Divergence Functions in Information Geometry

Domenico Felice[1(✉)] and Nihat Ay[1,2,3]

[1] Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22–04103, Leipzig, Germany
`felice@mis.mpg.de`
[2] Santa Fe Institute, Santa Fe, NM 87501, USA
[3] Faculty of Mathematics and Computer Science, University of Leipzig,
PF 100920, 04009 Leipzig, Germany

Abstract. A recently introduced canonical divergence $\mathcal{D}$ for a dual structure $(\mathrm{g}, \nabla, \nabla^*)$ on a smooth manifold M is discussed in connection to other divergence functions. Finally, general properties of $\mathcal{D}$ are outlined and critically discussed.

Keywords: Classical differential geometry · Riemannian geometries · Information Geometry

1 Introduction

The geometrical structure induced by a *divergence function* (or *contrast function*) on a smooth manifold M provides a unified approach to measurement of notions as information, energy, entropy, playing an important role in mathematical sciences to research random phenomena [1]. In the mathematical formulation, a divergence function $\mathcal{D}(p, q)$ on a smooth manifold M is defined by the first requirement for a distance:

$$\mathcal{D}(p, q) \geq 0, \qquad \mathcal{D}(p, q) = 0 \quad \text{iff} \quad p = q. \tag{1}$$

An important example of a divergence function is given by the Kullback-Leibler divergence $\mathrm{K}(p, q)$ in the context that p and q are the vectors of probabilities of disjoint events [2], namely

$$\mathrm{K}(p, q) = \sum_{i=1}^{n+1} p_i \, \log \left(\frac{p_i}{q_i} \right) \tag{2}$$

is a function on the n-simplex $\mathcal{S} := \{p = (p_1, \ldots, p_{n+1}) \,|\, p_i \geq 0, \sum_i p_i = 1\}$. Given a smooth n-dimensional manifold M, we assume that $\mathcal{D} : \mathrm{M} \times \mathrm{M} \to \mathbb{R}^+$ is a C^∞-differentiable function. Working with the local coordinates $\{\boldsymbol{\xi}_p := (\xi_p^1, \ldots, \xi_p^n)\}$

and $\{\boldsymbol{\xi}_q := (\xi_q^1, \ldots, \xi_q^n)\}$ at p and q, respectively, it follows from Eq. (1) that

$$\partial_i \, \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q} = 0, \quad \partial_i' \, \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q} = 0 \tag{3}$$

$$\partial_j \partial_i \, \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q} = - \, \partial_j' \partial_i \, \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q} = \partial_j' \partial_i' \, \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q}, \tag{4}$$

where $\partial_i = \frac{\partial}{\partial \xi_p^i}$ and $\partial_i' = \frac{\partial}{\partial \xi_q^i}$. Moreover, under the assumption that

$$\mathrm{g}_{ij} := \partial_i \partial_j \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q} > 0 \tag{5}$$

we can see that the manifold M is endowed, through the divergence function $\mathcal{D}$, with the Riemannian metric tensor given by $\mathrm{g} = \mathrm{g}_{ij} \, \mathrm{d}\xi^i \otimes \mathrm{d}\xi^j$, where the Einstein notation is adopted. The symmetry of g immediately follows from the requirement that $\mathcal{D}$ is a C^∞ function on M $\times$ M.

From Eq. (2) we can see that, in general, a divergence function $\mathcal{D}$ is not symmetric. The asymmetry of $\mathcal{D}$ leads to two different affine connections, ∇ and ∇^*, on M such that $1/2(\nabla + \nabla^*)$ is the Levi-Civita connection with respect to the metric tensor $\mathrm{g} = \mathrm{g}_{ij} \mathrm{d}\xi^i \otimes \mathrm{d}\xi^j$ defined by Eq. (5). More precisely, working with the local coordinates $\{\boldsymbol{\xi}_p\}$ and $\{\boldsymbol{\xi}_q\}$, we can define the symbols Γ_{ijk} and Γ_{ijk}^* of the connections ∇ and ∇^*, i.e. $\Gamma_{ijk} = \mathrm{g}\left(\nabla_{\partial_i}\partial_j, \partial_k\right)$ and $\Gamma_{ijk}^* = \mathrm{g}\left(\nabla_{\partial_i}^*\partial_j, \partial_k\right)$, by means of the following relations

$$\Gamma_{ijk}(p) = - \, \partial_i \partial_j \partial_k' \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q}, \qquad \Gamma_{ijk}^*(p) = - \, \partial_i' \partial_j' \partial_k \mathcal{D}(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)\big|_{p=q}. \tag{6}$$

To sum up, a divergence function $\mathcal{D}$ on a smooth manifold M induces a metric tensor on M by Eq. (5). In addition, the divergence $\mathcal{D}$ yields two linear torsion-free connections, ∇ and ∇^*, on TM which are dual with respect to the metric tensor g [2]:

$$X \, \mathrm{g}\left(Y, Z\right) = \mathrm{g}\left(\nabla_X Y, Z\right) + \mathrm{g}\left(Y, \nabla_X^* Z\right), \ \forall \ X, Y, Z \in \mathcal{T}(\mathrm{M}), \tag{7}$$

where $\mathcal{T}(\mathrm{M})$ denotes the space of vector fields on M. Finally, we refer to the quadruple $(\mathrm{M}, \mathrm{g}, \nabla, \nabla^*)$ as a *statistical manifold* [3].

1.1   The Inverse Problem Within Information Geometry

The inverse problem is to find a divergence $\mathcal{D}$ which generates a given geometrical structure $(\mathrm{M}, \mathrm{g}, \nabla, \nabla^*)$. For any such statistical manifold there exists a divergence $\mathcal{D}$ such that Eqs. (5) and (6) hold true [4]. However, this divergence is not unique and there are infinitely many divergences generating the same geometrical structure $(\mathrm{g}, \nabla, \nabla^*)$. When this structure is dually flat, namely the curvature tensors of ∇ and ∇^* are null $(\mathrm{R}(\nabla) \equiv 0$ and $\mathrm{R}^*(\nabla^*) \equiv 0)$, Amari and Nagaoka introduced a canonical divergence which is a Bregman divergence [5]. The canonical divergence has nice properties such as the generalized Pythagorean theorem and the geodesic projection theorem and it turns out to be of uppermost importance to define a canonical divergence in the general case. A first attempt to answer this fundamental issue is provided by Ay and Amari in [6] where a

canonical divergence for a general statistical manifold (M, g, ∇, ∇^*) is given by using the geodesic integration of the inverse exponential map. This one is understood as a *difference vector* that translates q to p for all $q, p \in M$ sufficiently close to each other.

To be more precise, the inverse exponential map supplies a generalization to M of the concept of difference vector in $\mathbb{R}^n$. In detail, let $p, q \in \mathbb{R}^n$, the difference between p and q is the vector $p - q$ pointing to p (see side (**A**) of Fig. 1). Then, the difference between p and q in M is provided by the inverse exponential map. In particular, given p, q suitably close in M, the difference vector from q to p is defined as (see (**B**) of Fig. 1)

$$\mathrm{X}_q(p) := \mathrm{X}(q, p) := \exp_q^{-1}(p) = \dot{\widetilde{\sigma}}(0), \tag{8}$$

where $\widetilde{\sigma}$ is the ∇-geodesic from q to p. Therefore, the divergence D introduced in [6] is defined as the path integral

$$\mathrm{D}(p, q) := \int_0^1 \langle \mathrm{X}_t(p), \dot{\widetilde{\sigma}}(t) \rangle_{\widetilde{\sigma}(t)} \, dt, \quad \mathrm{X}_t(p) := \mathrm{X}(\widetilde{\sigma}(t), p) := \exp_{\widetilde{\sigma}(t)}^{-1}(p), \tag{9}$$

where $\langle \cdot, \cdot \rangle_{\widetilde{\sigma}(t)}$ denotes the inner product induced by g on $\widetilde{\sigma}(t)$. After elementary computation Eq. (9) reduces to,

$$\mathrm{D}(p, q) = \int_0^1 t \, \|\dot{\sigma}(t)\|^2 \, dt, \tag{10}$$

where $\sigma(t)$ is the ∇-geodesic from p to q [6]. The divergence $\mathrm{D}(p, q)$ has nice properties such as the positivity and it reduces to the canonical divergence proposed by Amari and Nagaoka when the manifold M is dually flat. However, if we consider definition (9) for a general path γ, then $\mathrm{D}_\gamma(p, q)$ will be depending on γ. On the contrary, if the vector field $\mathrm{X}_t(p)$ is integrable, then $\mathrm{D}_\gamma(p, q) =: \mathrm{D}(p, q)$ turns out to be independent of the path from q to p.

2   The Recent Canonical Divergence

In this article, we discuss about a divergence function recently introduced in [7]. This turns out to be a generalization of the divergence introduced by Ay and Amari. The definition of the recent divergence (see below Eqs. (18), (19)) relies on an extended analysis of the intrinsic structure of the dual geometry of a general statistical manifold (M, g, ∇, ∇^*). In particular, we introduced a vector at q by modifying the definition (8) of the difference vector $\mathrm{X}(q, p)$. Consider $p, q \in M$ such that there exist both, a unique ∇-geodesic σ and a unique ∇^*-geodesic σ^*, connecting p with q. Moreover, let $\mathrm{X}_p(q) := \exp_p^{-1}(q) = \dot{\sigma}(0) \in T_pM$. Then, we ∇-parallel translate it along σ^* from p to q (see (**C**) of Fig. 1), and obtain

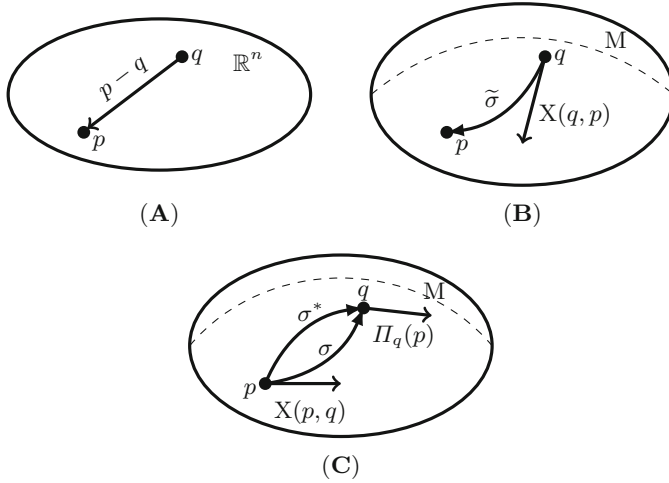$$\Pi_q(p) := \mathrm{P}_{\sigma^*} \mathrm{X}_p(q) \in T_qM. \tag{11}$$

Fig. 1. Illustration (**A**) of the difference vector in $\mathbb{R}^n$ pointing from q to p; and (**B**) the difference vector $\mathrm{X}(q,p) = \dot{\tilde{\sigma}}(0)$ as the inverse of the exponential map in q. The novel vector $\Pi_q(p)$ at q is described in (**C**) as the ∇ parallel transport of $\mathrm{X}(p,q) = \dot{\sigma}(0)$ along the ∇^*-geodesic σ^* from p to q.

(Note that $\Pi_q(p)$ corresponds to minus a difference vector.) Analogously, we introduce the dual vector of $\Pi_q(p)$ as the ∇^*-parallel transport of $\dot{\sigma}^*(0)$ along the ∇-geodesic σ,

$$\Pi_q^*(p) := \mathrm{P}_\sigma^* \mathrm{X}_p^*(q), \quad \mathrm{X}_p^*(q) := \overset{*}{\exp_p}^{-1}(q) = \dot{\sigma}^*(0), \tag{12}$$

where $\overset{*}{\exp}_p$ denotes the exponential map of the ∇^*-connection. A fundamental result obtained in [7] is that the sum of $\Pi_q(p)$ and $\Pi_q^*(p)$ is the gradient of a function $r_p(q) = r(p,q)$:

$$\Pi_q(p) + \Pi_q^*(p) = \mathrm{grad}_q r_p, \quad r(p,q) := \langle \exp_p^{-1}(q), \overset{*}{\exp_p}^{-1}(q) \rangle_p. \tag{13}$$

From here on, we refer to $r(p,q)$ as the pseudo-squared-distance between p and q. Note that the function $r(p,q)$, in general, is not symmetric in its arguments.

For $p \in \mathrm{M}$ fixed and q varying in M, we then obtain two vector fields whenever p and q are connected by a unique ∇-geodesic and a unique ∇^*-geodesic. Then, we can introduce two vector fields on an arbitrary path $\gamma : [0,1] \to \mathrm{M}$ connecting p and q in the following way. Let us firstly assume that for each $t \in [0,1]$ there exist a unique ∇-geodesic σ_t and a unique ∇^*-geodesic σ_t^* connecting p with $\gamma(t)$. Then, we define

$$\Pi_t(p) = \mathrm{P}_{\sigma_t^*} \mathrm{X}_p(t), \qquad \mathrm{X}_p(t) = \exp_p^{-1}(\gamma(t)) \tag{14}$$

$$\Pi_t^*(p) = \mathrm{P}_{\sigma_t}^* \mathrm{X}_p^*(t), \qquad \mathrm{X}_p^*(t) = \overset{*}{\exp_p}^{-1}(\gamma(t)). \tag{15}$$

From Eq. (13) we have that the sum

$$\int_0^1 \langle \Pi_t(p), \dot{\gamma}(t) \rangle_{\gamma(t)} \, dt + \int_0^1 \langle \Pi_t^*(p), \dot{\gamma}(t) \rangle_{\gamma(t)} \, dt = r(p,q) \qquad (16)$$

is independent of the particular path from p to q. This potential property of the pseudo-squared-distance $r(p,q)$ has been exploited for introducing a canonical divergence $\mathcal{D}$ as potential function of Π and a dual divergence $\mathcal{D}^*$ as potential function of Π^*, i.e. $\operatorname{grad} \mathcal{D} = \Pi$ and $\operatorname{grad} \mathcal{D}^* = \Pi^*$ [7]. However, this does not turn out in general. Indeed, on a general statistical manifold, the canonical divergence $\mathcal{D}$ and its dual function $\mathcal{D}^*$ are defined by the following unique and orthogonal decompositions,

$$\Pi = \operatorname{grad} \mathcal{D} + V, \qquad \Pi^* = \operatorname{grad} \mathcal{D}^* + V^*, \qquad (17)$$

where V is a vector field orthogonal to ∇-geodesics whereas V^* is orthogonal to ∇^*-geodesics *(we will prove later the orthogonality of these decompositions).*

When the vector field Π is integrated along the ∇-geodesic $\sigma(t)\,(0 \le t \le 1)$ connecting p with q, the recent canonical divergence assumes the following form

$$\mathcal{D}(p,q) = \int_0^1 \langle \Pi_t(p), \dot{\sigma}(t) \rangle_{\sigma(t)} \, \mathrm{d}t, \quad \Pi_t(p) = \mathrm{P}_{\sigma_t^*} \exp_p^{-1}(\sigma(t)), \qquad (18)$$

where $\sigma_t^*(s)\,(0 \le s \le 1)$ denotes the ∇^*-geodesic such that $\sigma_t^*(0) = p$ and $\sigma_t^*(1) = \sigma(t)$. On the contrary, if the vector field Π^* is integrated along the ∇^*-geodesic $\sigma^*(t)\,(0 \le t \le 1)$ connecting p with q, the dual divergence is given by

$$\mathcal{D}^*(p,q) = \int_0^1 \langle \Pi_t^*(p), \dot{\sigma}^*(t) \rangle_{\sigma^*(t)} \, \mathrm{d}t, \quad \Pi_t^*(p) = \mathrm{P}_{\sigma_t}^* \overset{*}{\exp}_p^{-1}(\sigma^*(t)), \qquad (19)$$

where $\sigma_t(s)\,(0 \le s \le 1)$ is the ∇-geodesic such that $\sigma_t(0) = p$ and $\sigma_t(1) = \sigma(t)$.

In this manuscript we review the relation of the canonical divergence $\mathcal{D}(p,q)$ of (18) to other divergence functions in Sect. 3. Finally, we outline in Sect. 4 the orthogonality and the symmetry properties of $\mathcal{D}$.

3 Comparison with Previous Divergence Functions

Given a general statistical manifold $(\mathrm{M}, \mathrm{g}, \nabla, \nabla^*)$, the basic requirement for a smooth function $\mathcal{D} : \mathrm{M} \times \mathrm{M} \to \mathbb{R}$ to be a divergence on M is its consistency with the dual structure $(\mathrm{g}, \nabla, \nabla^*)$ through Eqs. (5)–(6) and the positivity $\mathcal{D}(p,q) > 0$ for all $p, q \in \mathrm{M}$ sufficiently close to each other such that $p \ne q$. The recent canonical divergence (18) holds these properties [7].

In this section, we will show that the canonical divergence $\mathcal{D}$ can be interpreted as a generalization of the divergence D introduced by Ay and Amari. Indeed, we will see that these two divergences coincide on particular classes of

statistical manifolds. In order to achieve this result, we investigate some geometric properties of the vector field $\Pi_t(p)$ given by Eq. (14) aiming to split such a vector field in terms of the difference vector $X_t(p)$ given in Eq. (9). To be more precise, let us refer to Fig. 2 where the ∇-geodesic $\sigma(t)\,(0 \leq t \leq 1)$ connecting $\sigma(0) = p$ with $\sigma(1) = q$ is drawn. Then, for each $t \in [0,1]$ we can consider the ∇-geodesic $\sigma_t(s)\,(0 \leq s \leq 1)$ connecting p with $\sigma(t)$ and the ∇^*-geodesic $\sigma_t^*(s)\,(0 \leq s \leq 1)$ connecting p with $\sigma(t)$. The difference vector $X_t(p) = X(\sigma(t), p)$ at $\sigma(t)$ pointing to p is given in terms of the inverse exponential map by $X_t(p) := \exp_{\sigma(t)}^{-1}(p)$. Therefore, the opposite of $X_t(p)$ can be viewed as the ∇-parallel translation of $X_p(t) = \exp_p^{-1}(\sigma(t))$ along the ∇-geodesic σ_t, namely $-X_t(p) = P_{\sigma_t} X_p(t)$. Consider now the loop Σ_t based at p and given by first traveling from p to $\sigma(t)$ along the ∇^*-geodesic σ_t^* and then back from $\sigma(t)$ to p along the reverse of the ∇-geodesic σ_t. If Σ_t lies in a sufficiently small neighborhood of p, then [7]

$$P_{\Sigma_t} X_p(t) = X_p(t) + \mathcal{R}_{\Sigma_t}\left(X_p^*(t), X_p(t)\right),$$

where

$$\mathcal{R}_{\Sigma_t}\left(X_p^*(t), X_p(t)\right) := \int_{B_t} \frac{P\left[\mathcal{R}\left(X^*(t), X(t)\right) X(t)\right]}{\|X_p^*(t) \wedge X_p(t)\|}\, dA$$

with $X^*(t)$ and $X(t)$ being the ∇-parallel transport of $X_p^*(t)\,(:=\overset{*}{\exp}_p^{-1}(\sigma(t)))$ and $X_p(t)$, respectively, from p to each point of B_t along the unique ∇-geodesic joining them. Here, $\mathcal{R}$ is the curvature tensor of ∇, B_t denotes the disc defined by the curve Σ_t and $X_p^*(t)$, $X_p(t)$ are linearly independent. In addition, P within the integral denotes the ∇-parallel translation from each point in B_t to p along the unique ∇-geodesic segment joining them. Finally, by means of the property of the parallel transport, we obtain the following geometric relation between the vector $\Pi_t(p)$ and the opposite of the difference vector $X_t(p)$ [7],

$$\Pi_t(p) = P_{\sigma_t} X_p(t) + P_{\sigma_t}\left[\mathcal{R}_{\Sigma_t}\left(X_p^*(t), X_p(t)\right)\right]. \tag{20}$$

By noticing that $P_{\sigma_t} X_p(t) = \dot{\sigma}_t(1) = t\,\dot{\sigma}(t)$ and inserting Eq. (20) into the definition (18) of $\mathcal{D}$, we obtain

$$\mathcal{D}(p, q) = D(p, q) + \int_0^1 \langle P_{\sigma_t}\left[\mathcal{R}_{\Sigma_t}\left(X_p^*(t), X_p(t)\right)\right], \dot{\sigma}(t)\rangle_{\sigma(t)}\, dt, \tag{21}$$

where $D(p, q)$ is the divergence introduced in [6] and given by Eq. (10).

It is clear from Eq. (21) that particular conditions on the curvature tensor would lead to the required equivalence between $\mathcal{D}$ and D. Actually, in Information Geometry classes of statistical manifolds are characterized by the conditions on the curvature tensors of ∇ and ∇^* (see for instance Refs. [5], [8], [9]). In the Table 1 we can see the categories of statistical manifolds on which the canonical divergence $\mathcal{D}$ reduces to the divergence D introduced in [6]. A statistical manifold (M, g, ∇, ∇^*) is self-dual when $\nabla = \nabla^*$. Therefore, in this case M becomes a Riemannian manifold endowed with the Levi-Civita connection. Hence, the vectors $X_p(t)$ and $X_p^*(t)$ coincide for all $t \in [0, 1]$. Finally, the skew-symmetry of the
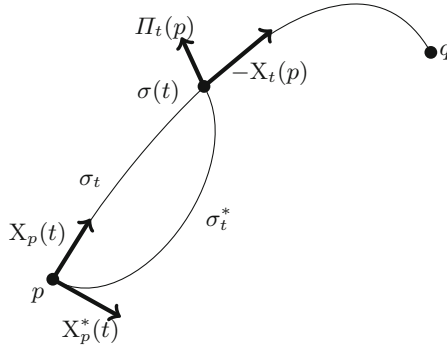
Fig. 2. The vector $\Pi_t(p)$ is obtained by ∇-parallel translating the vector $X_p(t) := \dot{\sigma}_t(0)$ along the ∇^*-geodesic σ_t^*. The opposite of the difference vector $X_t(p)$ at $\sigma(t)$ can be understood as the ∇-parallel translation of the vector $X_p(t)$ along the ∇-geodesic σ_t.

curvature tensor $\mathcal{R}$ yields the property $\mathcal{R}(X, X) = 0$ for any $X \in \mathcal{T}(M)$. When a manifold M is dually flat, it has a mutually dual affine coordinates $\{\boldsymbol{\theta}\}$, $\{\boldsymbol{\eta}\}$ and two potentials $\{\psi, \varphi\}$ such that $D(p, q) = \psi(\boldsymbol{\theta}_p) - \varphi(\boldsymbol{\eta}_q) + \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_q$ [6]. This claims that the canonical divergence $\mathcal{D}$ coincides with the canonical divergence of Bregman type introduced in [5] by Amari and Nagaoka on dually flat manifolds. The concept of a symmetric statistical manifold, that is the information geometric analogue to a symmetric space in Riemannian geometry, was introduced in [10]. There, the authors employed the following conditions on the curvature tensor, $\nabla \mathcal{R} = 0$ and $R(Y, X, X, X) := \langle \mathcal{R}(Y, X)X, X \rangle = 0$, in order to prove that their divergence function is independent of the particular path connecting any two points p, q sufficiently close to each other. The connection between the canonical divergence $\mathcal{D}$ and the divergence introduced by Henmi and Kobayashi is widely discussed in [7].

Table 1. The column on the left describes the category of statistical manifolds on which the canonical divergence $\mathcal{D}$ reduces to the divergence D. The column on the right shows the properties of the curvature tensor characterizing the corresponding manifolds and supplying the equivalence $\mathcal{D} \equiv D$ through the Eq. (21). Here, X, Y are vector fields on M, i.e. $X, Y \in \mathcal{T}(M)$.

Statistical manifold	Condition on $\mathcal{R}$
Self-dual	$\mathcal{R}(X, X) = 0$
Dually flat	$\mathcal{R} \equiv 0$
Symmetric	$\nabla \mathcal{R} \equiv 0, R(Y, X, X, X) = 0$

To summarize, Table 1 describes, from the top to the bottom, the statistical manifolds ordered from less generality to more generality where the equivalence

between $\mathcal{D}$ and D is achieved. In this view, we can consider $\mathcal{D}$ as an extension of the divergence D to the very general statistical manifold (M, g, ∇, ∇^*).

Since Eq. (2) we know that in general a divergence function is not symmetric in its argument. However, the symmetry property owned by the canonical divergence of Bregman type on dually flat manifolds, namely $D(q, p) = D^*(p, q)$, shows the way for the further investigation about symmetry properties of $\mathcal{D}$ in the very general context of Information Geometry.

4   General Properties of $\mathcal{D}$

In this section we prove that the decompositions given in Eq. (17) are orthogonal ones. To this aim, we rely on the *gradient–based approach* to divergence which was introduced in [6] and further developed in [3]. This approach has led to the following decompositions of $\Pi_q(p)$ and $\Pi_q^*(p)$ in terms of the canonical divergence gradient and its dual [7],

$$\Pi_q(p) = \mathrm{grad}_q \mathcal{D}_p(q) + V_q, \quad \langle V_q, \dot{\sigma}(1) \rangle_q = 0 \tag{22}$$

$$\Pi^* q(p) = \mathrm{grad}_q \mathcal{D}_p^*(q) + V_q^*, \quad \langle V_q^*, \dot{\sigma}^*(1) \rangle_{\gamma(t)} = 0, \tag{23}$$

where $\sigma(t), \sigma^*(t), (0 \le t \le 1)$ are the ∇ and ∇^* geodesics, respectively, from p to q. In order to prove that these decompositions are orthogonal ones, we exploit the theory of minimum contrast geometry by Eguchi [1]. This allows us to show that $\mathrm{grad}_q \mathcal{D}_p$ is parallel to the tangent vector of the ∇-geodesic starting from p whereas $\mathrm{grad}_q \mathcal{D}_p^*$ is parallel to the tangent vector of the ∇^*-geodesic starting from p. Thus, let us consider the level sets of $\mathcal{D}_p$ and $\mathcal{D}_p^*$:

$$\mathcal{H}(\kappa) = \{q \in M \,|\, \mathcal{D}_p(q) = \kappa\}, \quad \mathcal{H}^*(\kappa) = \{q \in M \,|\, \mathcal{D}_p^*(q) = \kappa\}. \tag{24}$$

Then to each $q \in \mathcal{H}(\kappa)$ we can define the minimum contrast leaf of $\mathcal{D}$ at q [2]:

$$L_q := \{p \in M \,|\, \mathcal{D}(p, q) = \min_{q' \in \mathcal{H}} \mathcal{D}(p, q')\}. \tag{25}$$

Let us now fix q. Since q minimizes the set $\{\mathcal{D}(p, q) \,|\, p \in L_q , q \in \mathcal{H}\}$ it follows that the derivative of $\mathcal{D}(p, q)$ at q along any direction U tangent to $\mathcal{H}$ vanishes, namely

$$\partial_U' \mathcal{D}(p, q) = 0, \quad \forall U \in T_q \mathcal{H}, \, p \in L_q,$$

where ∂_U' denotes the derivative at q along the direction U. Thus we have that $\langle U, \Xi \rangle_q = 0$ for all $U \in T_q \mathcal{H}$ and $\Xi \in T_q L_q$, or equivalently that the tangent space of L_q coincides with the normal space of $\mathcal{H}$ at q (see Fig. 3 for a cross-reference). In addition, by taking derivatives at q along directions Ξ, Υ normal to $\mathcal{H}$ we have that [1]

$$g\left(II(\Upsilon, \Xi), U\right) := -\partial_\Xi \partial_\Upsilon \partial_U' \mathcal{D}(p, q)\big|_{p=q} = 0, \quad \forall \Xi, \Upsilon \in T_q L_q, \, U \in T_q \mathcal{H}, \tag{26}$$

where the first relation defines the second fundamental tensor II with respect to the ∇-connection. This implies that the second fundamental tensor with respect

to ∇ for L_q vanishes at q. Therefore, according to the well-known Gauss formula [11]

$$\nabla_\Xi \Upsilon = \nabla_\Xi^{L_q} \Upsilon + II(\Xi, \Upsilon)$$

we can see from Eq. (26) that the family of all curves which are orthogonal to the level set $\mathcal{H}$ are all ∇-geodesics ending at q (with a suitable choice of the parameter). Since $\mathrm{grad}_q \mathcal{D}_p$ is orthogonal to $\mathcal{H}$ at q, we finally obtain that $\mathrm{grad}_q \mathcal{D}_p$ is parallel to the ∇-velocity $\dot{\sigma}(1)$. This proves that Eq. (22) provides an orthogonal decomposition.

Analogously, we have that the family of all curves which are orthogonal to the level set $\mathcal{H}^*$ are all ∇^*-geodesics ending at q (with a suitable choice of the parameter). Since $\mathcal{H}^*$ is the hypersurface of constant $\mathcal{D}^*$ it turns out that $\mathrm{grad}_q \mathcal{D}^*$ is orthogonal to $\mathcal{H}^*$ at q. Therefore, we can conclude that the decomposition in Eq. (23) is an orthogonal one.
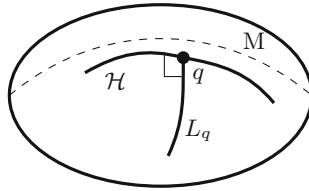

Fig. 3. According to the Eguchi's theory [1], the minimum contrast leaf L_q turns out to be orthogonal at q to the level-set $\mathcal{H}$ generated by the canonical divergence $\mathcal{D}$.

In general, the canonical divergence $\mathcal{D}(p, q)$ is not symmetric in its argument, i.e. $\mathcal{D}(p, q) \neq \mathcal{D}(q, p)$. However, the Kullback-Leibler divergence $\mathrm{K}(p, q)$ and the canonical divergence $\mathrm{D}(p, q)$ of Bregman type suggest a symmetry property that a divergence should satisfy, namely $\mathrm{K}^*(p, q) = \mathrm{K}(q, p)$ and $\mathrm{D}^*(p, q) = \mathrm{D}(q, p)$ [5]. In order to answer what the relation of $\mathcal{D}(p, q)$ and $\mathcal{D}(q, p)$ is, let us set $\widetilde{\mathcal{D}}(p, q) := \mathcal{D}(q, p)$. Consider the hypersurface $\widetilde{\mathcal{H}}$ of constant $\widetilde{\mathcal{D}}$ given by $\widetilde{\mathcal{H}} = \{q \in \mathrm{M} \mid \widetilde{\mathcal{D}}(p, q) = \kappa\}$. Then, by noticing that $-\partial_i \partial_j \partial_k' \widetilde{\mathcal{D}}(p, q) = \Gamma_{ijk}^*$ we have that $\mathrm{grad}_q \widetilde{\mathcal{D}}_p$ is parallel to $\dot{\sigma}^*(1)$ as well as $\mathrm{grad}_q \mathcal{D}_p^*(q)$. Here, $\sigma^*(t)$ is the ∇^*-geodesic from p to q. Recall that, since both functions, $\widetilde{\mathcal{D}}_p(q)$ and $\mathcal{D}_p^*(q)$, are divergences, we have that $\widetilde{\mathcal{D}}(p, q) \geq 0$, $\mathcal{D}^*(p, q) \geq 0$ and $\widetilde{\mathcal{D}}(p, q) = 0$, $\mathcal{D}^*(p, q) = 0$ if and only if $p = q$. Furthermore, the gradient flows of these functions are identical as a family of curves. This implies that, let p fixed, both, $\mathrm{grad}_q \widetilde{\mathcal{D}}_p(q)$ and $\mathrm{grad}_q \mathcal{D}_p^*(q)$, point from q to the opposite direction, along the curve $\sigma^*(t)$, with respect to p. Therefore, we can find a positive constant depending on p and q such that $\mathrm{grad}_q \widetilde{\mathcal{D}}_p = c(p, q)\, \mathrm{grad}_q \mathcal{D}_p^*$. This proves that there exists a function $f : [0, \infty] \to \mathbb{R}^+$ satisfying $f(0) = 0$ and $f'(0) > 0$ such that [7]

$$\mathcal{D}(q, p) = f\left(\mathcal{D}^*(p, q)\right).$$

Though this relation holds for a very general statistical manifold (M, g, ∇, ∇^*), this result is still not satisfactory. However, it allows to face the issue of the symmetry property originally established by the canonical divergence of Bregman type on dually flat manifolds. By analysing the cases where it holds true, we may conjecture through our approach that it is satisfied whenever $r(p, q) = \mathcal{D}(p, q) + \mathcal{D}^*(p, q)$. Or equivalently, whenever the pseudo-squared-distance is symmetric in its argument [7]. Forthcoming investigation will address this fundamental issue of Information Geometry.

References

1. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
2. Eguchi, S.: A differential geometric approach to statistical inference on the basis of contrast functions. Hiroshima Math. J. **15**, 341–391 (1985)
3. Ay, N., Jost, J., Van Le, H., Schwachhöfer, L.: Information Geometry, 1st edn. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56478-4
4. Matumoto, T.: Any statistical manifold has a contrast function-On the C3-functions taking the minimum at the diagonal of the product manifold. Hiroshima Math. J. **23**, 327–332 (1993)
5. Amari, S.-I., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
6. Ay, N., Amari, S.-I.: A novel approach to canonical divergences within information geometry. Entropy **17**, 8111–8129 (2015)
7. Felice, D., Ay, N.: Towards a canonical divergence within information geometry. arXiv:1806.11363 [math.DG] (2018)
8. Lauritzen, S.L.: Differential geometry in statistical inference. Lect. Notes-Monogr. Ser. **10**, 163–218 (1987)
9. Zhang, J.: A note on curvature of α-connections of a statistical manifold. Ann. Inst. Stat. Math. **59**, 161–170 (2007)
10. Henmi, M., Kobayashi, R.: Hooke's law in statistical manifolds and divergences. Nagoya Math. J. **159**, 1–24 (2000)
11. Lee, M.J.: Riemannian Manifolds: An Introduction to Curvature, 1st edn. Springer, New York (1997). https://doi.org/10.1007/b98852