow for the introduction of a Riemannian metric and a dual connection structure on a finite dimensional manifold of probability distributions. They also allow to define, in a canonical way, a symplectic structure on the square of the above manifold of probability distributions, a property that has received less attention in the literature until recent contributions. In this paper, we hint at a possible application: we study Lagrangian submanifolds of this symplectic structure and show that they are useful for describing the manifold of solutions of the Maximum Entropy principle.

Keywords: canonical divergence; Lagrangian submanifolds; Morse family; constrained optimization; geometric phase transitions

1. Introduction

Information Geometry [1,2] provides a sound and fruitful framework for interpreting statistics using classical differential geometry notions [3]. A principal object in Information Geometry is the notion of contrast or divergence function, which (informally speaking) measures the degree of separation between two probability distributions [4–6]. The main thrust of divergence functions is that they allow to define a Riemannian structure on a finite dimensional submanifold M of probability distributions endowed with a dual coordinate system, with far reaching implications. A less-studied spin off of contrast function is the possibility of introducing a symplectic structure on the *square* of M by the pull-back of the canonical symplectic structure defined on the cotangent bundle T^*M. This procedure was introduced in 1995 in the pioneering paper [7], suggesting that symplectic geometry may have a natural role to play in statistics. In recent times there has been a renewed interest in possible applications of the symplectic structures introduced, as in [7] for example, to studying the analogies with the discrete Lagrangian mechanics (see in [8]) or the relations with completely integrable systems of Hamiltonian mechanics (see in [9,10]).

In this paper, we try to look at a possible role for Lagrangian submanifolds of the above-discussed symplectic structure on M^2 in the case that M is an *exponential family* $M(h,k)$. Exponential families are prototypical examples of finite dimensional manifolds admitting a dually flat canonical structure defined by the *canonical divergence*, and they play a relevant role in information geometry and statistics [1,2]. For our argument, their importance is due to the fact that they represent the manifold of solutions of the variational problem associated to the Maximum Entropy Principle (MEP) with linear constraints ([11,12]). In some applications to statistical mechanics, e.g., in the descriptions of phase transitions in Ising spin systems, MEP with *nonlinear* constraints is considered, see, e.g., in [13–15]. In this case, the set of possible solutions has a richer structure, which is well captured by a Lagrangian submanifold of $T^*M(h,k)$. In this work, we are concerned with the Lagrangian submanifolds defined in the square of $M(h,k)$ via the canonical pull-back hinted at above.

The structure of the paper is as follows. In Section 2, we recall the needed tools of Symplectic Geometry, and in Section 2.1 we review the canonical pull-back construction via divergence function construction exposed in [7]. In Section 3, we consider the special case of exponential families associated with MEP with nonlinear constraints.

2. Synopsis of Symplectic Geometry

We briefly recall the basic facts of symplectic geometry that are necessary for introducing our argument referring to classical textbooks for the proof of the results. A symplectic manifold (M, ω) is a smooth even-dimensional manifold M equipped with a non-degenerate, closed two-form ω ($d\omega = 0$, where d is the external derivation operator). A submanifold L of M is a *Lagrangian* submanifold if $2 \dim L = \dim M$ and the two-form restricted to L is vanishing, $\omega_{|L} = 0$. A prototypical example of symplectic manifold is the cotangent bundle T^*S of a manifold S. If $x = (x^1, \ldots, x^n)$ are local coordinates on S, and (x, λ) are local coordinates on T^*S, then the Liouville one-form θ_c on T^*S has the local expression $\theta_c = \lambda_i dx^i$ (summation over repeated indices is understood) and the symplectic two form is

$$\omega = d\theta_c = d\lambda_i \wedge dx^i. \tag{1}$$

A classical theorem of Darboux says that every symplectic manifold (M, ω) admits an atlas of local coordinates (x, λ) such that locally ω has the representation (1). A relevant example of Lagrangian submanifold of T^*S is the graph of the differential of a function $g : S \to \mathbb{R}$, that is,

$$L_g = \{(x, \lambda(x)) \in T^*S \ : \ \lambda(x) = dg(x), \quad x \in S\}.$$

Note that L_g is a n-dimensional submanifold which is *transversal* to the fibers of the fibration $\pi : T^*S \to S$, that is, its tangent bundle TL_g is transversal to the vertical bundle $\ker T\pi$.

According to a theorem of Maslov–Hormander ([16,17]), a general (i.e. not necessarily trasversal) Lagrangian submanifold of T^*S can be locally described as the graph of a smooth function G depending on extra parameters. Let us sketch briefly this construction along the lines of the works in [18,19].

Let U be a k-dimensional manifold called supplementary manifold, and let $G : S \times U \to \mathbb{R}$ be a smooth function whose representation in a local chart is $G(x, u)$. We define the *critical set* of G as (we use the notation $(G_x)_i = \partial G / \partial x^i$ and $(G_{xy})_{ij} = \partial^2 G / \partial x^i \partial y^j$) for partial derivatives)

$$\mathcal{E} = \{(x, u) \ : \ G_u(x, u) = 0\}. \tag{2}$$

If dG_u has maximal rank over $\mathcal{E}$, that is,

$$\operatorname{rk} dG_u = \operatorname{rk}\begin{pmatrix} G_{xu} & G_{uu} \end{pmatrix} = k \quad \text{for all} \quad (x, u) \in \mathcal{E} \tag{3}$$

then G is called *Morse family* and the following Λ_G is a Lagrangian submanifold of T^*S,

$$\Lambda_G = \{(x, G_x(x, u)) \in T^*S \quad \text{where} \quad (x, u) \in \mathcal{E}\}. \tag{4}$$

If there are no extra parameters $k = 0$, then Λ_G is the graph of a differential and thus Λ_G is a transversal submanifold. Note that the above rank condition (3) can be satisfied if the square submatrix G_{uu} has maximal rank, i.e., $\det G_{uu} \neq 0$ on $\mathcal{E}$. In this case, by the implicit function theorem there exist a locally defined function $u = u(x)$ such that $\mathcal{E}$ is the graph of u and setting $\hat{G}(x) = G(x, u(x))$ we have that

$$\hat{G}_x(x) = G_x(x, u(x)) + G_u(x, u(x))u_x(x) = G_x(x, u(x)) \quad \text{for all } (x, u) \in \mathcal{E}.$$

Therefore, where $\det G_{uu} \neq 0$ on $\mathcal{E}$, all the parameters u can be eliminated and $\Lambda_{\hat{G}}$ is locally transversal to the fibers. The set of points of S where $\det G_{uu}(x, u) = 0$ for $(x, u) \in \mathcal{E}$ is called the *caustic* of Λ_G.

These are the points where the Lagrangian submanifold is *tangent* to the fibers of $\pi : T^*S \to S$ and trasversality is lost.

2.1. Symplectic Structures Defined by Divergence Functions

Given a smooth n-dimensional manifold, M, let us denote with $M^2 = M \times M$ the square of M and with $\Delta_M \subset M^2$ the diagonal of M^2. We will use local coordinates $x = (x^1, \dots, x^n)$ on M and $(x, y) = (x^1, \dots, x^n, y^1, \dots, y^n)$ on M^2.

Let $D : M^2 \to [0, +\infty)$ be a smooth non-negative function whose representation in a local chart is $D(x, y) \geq 0$. We use the notations

$$(D_x)_i = \frac{\partial D}{\partial x^i}, \quad (D_y)_j = \frac{\partial D}{\partial y^j}, \quad (D_{yx})_{ji} = \frac{\partial}{\partial y^j}\left(\frac{\partial D}{\partial x^i}\right) = -\phi_{ji}$$

for first and second order derivatives of D. The function D is a *yoke* (see [7]) if the following conditions hold and D is a *divergence* (see [8]) if iii) below holds on the whole M^2.

(i) $D = 0$ only on Δ_M
(ii) $D_x = 0$ and $D_y = 0$ on Δ_M
(iii) $\phi = -D_{xy}$ is positive definite on Δ_M

thus points of Δ_M are minima of D. A divergence function act as a pseudo-distance but it does not satisfy the symmetry nor the triangle inequality conditions. In [7], the following fibered map $F_D : M^2 \to T^*M$ over M is considered, whose representation in a local chart is

$$F_D(x, y) = (x, D_x(x, y)). \tag{5}$$

By condition (iii) above there exist a neighborhood W of Δ_M, where F_D has a smooth inverse

$$F_D^{-1}(x, \lambda) = (x, y(x, \lambda)).$$

Using the local diffeomorphism F_D a symplectic structure (W, ω_D) is defined in [7] via the pull-back $\omega_D = F_D^* \omega$ of the canonical two form (1) on T^*M. The local form of ω_D can be computed as follows,

$$\omega_D = F_D^* \omega = F_D^*(d\theta_c) = d(F_D^* \theta_c) = d\big((D_x)_i dx^i\big) \tag{6}$$

thus (see Section 3.2 in [7])

$$\omega_D = \frac{\partial^2 D}{\partial x^j \partial x^i} dx^j \wedge dx^i + \frac{\partial^2 D}{\partial y^j \partial x^i} dy^j \wedge dx^i = -\phi_{ji} dy^j \wedge dx^i$$

because the first term $\partial^2 D / \partial x^j \partial x^i$ is symmetric in the i, j indices. For the applications that we have in mind of the above theory, we will assume in (iii) above that $-D_{yx}$ is positive definite on the *whole* M^2 so that F_D is a global diffeomorphism.

Simple examples of Lagrangian submanifolds of M^2 with respect to ω_D are (with a little abuse of notation) the n-dimensional submanifolds $M_x = M \times \{y\} \approx M$, which are also transversal to the fibers of $\pi_1 : M^2 \to M$, $\pi_1(x, y) = x$. Moreover, as $\omega_D(u, u) = 0$, Δ_M is also a Lagrangian submanifold.

Note also that (6) implies that F_D is a symplectomorphism, thus $L = F_D^{-1}(\Lambda)$ is a Lagrangian submanifold of M^2 whenever $\Lambda \subset T^*M$ is a Lagrangian submanifold. In this paper, we will be mainly concerned with the study of Lagrangian submanifolds of M^2 defined in this way.

In the following Section 2.2, we will compute the above introduced objects for the relevant case of exponential families of probability distributions and canonical divergence.

In [7], the Hamiltonian $H : T^*M \to [0, +\infty)$ associated to a divergence function is defined as $H = D \circ F_D^{-1}$ and locally it has the form

$$H(x, \lambda) = D(x, y(x, \lambda)). \tag{7}$$

2.2. Canonical Divergence and Exponential Families

In this section, we recall the basic definitions of exponential family and canonical divergence, as described, e.g., in [1,2]. Let $(X, \mathcal{B}, dx)$ be a probability space, where X may be a discrete set or $X = \mathbb{R}^k$. We stipulate that in case of a discrete set the integrals over X with respect to the measure dx are substituted by summations. Let

$$\mathcal{P}(X) = \{p : X \to [0, +\infty),\ p(x) \geq 0, \int_X p dx = 1\}$$

and suppose that $q \in \mathcal{P}(X)$ for suitable k, where $q(x) = e^{k(x)} > 0$. Consider n independent *observables*

$$h : X \to \mathbb{R}^n, \quad \text{rk}\, dh(x) = n \quad \forall\, x \in X$$

and define the related *free energy* $\psi : \Theta \subset \mathbb{R}^n \to \mathbb{R}$ as (here $\theta \cdot h = \theta^i h_i$)

$$e^{\psi(\theta)} = \int_X e^{\theta \cdot h(x) + k(x)} dx. \tag{8}$$

The n real numbers θ^i are called *canonical* parameters. They define uniquely a probability distribution $p(\cdot; \theta)$ which belongs to the *exponential family* defined by h, k,

$$M(h, k) = \{\ p(x; \theta) = e^{\theta \cdot h(x) + k(x) - \psi(\theta)},\ \theta \in \Theta\} \subset \mathcal{P}(X). \tag{9}$$

The relevant fact is that $M(h, k)$ is a n-dimensional submanifold of the infinite dimensional set $\mathcal{P}(X)$ and that the canonical parameters θ are local coordinates. Note that $q \in M(h, k)$ as $\psi(0) = 0$ and $q(x) = p(x; 0)$. Another system of local coordinates is provided by the so-called *expectation* parameters defined by

$$\eta = \psi_\theta(\theta) = \mathbb{E}_{p_\theta}[h] = \int_X h(x) p(x; \theta) dx.$$

As ψ is a convex function, the gradient map $\psi_\theta(\theta) = \eta$ is globally invertible with inverse $\theta = \hat{\theta}(\eta)$, which is also a gradient map $\hat{\theta}(\eta) = \varphi_\eta(\eta)$, where

$$\varphi(\eta) = \hat{\theta}(\eta) \cdot \eta - \psi(\hat{\theta}(\eta)) \tag{10}$$

is the Legendre transform of ψ (see, e.g., in [1]). We will denote with $p(x; \eta)$ the point in $M(h, k)$ associated to η. The Kullback–Leibler divergence is defined for general $(p, \tilde{p})$ in $\mathcal{P}(X)^2$ as

$$D_{KL}(p, \tilde{p}) = \int_X p(x) \log \frac{p(x)}{\tilde{p}(x)} dx.$$

The restriction of D_{KL} to $M(h, k)^2$, the square of $M(h, k)$, $D_{KL} : M(h, k)^2 \to [0, +\infty)$ is called *canonical* divergence. It can be shown (see in [1]) that when $M(h, k)$ is *referred to the coordinates* (η, θ), D_{KL} has the local representation

$$D(\eta, \theta) = \varphi(\eta) + \psi(\theta) - \eta \cdot \theta. \tag{11}$$

Note that as $p(\cdot; \theta) = q$ for $\theta = 0$

$$D_{KL}(p, q) = D_{KL}(p(\cdot; \eta), p(\cdot; 0)) = \varphi(\eta) + \psi(0) - \eta \cdot 0 = \varphi(\eta). \tag{12}$$

A key object is the map F_D introduced in (5) associated to $M(h,k)$ and the canonical divergence (11). It has the local form in coordinates (η,θ), see (5) and (11),

$$F_D(\eta,\theta) = (\eta, D_\eta) = (\eta, \varphi_\eta(\eta) - \theta), \tag{13}$$

with the *explicit* inverse, using local coordinates (η,λ) in $T^*M(h,k)$,

$$F_D^{-1}(\eta,\lambda) = (\eta,\theta(\eta,\lambda)) = (\eta,\varphi_\eta(\eta)-\lambda) = (\eta,\hat\theta(\eta)-\lambda). \tag{14}$$

A simple but elegant result of the above-introduced framework is the following.

Proposition 1. *Let Λ_G be a Lagrangian submanifold of $T^*M(h,k)$ described by the Morse family $G(\eta,u)$ as in (4). Then, $L_S = F_D^{-1}(\Lambda_G)$ is a Lagrangian submanifold of $M(h,k)^2$ described by the Morse family $S(\eta,u) = \varphi(\eta) - G(\eta,u)$.*

Proof. From (4) we have that $\lambda = G_\eta(\eta,u)$ on Λ_G and from (14)

$$F_D^{-1}(\Lambda_G) = \{(\eta,\theta) = (\eta,\varphi_\eta(\eta) - G_\eta(\eta,u)) = (\eta, S_\eta(\eta,u)),\ (\eta,u)\in\mathcal{E}\}$$

where $S(\eta,u) = \varphi(\eta) - G(\eta,u)$. Moreover, as $S_u(\eta,u) = G_u(\eta,u)$ the critical set $\mathcal{E}$ in (2) is the same. □

As a consequence of the above proposition, if Λ_G is transversal to the fibers of $T^*M(h,k)$ (no extra parameters u), then its image in $M(h,k)^2$ is transversal to the fibers of π_1.

Another interesting consequence is that the zero section of the cotangent bundle $T^*M(h,k)$, locally represented as $Z = \{(\eta,0)\ :\ \eta\in E\}$, is mapped by F_D^{-1} into

$$Z_0 = F_D^{-1}(Z) = \{(\eta,\hat\theta(\eta))\ :\ \eta\in E\}$$

which is contained into $D^{-1}(0)$, the zero-level set of the canonical divergence. Indeed, from (10) and (11) we have that

$$D(\eta,\hat\theta(\eta)) = \varphi(\eta) + \psi(\hat\theta(\eta)) - \eta\cdot\hat\theta(\eta) = \varphi(\eta) - \varphi(\eta) \equiv 0 \tag{15}$$

thus $Z_0 \subset D^{-1}(0)$ in the general case and $Z_0 = D^{-1}(0)$ if $n = 1$. For later use, we compute from (7) the Hamiltonian associated to the canonical divergence

$$H(\eta,\lambda) = D\circ F_D^{-1}(\eta,\lambda) = \varphi(\eta) + \psi(\hat\theta(\eta)-\lambda) - \eta\cdot(\hat\theta(\eta)-\lambda).$$

We set for the sake of simplicity $\hat\theta(\eta) = \hat\theta$ and we compute from (8) the free energy $\psi(\hat\theta(\eta)-\lambda)$

$$\begin{aligned} e^{\psi(\hat\theta-\lambda)} &= \int_X e^{(\hat\theta-\lambda)\cdot h + k}dx = \int_X e^{(\hat\theta-\lambda)\cdot h + k + \psi(\hat\theta)-\psi(\hat\theta)}dx \\ &= e^{\psi(\hat\theta)}\int_X e^{-\lambda\cdot h}e^{\hat\theta\cdot h + k - \psi(\hat\theta)}dx = e^{\psi(\hat\theta)}\mathbb{E}_{p_{\hat\theta}}[e^{-\lambda\cdot h}]. \end{aligned} \tag{16}$$

Using (15) and (16), the Hamiltonian can be written using relation (10) as

$$\begin{aligned} H(\eta,\lambda) &= \varphi(\eta) + \psi(\hat\theta) + \ln\mathbb{E}_{p_{\hat\theta}}[e^{-\lambda\cdot h}] - \eta\cdot\hat\theta + \eta\cdot\lambda \\ &= \ln\mathbb{E}_{p_{\hat\theta}}[e^{-\lambda\cdot h}] + \eta\cdot\lambda. \end{aligned} \tag{17}$$

It is interesting to investigate more in detail the structure of the Lagrangian submanifold $L_S = F_D^{-1}(\Lambda_G) \subset M(h,k)^2$ by studying the form of the two probability distributions

$F_D^{-1}(\eta, \lambda) = (\eta, \hat{\theta} - \lambda)$ in L_S associated to the coordinates respectively η and $\hat{\theta} - \lambda$. We compute from (9)

$$p(x; \eta) = e^{\hat{\theta} \cdot h(x) + k(x) - \psi(\hat{\theta})}$$

and using (17)

$$\begin{aligned} p(x; \hat{\theta} - \lambda) &= e^{(\hat{\theta} - \lambda) \cdot h + k - \psi(\hat{\theta} - \lambda)} \\ &= e^{\hat{\theta} \cdot h - \lambda \cdot h + k - \psi(\hat{\theta}) - \ln \mathbb{E}_{p_{\hat{\theta}}}[e^{-\lambda \cdot h}]} \\ &= p(x; \eta) \frac{e^{-\lambda \cdot h(x)}}{\mathbb{E}_{p_{\hat{\theta}}}[e^{-\lambda \cdot h}]} . \end{aligned} \tag{18}$$

Note that setting

$$p(x; \lambda) = \frac{e^{-\lambda \cdot h(x)}}{Z(\lambda)} = \frac{e^{-\lambda \cdot h(x)}}{\int_X e^{\lambda \cdot h(x) dx}}$$

relation (18) can be given the form

$$p(x; \hat{\theta} - \lambda) = \frac{p(x; \eta) e^{-\lambda \cdot h(x)}}{\int_X p(x; \eta) e^{-\lambda \cdot h(x)} dx} = \frac{p(x; \eta) p(x; \lambda)}{\int_X p(x; \eta) p(x; \lambda) dx} . \tag{19}$$

We will give an interpretation of this relation in the case of discrete probability distributions in Section 3.2 below.

3. Application to Maximum Entropy Principle with Nonlinear Constraints and Phase Transitions

A relevant application of the above-introduced framework concerns the use of the Maximum Entropy Principle with *nonlinear* constraints. Let us consider a physical system X whose description is given in terms of a probability distribution $q \in \mathcal{P}(X)$. The Maximum Entropy Principle (E.T. Jaynes, see in [11,12]) is a general inference procedure that allows to update an initial probability distribution q on the basis of subsequent information on the system represented by the average values $\mathbb{E}_p[h]$ of some observables h of interest for the system. The sought distribution p is the one that minimizes the relative entropy $D_{KL}(p, q)$ on the set of the distributions which satisfy the constraints on $\mathbb{E}_p[h]$. From a mathematical point of view, we are faced with a constrained extremization problem to be solved below using the Lagrange multipliers method.

We will see that the set of solutions for different values of the constraints defines a Lagrangian submanifold of a cotangent space of a manifold $M(h, k)$. We are interested in describing the corresponding Lagrangian submanifold in $M(h, k)^2$.

This section has a pedagogical character, so for the sake of simplicity we will avoid technicalities and assume that $X = \{1, \ldots, n\}$ is a discrete space and that there is only one observable of interest defined by assigning $h = (h_1, \ldots, h_n)$. The case of k observables can be dealt with along the same lines with no extra effort. The case of a continuous space $X \subset \mathbb{R}^n$ presents more technical difficulties and it is considered in [20].

Let $q_i = e^{k_i} \in \mathcal{P}(X)$ be the a priori distribution describing X. The Kullback–Leibler divergence is called *relative entropy* in this setting and has the form

$$D(p, q) = \sum_i p_i \ln \frac{p_i}{q_i}.$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth globally non-invertible function (think for example of a cubic $f(x) = x(x^2 - a^2)$ for $a \in \mathbb{R}$, see Figure 1 below). We look for the minima of D on the set of $p \in \mathcal{P}(X)$ that satisfy the nonlinear constraint on p in the form $g : \mathbb{R}^n_+ \to \mathbb{R}$, $g(p) = y$ that is

$$g(p) = f(\mathbb{E}_p[h]) = f(\sum_{i=1}^{n} h_i p_i) = y. \tag{20}$$

The choice of this type of constraints is motivated by classical applications in statistical physics. For example in the Ising model in the Curie–Weiss (mean field) approximation the average energy of the spin lattice is a quadratic function of the average magnetization $\mathbb{E}_p[s]$, see [14,15]. We have that

$$dg(p) = f'(\mathbb{E}_p[h])h = (f'(\mathbb{E}_p[h])h_1, \ldots, f'(\mathbb{E}_p[h])h_n). \tag{21}$$

Note that we do not take into account at this stage of the procedure the normalization constraint stipulating that we will enforce it by dividing any candidate extremum point $\hat{p}$ by $\sum_i \hat{p}_i$. After introducing the Lagrange function where λ is the Lagrange multiplier associated to the constraint (20)

$$G(y, p, \lambda) = D(p, q) - \lambda(f(\mathbb{E}_p[h]) - y) \tag{22}$$

we see that the candidate extrema are the solutions (p, λ) for given y of (here $i = 1, \ldots, n$)

$$(G_p)_i = \ln \frac{p_i}{q_i} + 1 - \lambda f'(\mathbb{E}_p[h])h_i = 0, \quad G_\lambda = f(\mathbb{E}_p[h]) - y = 0 \tag{23}$$

that is, setting $q_i = e^{k_i}$, we have to face a trascendental equation for the unnormalized probability

$$p_i = c\, e^{\lambda f'(\mathbb{E}_p[h])h_i + k_i}, \quad f(\mathbb{E}_p[h]) = y. \tag{24}$$

After normalization, $(24)_1$ becomes

$$p_i = e^{\lambda f'(\mathbb{E}_p[h])h_i + k_i - \psi(\lambda, p)}, \qquad e^{\psi(\lambda, p)} = \sum e^{\lambda f'(\mathbb{E}_p[h])h_i + k_i}. \tag{25}$$

Let us denote with $f^{\leftarrow}(y) \subset \mathbb{R}$ the set of pre-images of y along f (see, e.g., Figure 1 below)

$$f^{\leftarrow}(y) = \{\eta \in \mathbb{R} \,:\, f(\eta) = y\} = \{\eta_1, .., \eta_\alpha, .., \eta_A\}, \quad \eta_\alpha = \eta_\alpha(y) \tag{26}$$

where we have supposed that, for every y, $f^{\leftarrow}(y)$ is a *finite* set of cardinality $A(y) < +\infty$. The crux is that we can substitute the constraint $f(\mathbb{E}_p[h]) = y$ in $(24)_2$ with the following equivalent one

$$f(\mathbb{E}_p[h])) = y \quad \Longleftrightarrow \quad \mathbb{E}_p[h] \in f^{\leftarrow}(y)$$

therefore we can describe the—possibly non-unique—solution (25) of the extremum problem (23) as

$$p_i^\alpha = e^{\lambda f'(\eta_\alpha)h_i + k_i - \psi(\lambda, \alpha)}, \quad e^{\psi(\lambda, \alpha)} = \sum e^{\lambda f'(\eta_\alpha)h_i + k_i} \tag{27}$$

where $\alpha = 1, \ldots, A(y)$, showing that the candidate solution belongs to an exponential family $M(h, k)$. Note that in Information Geometry, the critical points of the MEP extremum problem are computed as *geodesic projections* over a submanifold which is an exponential family and multiplicity of solutions are related to the non-uniqueness of the geodesic projection, see in [1,15].

Note that *where* $f'(\eta_\alpha) \neq 0$ setting $\lambda f'(\eta_\alpha) = \theta_\alpha$ the solution (27) can be given the standard form (see in [1,14]) of MEP solution

$$\hat{p}_i = e^{\hat{\theta} h_i + k_i - \psi(\hat{\theta})}, \quad \hat{\theta}(\eta) = \varphi_\eta(\eta)$$

with *linear* constraint $\mathbb{E}[h] = \eta_\alpha$, hence (25) becomes

$$p_i^\alpha = e^{\theta_\alpha h_i + k_i - \psi(\theta_\alpha)}, \qquad e^{\psi(\theta_\alpha)} = \sum e^{\theta_\alpha h_i + k_i}. \tag{28}$$

The multipliers $\theta_\alpha = \hat{\theta}(\eta_\alpha(y)),\ \alpha = 1, \dots, A(y)$ are uniquely determined (see (10)) by the equation

$$\psi_\theta(\theta) = \eta \qquad \text{i.e.} \qquad \hat{\theta}(\eta) = \varphi_\eta(\eta) \tag{29}$$

for $\eta = \eta_\alpha(y)$ and accordingly we can compute the multipliers λ as

$$\lambda_\alpha(y) = \frac{\hat{\theta}(\eta_\alpha(y))}{f'(\eta_\alpha(y))}. \tag{30}$$

Note that the solution to our constrained extremization problem (28) has the form of a curved exponential family (see [1]) with respect to the discrete parameter α. We will see in the next Section 3.1 that the framework of Lagrangian submanifold is useful to describe the global picture of the solutions in case of multiple solutions.

3.1. The Global Picture via Lagrange Submanifold

If we set in the Lagrange function (22) $(p, \lambda) = u$, we see that for $G(y, u)$ the set of points (y, u) satisfying the first order necessary condition for unconstrained extremum (23) is the *critical* set

$$\mathcal{E} = \{(y, u) : G_u(y, u) = 0\}.$$

We can check if the Lagrange function $G(y, u)$ defines a Morse family using the rank condition (3)

$$\operatorname{rk}\begin{pmatrix} G_{yu} & G_{uu} \end{pmatrix} = n + 1 \quad \text{for all} \quad (y, u) \in \mathcal{E}$$

where in this case

$$\begin{pmatrix} G_{yu} & G_{uu} \end{pmatrix} = \begin{pmatrix} 0 & G_{pp} & -dg^T \\ 1 & -dg & 0 \end{pmatrix} \tag{31}$$

and G_{pp} is the n-dimensional Hessian matrix (here δ_{ij} is Kronecker symbol)

$$(G_{pp})_{ij} = (D_{pp})_{ij} - \lambda f''(\mathbb{E}_p[h]) h_i h_j = \frac{\delta_{ij}}{p_i} - \lambda f''(\mathbb{E}_p[h]) h_i h_j. \tag{32}$$

If $G(y, u)$ is a Morse family, then by Maslov–Hormander theorem

$$\Lambda_G = \{(y, G_y) \quad \text{where } (y, u) \in \mathcal{E}\} \tag{33}$$

is a Lagrangian submanifold of $T^*\mathbb{R}$. We claim that (33) provides a global description of the set of solutions (28). We have seen in Section 1 that a sufficient condition for the elimination of all extra parameters u is that G_{uu} has maximal rank for all $(y, u) \in \mathcal{E}$. A criterion for this is given by the following classical result in constrained optimization theory, here adapted to our notations, which express the second order sufficient condition for maxima or minima (see in [14,21] for the proof).

Proposition 2. *If the symmetric matrix G_{pp} in (32) is (positive or negative) definite on* $\ker dg$ *for $(y, u) \in \mathcal{E}$, then the square matrix G_{uu} in (31) has maximal rank.*

From (21), we have that for $(y, u) \in \mathcal{E}$

$$\ker dg(p) = \{u \in \mathbb{R}^n \ : \ f'(\eta_\alpha) h \cdot u = 0\}$$

and from (32), that

$$G_{pp}u \cdot u = \Big(\sum_i \frac{u_i^2}{p_i}\Big) - \lambda f''(\eta_\alpha)(h \cdot u)^2.$$

It is straightforward to derive from the above relations that the two cases below hold

$$\begin{cases} f'(\eta_\alpha) \neq 0 \;\Rightarrow\; \ker dg(p) = \{u \;:\; h \cdot u = 0\} \;\Rightarrow\; G_{pp}u \cdot u > 0 \;\; \forall\, u \neq 0, \\ \\ f'(\eta_\alpha) = 0 \;\Rightarrow\; \ker dg(p) = \mathbb{R}^n \;\Rightarrow\; G_{pp}u \cdot u \in \mathbb{R}. \end{cases}$$

Therefore, at points $(y,u) \in \mathcal{E}$ where $f'(\eta_\alpha) \neq 0$ the Lagrangian submanifold Λ_G in (33) is transversal. At points in $\mathcal{E}$ where $f'(\eta_\alpha) = 0$, we have $dg = f'(\eta_\alpha)h = 0$, see (21), thus transversality is lost as—see the form of G_{uu} in (31)—for these points

$$\det G_{uu}(p,\lambda) = 0, \;\text{ and }\; (y,u) \in \mathcal{E}.$$

We remark that the above introduced framework is able to give the global description of the set of solutions (28), (30) in terms of the Lagrangian submanifold locally described as

$$\Lambda_f^{(y)} = \{(y, G_y) = \big(y, \lambda(\eta_\alpha(y))\big) = \big(y, \frac{\hat\theta(\eta_\alpha(y))}{f'(\eta_\alpha(y))}\big)\} \subset T^*\mathbb{R}_y \tag{34}$$

where $\lambda(\eta_\alpha(y))$ is given by (30). If we consider $f : E \subset \mathbb{R}_\eta \to \mathbb{R}_y$, $y = f(\eta)$ as a local change of coordinates on $M(h,k)$ (since f is locally invertible where $f'(\eta) \neq 0$) it is easy to prove that

Proposition 3. *The submanifold $\Lambda_f^{(y)} \subset T^*\mathbb{R}_y$ in (34) is the image $\Lambda_f^{(y)} = T^*f(\Lambda_f)$ of*

$$\Lambda_f = \{(\eta, \hat\theta_\alpha(\eta)) \;:\; \eta \in E\} \subset T^*M(h,k) \tag{35}$$

where $\hat\theta_\alpha(\eta)$ is the multiplier in (29) associated to the constraint $\mathbb{E}_p[h] = \eta_\alpha$ and $\eta_\alpha \in I(\eta) = f^{\leftarrow}(f(\eta))$.

Proof. If $y = f(\eta)$ is the local change of coordinates in $M(h,k)$, then the tangent map $Tf : T\mathbb{R}_\eta \to T\mathbb{R}_y$ has the local form $(y, \dot y) = Tf(\eta, \dot\eta) = (f(\eta), f'(\eta)\dot\eta)$ and the cotangent map $T^*f : T^*\mathbb{R}_\eta \to T^*\mathbb{R}_y$ has the local form

$$(y, \lambda) = T^*f(\eta, \beta) = \big(f(\eta), \frac{\beta}{f'(\eta)}\big)$$

if we want that the Liouville one-form (see above (1)) has the same canonical form $\theta_c = \lambda dy = \beta d\eta$ in the two coordinate charts. See, e.g., in [19] for a proof of this last classical result of differential geometry. □

We want to study the Lagrangian submanifold Λ_f defined in (35) and its image $L_f = F_D^{-1}(\Lambda_f) \subset M(h,k)^2$, where F_D^{-1} is defined in (14), whose local expression is

$$L_f = \{(\eta, \hat\theta(\eta) - \hat\theta(\eta_\alpha)) \;:\; \eta \in E\}. \tag{36}$$

First we consider the case that f is a globally invertible function. In this case, $I(\eta) = f^{\leftarrow}(f(\eta)) = \{\eta\}$ and $\hat\theta(\eta) = \varphi_\eta(\eta)$. The Lagrangian submanifold Λ_f in (35) is the graph of the differential $\varphi_\eta(\eta)$ and it is transversal, see Figure 2a. Moreover, see below (9), if $\eta = \eta_0 = \mathbb{E}_q[h]$ then $\hat\theta(\eta_0) = 0$. As $\psi_\theta(\theta) = \eta$ is invertible with inverse $\theta = \hat\theta(\eta)$, we have

$$\frac{d\hat\theta}{d\eta}(\eta) = (\frac{d^2\psi}{d\theta^2})^{-1} = \mathrm{var}_{\hat p}(h) = \mathbb{E}_{\hat p}[h^2] - \eta^2 > 0$$

and $\hat{\theta}(\eta)$ is a monotonically increasing function, see Figure 2a. Its image (36) is $L_f = M(h,k) \times \{0\}$, see Figure 2b.

If we consider a globally non invertible function f as the one depicted in Figure 1, then $I(\eta)$ contains multiple points and Λ_f is non transversal at points where $f'(\eta) = 0$, see Figure 3a. The corresponding image L_f has multiple branches and it is not a manifold at points (b, c) where transversality fails, see Figure 3b).

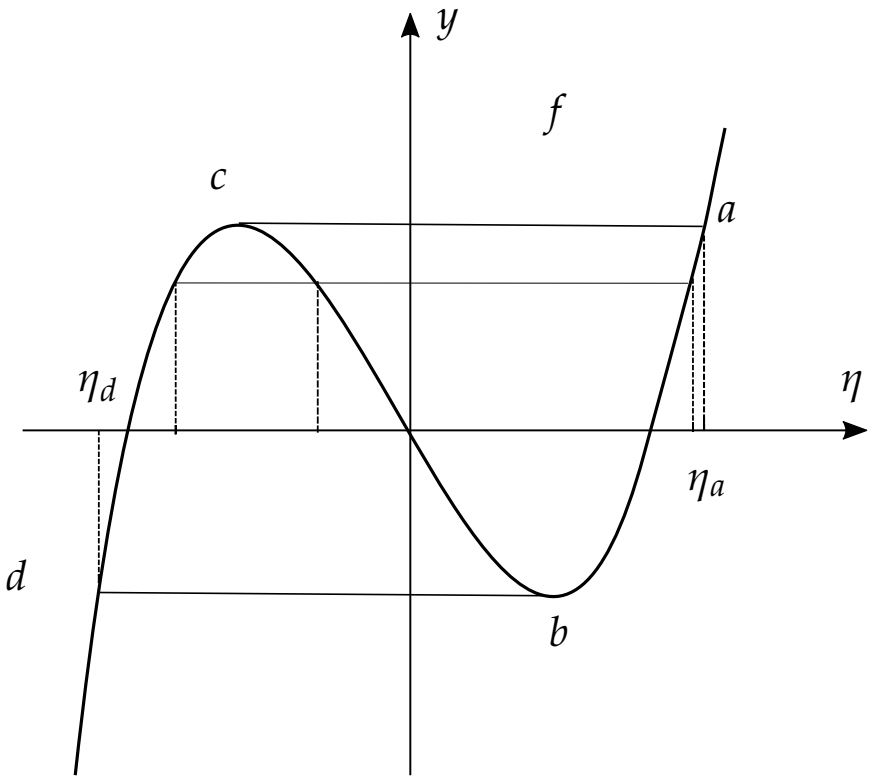
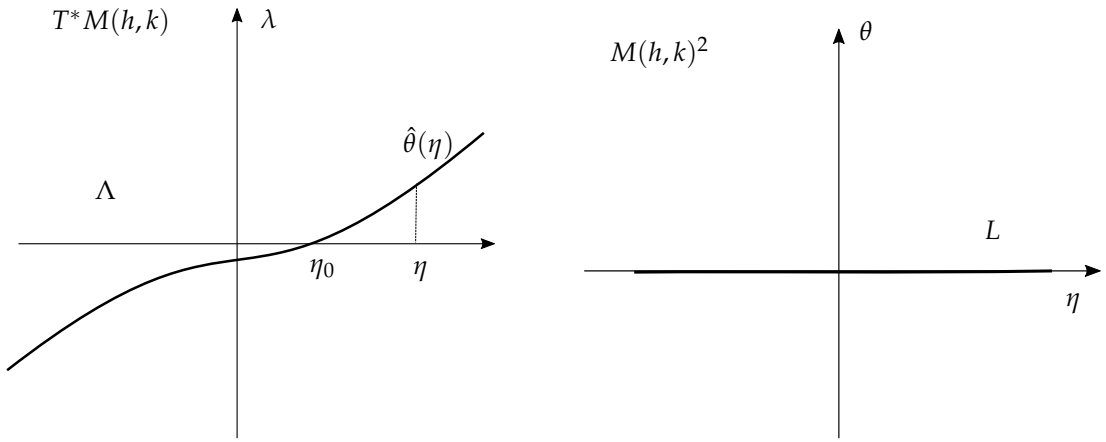


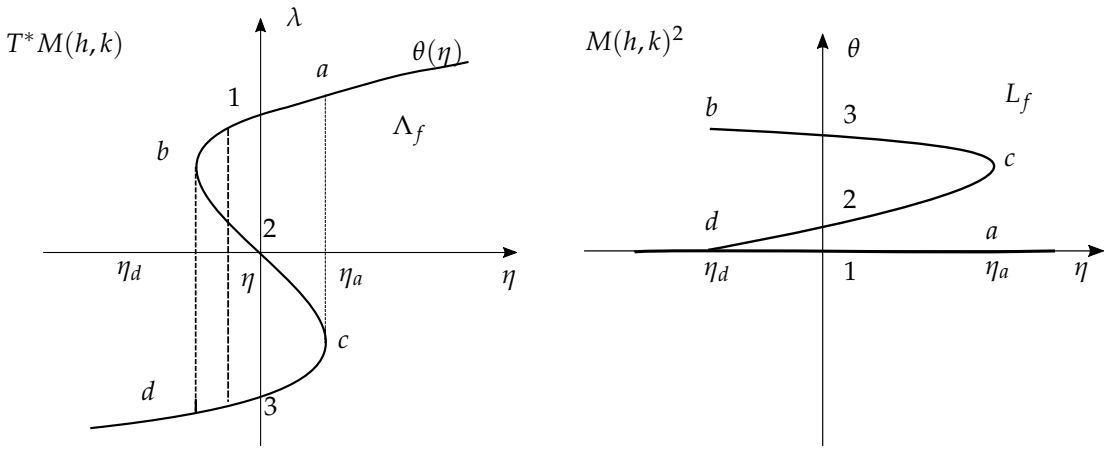
Figure 1. Plot of $y = f(\eta) = \eta(\eta^2 - a^2)$. Points b, c correspond to points where $f'(\eta) = 0$.



(a) A transversal Lagrangian submanifold Λ of $T^*M(h,k)$ generated by the constraint $\mathbb{E}_p(h) = \eta$

(b) The image $L = F_d^{-1}(\Lambda)$ in $M(h,k)^2$

Figure 2. The case of a transversal Lagrangian submanifold.



(a) Lagrangian submanifold Λ_f of $T^*M(h,k)$ generated by the nonlinear constraint $f(\mathbb{E}_p(h)) = y$. Points b, c corresponds to points where $f'(\eta) = 0$ and Λ_f is non transversal

(b) The image $L_f = F_d^{-1}(\Lambda_f)$ in $M(h,k)^2$

Figure 3. The case of a folded, i.e., non transversal Lagrangian submanifold.

3.2. Probability Distributions in L_f

In this section, we study the structure of the probability distributions in L_f. In the local coordinate systems (η, θ) of $M(h,k)^2$, η and $\hat{\theta}(\eta)$ describe the same probability distribution that we write for brevity as $p_i(\eta) = p_i(\hat{\theta})$. Therefore, the probability distributions in L_f in (36) associated to η and $\hat{\theta}(\eta) - \hat{\theta}(\eta_\alpha)$ are, respectively,

$$p_i(\eta) = e^{\hat{\theta}h_i - k_i - \psi(\hat{\theta})} \tag{37}$$

and, see (18),

$$p_i(\hat{\theta} - \hat{\theta}(\eta_\alpha)) = p_i(\eta)\frac{e^{-\hat{\theta}(\eta_\alpha)h_i}}{\sum_i p_i(\eta)e^{-\hat{\theta}(\eta_\alpha)h_i}}. \tag{38}$$

Setting

$$\tilde{p}_i(\eta_\alpha) = \frac{e^{-\hat{\theta}(\eta_\alpha)h_i}}{Z(\lambda)}, \quad Z(\lambda) = \sum_i e^{-\hat{\theta}(\eta_\alpha)h_i},$$

the above (38) can be rewritten as the discrete version of (19), that is,

$$p_i(\hat{\theta} - \hat{\theta}(\eta_\alpha)) = \frac{p_i(\eta)\tilde{p}_i(\eta_\alpha)}{\sum_i p_i(\eta)\tilde{p}_i(\eta_\alpha)}. \tag{39}$$

This last formula can be interpreted as follows; let A and B be two independent random variables A, B: $\Omega \to X$, where $X = \{1, \ldots, n\}$ is the discrete state space, described by the probability distributions p_i and $\tilde{p}_i$, respectively (for example, A and B describe two dices with n faces). Then, $\sum_i p_i\tilde{p}_i$ is the probability that A and B are found in the *same* state and

$$Prob(A = i, B = i | A = B) = \frac{p_i\tilde{p}_i}{\sum_i p_i\tilde{p}_i}$$

in (39) is the conditional probability that A and B are found in the state i provided that they are found in the same state. Note that for $p_i(\eta)$ in (37) we have $e^{k_i} = q_i$, thus (37) can be rewritten as

$$p_i(\eta) = q_i\frac{e^{\hat{\theta}h_i}}{\sum_i q_i e^{\hat{\theta}h_i}} = \frac{q_i\tilde{p}(\hat{\theta})}{\sum_i q_i\tilde{p}_i(\hat{\theta})}$$

and (39) above is equal to

$$p_i(\hat{\theta} - \hat{\theta}(\eta_\alpha)) = \frac{q_i p_i \tilde{p}_i}{\sum_i q_i p_i \tilde{p}_i} = Prob(A = i, B = i, C = i | A = B = C)$$

where A, B, C are described by q_i , $p_i = \tilde{p}_i(\hat{\theta}(\eta))$, $\tilde{p}_i = \tilde{p}_i(\hat{\theta}(\eta_\alpha))$.

4. Discussion

Canonical coordinates η and θ associated to an exponential family $M(h,k)$ are dually flat coordinates with respect to the duality defined by the canonical divergence. With respect to these coordinates, a generalization of the Pitagorean theorem is proved in Information Geometry which provides a generalized formulation of the Maximum Entropy Principle with linear constraints as a geodesic projection problem (see [2]). Multiplicity of the solutions $\hat{\theta}(\eta)$ of the Maximum Entropy problem are due to the non uniqueness of the projection. In this paper, we have shown that the set of couples $(\eta, \hat{\theta}(\eta))$ defines a transversal Lagrangian submanifold Λ of $T^*M(h,k)$, and we have seen with an example that if nonlinear constraints are considered the set of possible multiple solutions to the Maximum Entropy problem is globally described by a folded (i.e., a possibly non-trasversal) Lagrangian submanifold Λ_f. We have computed their pull-back to the square manifold $M(h,k)^2$ via the map F_D^{-1}. We think that this framework offers a complementary view to the generalized Pitagorean Theorem. We plan to address in a subsequent paper a generalization of the theory presented here to a more general form of nonlinear constraint.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Amari, S. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 194.
2. Amari, S.; Hiroshi, N. *Methods of Information Geometry*; American Mathematical Soc.: Providence, RI, USA, 2007; Volume 191.
3. Murray, M.K.; Rice, J.W. *Differential Geometry and Statistics*; CRC Press: Boca Raton, FL, USA, 1993; Volume 48.
4. Amari, S.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci. Tech.* **2010**, *58*, 183–195.
5. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391.
6. Ay, N.; Amari, S. A novel approach to canonical divergences within information geometry. *Entropy* **2015**, *17*, 8111–8129.
7. Barndorff-Nielsen, O.E.; Jupp, P.E. Statistics, yokes and symplectic geometry. *Ann. Fac. Sci. Toulouse Math.* **1997**, *6*, 389–427.
8. Leok, M.; Zhang, J. Connecting information geometry and geometric mechanics. *Entropy* **2017**, *19*, 518.
9. Noda, T. Symplectic structures on statistical manifolds. *J. Aust. Math. Soc.* **2011**, *90*, 371–384
10. Nakamura, Y. Completely integrable gradient systems on the manifolds of Gaussian and multinomial distributions. *Jpn. J. Ind. Appl. Math.* **1993**, *10*, 179.
11. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630
12. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
13. Brot, R. Phase Transitions. In *Statistical Physics. Phase Transitions and Superfluidity*; Brandeis University Summer Institute in Theoretical Physics, Gordon and Breach Science Publishers: London, UK1966; pp. 5–103
14. Favretti, M. Lagrangian submanifolds generated by the Maximum Entropy principle. *Entropy* **2005**, *7*, 1–14.
15. Fujiwara, A.; Shigeru, S. Hereditary structure in Hamiltonians: Information geometry of Ising spin chains. *Phys. Lett. A* **2010**, *374* ,911–916.
16. Maslov, V.P.; Bouslaev,V.C.; Arnol'd, V.I. *Theorie des Perturbations et Methodes Asymptotiques*; Dunod: Paris, France, 1972.
17. Hormander, L. Fourier integral operators. I. *Acta Math.* **1971**, *127*, 79.

18. Weinstein, A. *Lectures on Symplectic Manifolds*; American Mathematical Soc.: Providence, RI, USA, 1977; No. 29.
19. Cardin, F. *Elementary Symplectic Topology and Mechanics*; Springer: Berlin/Heidelberg, Germany,2015.
20. Favretti, M. Isotropic submanifolds generated by the Maximum Entropy Principle and Onsager reciprocity relations. *J. Funct. Anal.* **2005**, *227*, 227–243.
21. Bertsekas, D.P. *Constrained Optimization and Lagrange Multiplier Methods*; Academic Press: Cambridge, MA, USA, 2014.