CrossMark

Logarithmic divergences from optimal transport and Rényi geometry

Ting-Kam Leonard Wong[1]

Abstract

Divergences, also known as contrast functions, are distance-like quantities defined on manifolds of non-negative or probability measures. Using the duality in optimal transport, we introduce and study the one-parameter family of $L^{(\pm\alpha)}$-divergences. They extrapolate between the Bregman divergence corresponding to the Euclidean quadratic cost, and the L-divergence introduced by Pal and the author in connection with portfolio theory and a logarithmic cost function. They admit natural generalizations of exponential family that are closely related to the α-family and q-exponential family. In particular, the $L^{(\pm\alpha)}$-divergences of the corresponding potential functions are Rényi divergences. Using this unified framework we prove that the induced geometries are dually projectively flat with constant sectional curvatures, and a generalized Pythagorean theorem holds true. Conversely, we show that if a statistical manifold is dually projectively flat with constant curvature $\pm\alpha$ with $\alpha > 0$, then it is locally induced by an $L^{(\mp\alpha)}$-divergence. We define in this context a canonical divergence which extends the one for dually flat manifolds.

Keywords Statistical manifold · Optimal transport · Exponential concavity and convexity · Spaces of constant curvature · Projective flatness · Rényi divergence, α-divergence

1 Introduction

In this paper we study geometric properties of statistical manifolds defined by solutions to some optimal transport problems. We show that they characterize dually projectively flat manifolds with constant curvatures and provide natural geometries to generalizations of exponential family. Before stating our main results let us describe the background and motivations of our study.

✉ Ting-Kam Leonard Wong
tkl.wong@utoronto.ca

1   University of Toronto, Toronto, Canada

1.1 Background

Information geometry studies manifolds of probability distributions and measures using geometric ideas. For modern introductions to its theory and applications we refer the reader to the recent monographs [3,9]. Typically the geometry is induced by a divergence which is a distance-like quantity $\mathbf{D}[q:p] \geq 0$ defined on the underlying manifold $\mathcal{M}$. Of particular importance is the Bregman divergence. Given a differentiable concave function $\varphi(\xi)$ defined on a convex domain Ω in $\mathbb{R}^d$, the Bregman divergence of φ is defined for $\xi, \xi' \in \Omega$ by

$$\mathbf{D}^{(0+)}\left[\xi:\xi'\right] = D\varphi(\xi')\cdot(\xi-\xi') - \left(\varphi(\xi) - \varphi(\xi')\right), \tag{1}$$

where D is the Euclidean gradient and $a \cdot b$ is the dot product (the meaning of the superscript 0+ will become clear in Sect. 1.2). For example, if $\mathcal{M}$ is an exponential family of probability densities where ξ is the natural parameter, the relative entropy, also known as the Kullback–Leibler divergence, can be expressed as the Bregman divergence of the cumulant generating function $\varphi(\xi)$ (which is convex, so in (1) we may consider $-\varphi$ or use $\mathbf{D}^{(0-)}$ to be defined in Sect. 3). The differential geometry of Bregman divergence was first studied by Nagaoka and Amari in [27].

Following the general framework established by Eguchi [16,17], the geometry induced on $\mathcal{M}$ by a given divergence $\mathbf{D}[\cdot:\cdot]$ consists of a Riemannian metric g and a dual pair (∇, ∇^*) of torsion-free affine connections (see Definition 10). We call the quadruplet $(\mathcal{M}, g, \nabla, \nabla^*)$ a dualistic structure or a statistical manifold. The Riemannian metric is given by the quadratic approximation of $\mathbf{D}[q:p]$ when $q \approx p$, whereas the two affine connections capture higher order local properties and allow us to define primal and dual geodesics that are in some sense compatible with the divergence. In contrast with usual Riemannian geometry where the Levi-Civita connection is the canonical connection defined by the metric, in information geometry we usually use a pair of connections (∇, ∇^*) to reflect the asymmetry of the divergence: generally we have $\mathbf{D}[p:q] \neq \mathbf{D}[q:p]$. When the divergence is symmetric (for example when $\mathbf{D}[q:p] = \frac{1}{2}d(p,q)^2$ where d is a metric) both ∇ and ∇^* coincide with the Levi-Civita connection. From this geometric perspective, the Bregman divergence is fundamental in the sense that it is the canonical divergence which generates a dually flat geometry, i.e., both the primal and dual connections ∇ and ∇^* have zero curvature (see for example [3, Sect. 6.6] and [9, Sect. 4.2]; this is also a limiting case of Theorem 3). Moreover, the Bregman divergence satisfies a generalized Pythagorean theorem which enables explicit computation and geometric interpretation of projections based on the divergence. By now these results are well-known and numerous applications can be found in the two monographs cited above.

Motivated by portfolio theory, in a series of papers [31–34,41,42] we introduced and studied the L-divergence (L stands for logarithmic) defined by

$$\mathbf{D}^{(1)}\left[\xi:\xi'\right] = \log(1 + D\varphi(\xi')\cdot(\xi-\xi')) - \left(\varphi(\xi) - \varphi(\xi')\right), \tag{2}$$

where now the potential function φ is exponentially concave, i.e, $\Phi = e^\varphi$ is concave. Being the logarithm of the concave function Φ, φ is itself concave and the L-divergence

has a logarithmic correction term that captures the extra concavity. An important example of L-divergence is the "excess growth rate" defined for probability vectors $p = (p^i), q = (q^i)$ (here $i \in \{0, 1, \ldots, d\}$) by

$$\mathbf{D}\left[q : p\right] = \log \left(\sum_{i=0}^{d} \frac{1}{d+1} \frac{q^i}{p^i} \right) - \sum_{i=0}^{d} \frac{1}{d+1} \log \frac{q^i}{p^i}, \qquad (3)$$

where $\varphi(p) = \sum_{i=0}^{d} \frac{1}{d+1} \log p^i$. Exponentially concave functions appear naturally in many recent applications in analysis, probability and statistics. In particular, let us mention the paper [18] on Bochner's inequality on metric measure spaces. Other examples can be found in the references of [33].

In [33] we established the following results where the underlying manifold is the open unit simplex:

(i) The dualistic structure induced by the L-divergence (2) is dually projectively flat with constant curvature -1, and the generalized Pythagorean theorem (which has a financial interpretation in terms of optimal trading frequency) holds true.

(ii) The L-divergence can be expressed in terms of the optimal transport map with respect to a logarithmic cost function, whereas the Bregman divergence corresponds to the classical quadratic cost. Furthermore, displacement interpolations for these transport problems are related to the dual geodesics of the induced dualistic structure.

It is natural to ask if there is a unified framework that covers both Bregman and L-divergences. Moreover, can we characterize statistical manifolds that satisfy these (very strong) geometric properties? On the other hand, Amari and Nagaoka [2] considered α-families of probability distributions and the corresponding α-divergences that generalize the exponential family and the Kullback–Leibler divergence. We will see that these questions are intimately related.

The connection between optimal transport and information geometry is a recent topic that has started to receive serious attention. For example, the recent paper [4] studies the manifold of entropy-relaxed optimal transport plans and defines a divergence which interpolates between the Kullback–Leibler divergence and the Wasserstein metric (optimal transport cost). Dynamic approach are considered in [14] and the recent papers [13,24]. Also see [30] which builds on the ideas of [33] and embeds a large family of optimal transport problems in a generalized simplex.

1.2 Summary of main results

The key objects of study in this paper are the $L^{(\pm\alpha)}$-divergences defined for $\alpha > 0$. The $L^{(\alpha)}$-divergence is defined by

$$\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = \frac{1}{\alpha} \log(1 + \alpha \nabla\varphi(\xi') \cdot (\xi - \xi')) - \left(\varphi(\xi) - \varphi(\xi')\right), \qquad (4)$$

where the function φ is α-exponentially concave, i.e., $e^{\alpha\varphi}$ is concave. To cover the case of positive curvature, we will also consider a local $L^{(-\alpha)}$-divergence $\mathbf{D}^{(-\alpha)}$ where φ is

α-exponentially convex, i.e., $e^{\alpha\varphi}$ is convex. When $\alpha = 1$ we have the $L^{(1)}$-divergence (2), and when $\alpha \to 0^+$ it reduces to the Bregman divergence (1) of a concave function ($\mathbf{D}^{(0-)}$ is the Bregman divergence of a convex function). Based on financial arguments, the $L^{(\alpha)}$-divergence was introduced recently in [42] as the canonical interpolation between the Bregman and L-divergences. In this paper we consider also the $L^{(-\alpha)}$-divergence and show that they are of fundamental importance. We note that there exists other natural extrapolations of the Bregman divergence.[1] For example, Zhang [43] (also see [29,44,45]) introduced a parameterized family $\mathcal{D}_\Phi^{(\alpha)}$ of divergences for a given convex function Φ, which includes the Bregman and α-divergences among others, and introduced the concepts of referential duality and representational duality. The biduality of divergence also plays a role in this paper (see Corollary 1).

While we worked on the open unit simplex in our previous paper [33], here we consider a general open convex domain Ω in $\mathbb{R}^d$. This answers a question asked in [33, Sect. 1] about extending the results to general domains. By a translation if necessary, we may assume that $0 \in \overline{\Omega}$ [this is used for the normalization of the function $c^{(\alpha)}$ in (5)]. Nevertheless, we note that exponential concavity/convexity impose certain restrictions on the domains of φ and the $L^{(-\alpha)}$-divergence.

In Sects. 2 and 3 we prove the following result which connects the $L^{(\pm\alpha)}$-divergence with optimal transport and generalizes the classical Legendre duality for Bregman divergence.

Theorem 1 *For $\alpha > 0$, consider the function*

$$c^{(\alpha)}(x, y) = \frac{1}{\alpha} \log\left(1 + \alpha x \cdot y\right), \tag{5}$$

where $x \cdot y$ is the Euclidean dot product. Then the $L^{(\alpha)}$-divergence (28) admits the self-dual representation

$$\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = c^{(\alpha)}(\xi, \eta') - \varphi(\xi) - \psi(\eta'), \tag{6}$$

where ψ is the α-conjugate of φ and is (locally) α-exponentially concave, and $\eta = D^{(\alpha)}\varphi(\xi)$ is the α-gradient which generalizes the Legendre transformation.

Similarly, when φ is α-exponentially convex, the $L^{(-\alpha)}$-divergence has the self-dual representation

$$\mathbf{D}^{(-\alpha)}\left[\xi : \xi'\right] = \varphi(\xi) + \psi(\eta') - c^{(\alpha)}(\xi, \eta'), \tag{7}$$

where now φ and ψ are α-exponentially convex.

Precise statements corresponding to Theorem 1 are given in Theorems 6, 8, 9 and 11. The self-dual expressions (6) and (7) are motivated by optimal transport which identifies the dual coordinate system η as the image of ξ under the optimal transport map. This idea comes from our previous paper [33].

Motivated by the relationships between Bregman divergence and exponential family (see [10]), in Sect. 4 we consider generalizations of exponential family that are closely

---

related to the α-family and q-exponential family [2,5,28]. We show that the analogue of the cumulant generating function is α-exponentially concave/convex (Propositions 2, 3), and the corresponding $L^{(\pm\alpha)}$-divergence is the Rényi divergence (Theorem 13). Moreover, the dual function is the Rényi entropy. Our results thus provide a new approach to the geometry of Rényi and α-divergences. When $\alpha \to 0$ we recover the dually flat geometry of exponential family. The case of finite simplex is discussed in Sect. 6.4.

In Sects. 5 and 6 we study the statistical manifold $(\mathcal{M} = \Omega, g, \nabla, \nabla^*)$ induced by the $L^{(\pm\alpha)}$-diverence. Our approach, based on the duality in Theorem 1, gives a unified treatment covering both the Bregman and L-divergences and simplifies the proofs in [33]. The following result summarizes Theorems 14, 15 and 16.

Theorem 2 *For $\alpha > 0$, the dualistic structure induced by the $L^{(\pm\alpha)}$-divergence defined on an open convex set $\Omega \subset \mathbb{R}^d$ with $d \geq 2$ is dually projectively flat, has constant sectional curvature $\mp\alpha$, and the generalized Pythagorean theorem holds (when $\alpha \to 0^+$ we reduce to the dually flat case).*

By dual projective flatness, we mean that there exist 'affine' coordinate systems (ξ and η) under which the primal and dual geodesics are, respectively, straight lines up to time reparameterizations (for the precise technical statement see Definition 11). Since the properties in Theorem 2 are very strong and spaces of constant curvatures play fundamental roles in differential geometry, it is natural to ask if Theorem 2 has a converse. Indeed there is one and it will be proved in Sect. 7 (see Theorems 18, 19).

Theorem 3 *Consider a dualistic structure $(\mathcal{M}, g, \nabla, \nabla^*)$ which is dually projectively flat and has constant curvature $-\alpha < 0$. Then locally there exist affine coordinate systems ξ and η for ∇ and ∇^* respectively, and α-exponentially concave functions $\varphi(\xi)$ and $\psi(\eta)$ which satisfy the generalized Fenchel identity*

$$\varphi(\xi) + \psi(\eta) \equiv c^{(\alpha)}(\xi, \eta).$$

Moreover, the self-dual representation (6) defines locally a canonical divergence which induces the given dualistic structure. Analogous statements hold when the curvature is $\alpha > 0$.

Our results establish the $L^{(\pm\alpha)}$-divergence as the canonical divergence for a dually projectively flat statistical manifold with constant curvature, thus generalizing the Bregman divergence for dually flat manifolds (see [21] for another characterization based on affine differential geometry). Recently, Ay and Amari [8] (also see [19]) defined a canonical divergence for an arbitrary statistical manifold and showed that on a dually flat manifold (i.e., $\alpha = 0$) it reduces to the Bregman divergence. It is interesting to know if their canonical divergence is consistent with our $L^{(\pm\alpha)}$-divergences. Since the $L^{(\pm\alpha)}$-divergences have properties analogous to those of the Bregman divergence, explicit computations are tractable and a natural question is to study how they perform in applied problems such as clustering [10] and statistical estimation. Projectively flat connections, statistical manifolds with constant curvatures and relations with affine differential geometry have been studied in the literature; see in particular [12,15,21,

22,25]. Using ideas of optimal transport and the $L^{(\pm\alpha)}$-divergence, we consider these properties as a whole and are able to give a new and elegant characterization of the geometry.

In optimal transport, it is well known that the quadratic cost, which corresponds in our setting to the Bregman divergence, is intimately related to Brownian motion and the heat equation (via the Wasserstein gradient flow of entropy, see [7,20]) as well as large deviations [1] and the Schrödinger problem [23]. A programme suggested in [33] is to study the analogues for our logarithmic cost functions, and in [34] we took a first step by studying a multiplicative Schrödinger problem. We plan to address some of these questions in future research. Another intriguing problem is to connect other information-geometric structures (such as curvature and the ρ-τ embedding, see [43]) with optimal transport.

2 Divergences induced by optimal transport maps

A major theme of this paper is that the $L^{(\pm\alpha)}$-divergences are intimated related to optimal transport maps. As a motivating example we first consider the classical Bregman divergence which corresponds to the quadratic cost. Then we introduce the general framework of c-divergence following the ideas of [33].

2.1 The Monge–Kantorovich optimal transport problem

We begin by introducing some basic terminologies of optimal transport. For further details we refer the reader to standard references such as [6,37,39,40]. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces (i.e., complete and separable metric spaces) interpreted respectively as the source and target spaces of the optimal transport problem. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function. Given Borel probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the Monge–Kantorovich optimal transport problem is

$$\inf_{\gamma} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y), \tag{8}$$

where the infimum is taken over joint distributions $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ whose marginals are μ and ν respectively. We call γ a coupling of the pair (μ, ν) and write $\gamma \in \Pi(\mu, \nu)$. If γ attains the infimum in (8), we say that it is an optimal coupling. For certain cost functions and under suitable conditions on μ and ν, the optimal coupling has the form

$$\gamma = (\text{Id} \times T)_{\#} \mu$$

for some measurable map $T : \mathcal{X} \to \mathcal{Y}$. Here we use $F_{\#}\mu = \mu \circ F^{-1}$ to denote the pushfoward of a measure. In this case the optimal coupling is deterministic, and we call T an optimal transport map.

Remark 1 Note that if we replace the cost function c by

$$\tilde{c}(x, y) = c(x, y) + h(x) + k(y), \tag{9}$$

where h and k and real-valued functions on $\mathcal{X}$ and $\mathcal{Y}$ respectively, then for any coupling $\gamma \in \Pi(\mu, \nu)$ we have

$$\int \tilde{c} d\gamma = \int c d\gamma + \int h d\mu + \int k d\nu.$$

The last two integrals, if they exist, are determined once μ and ν are fixed and are independent of the coupling γ. This means that we are free to modify the cost function in the manner of (9) without changing the optimal couplings.

2.2 Quadratic cost and Bregman divergence

A fundamental example is where $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ ($d \geq 1$) and

$$c(x, y) = \frac{1}{2}\|x - y\|^2 = \frac{1}{2}\sum_{i=1}^{d}(x^i - y^i)^2 \tag{10}$$

is the quadratic cost (where we write $x = (x^1, \ldots, x^d)$). Suppose $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have finite second moments, i.e.,

$$\int \|x\|^2 d\mu(x) < \infty, \quad \int \|y\|^2 d\nu(y) < \infty,$$

and suppose μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. By Brenier's theorem [11], there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that its gradient $D\varphi$ (which is defined μ-a.e.) pushforwards μ to ν. Moreover, the deterministic coupling

$$\gamma = (\text{Id} \times \nabla\varphi)_\# \mu$$

solves the Monge–Kantorovich problem. In fact, it can be shown that the optimal transport map $T(x) = D\varphi(x)$, which is a Legendre trasnformation, is unique μ-a.e.

To see how the Bregman divergence comes into play, suppose that φ is C^2 and $D^2\varphi$ (the Hessian) is strictly positive definite. Then the Legendre transform $y = T(x)$ is a diffeomorphism (see [3, Chapter 1]), and we have $x = D\varphi^*(y)$ and φ^* is the convex conjugate of φ. By Fenchel's inequality, for any x and y' we have

$$\varphi(x) + \varphi^*(y') \geq x \cdot y', \tag{11}$$

and equality holds if and only if $y' = D\varphi(x)$. Intuitively, the inequality (11) quantifies the inefficiency of coupling $x' = D\varphi^*(y')$ with $y = D\varphi(x)$ when compared to y'.

Since $\varphi^*(y') = x' \cdot y' - \varphi(x')$, we may rearrange (11) to get

$$\varphi(x) - \varphi(x') - D\varphi(x') \cdot (x - x') \geq 0,$$

which is nothing but the Bregman divergence of φ.

2.3 *c*-divergence

Now we show that this idea can be formulated in an abstract framework using generalized concepts of convex analysis. These concepts were introduced by Moreau in [26] and later become fundamental in optimal transport (see [35, Sect. 3.3] for a discussion). In the following definition we fix a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Definition 1 Let $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $g : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$.

(i) The *c*-transforms of f and g are defined respectively by

$$\begin{aligned} f^c(y) &= \inf_{x \in \mathcal{X}} \left(c(x, y) - f(x) \right), \quad y \in \mathcal{Y}, \\ g^c(x) &= \inf_{y \in \mathcal{Y}} \left(c(x, y) - g(x) \right), \quad x \in \mathcal{X}. \end{aligned} \tag{12}$$

(ii) We say that f (respectively g) is *c*-concave if $f^{cc} = f$ (respectively $g^{cc} = g$).
(iii) If f and g are *c*-concave, their *c*-superdifferentials are defined by

$$\begin{aligned} \partial^c f &= \{(x, y) \in \mathcal{X} \times \mathcal{Y} : f(x) + f^c(y) = c(x, y)\}, \\ \partial^c g &= \{(x, y) \in \mathcal{X} \times \mathcal{Y} : g^c(x) + g(y) = c(x, y)\}. \end{aligned} \tag{13}$$

If f is *c*-concave on $\mathcal{X}$, then

$$f(x) + f^c(y) \leq c(x, y) \tag{14}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and equality holds if and only if $(x, y) \in \partial^c f$. We call (14) the (generalized) Fenchel inequality (or identity when equality holds) which generalizes (11). The following result (see for example [6, Theorem 2.13]) is sometimes called the Fundamental Theorem of Optimal Transport.

Theorem 4 *Suppose the cost function c is continuous and bounded below, and there exist $a \in L^1(\mu)$ and $b \in L^1(\nu)$ such that $c(x, y) \leq a(x) + b(y)$. Then, a coupling γ of the pair (μ, ν) solves the Monge–Kantorovich problem (8) if and only if there exists a c-concave function f on $\mathcal{X}$ such that $\mathrm{supp}(\gamma) \subset \partial^c f$.*

By Theorem 4, a *c*-concave function f on $\mathcal{X}$ (or, rather, its *c*-superdifferential) can be regarded as encoding the solution to an optimal transport problem.

Let f is a *c*-concave function on $\mathcal{X}$. Assume that f is *c*-differentiable in the following sense: for each $x \in \mathcal{X}$ there is a unique element $y \in \mathcal{Y}$ such that $(x, y) \in \partial^c f$. We call $y = D^c f(x)$ the *c*-gradient of f. This is a rather strong assumption but is satisfied for the cost functions considered in this paper. By Theorem 4 we may

regard $D^c f : \mathcal{X} \rightarrow \mathcal{Y}$ as an optimal transport map. For a detailed discussion of the smoothness properties of optimal transport maps see [40, Chapter 11].

Using the definition of c-concavity, for any $x, x' \in \mathcal{X}$ we have

$$f(x') + c(x, y') - c(x', y') \geq f(x), \quad y' = D^c f(x'). \tag{15}$$

Analogous definitions and results hold for a c-concave function g on $\mathcal{Y}$. The inequality (15) motivates the definition of c-divergence (a more suggestive name may be transport divergence) which was first introduced in [33].

Definition 2 (*c-divergence*) Let f be c-concave and c-differentiable on $\mathcal{X}$. The c-divergence of f is the functional $\mathbf{D}_f : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ defined by

$$\mathbf{D}_f \left[x : x' \right] = c(x, y') - c(x', y') - (f(x) - f(x')), \quad x, x' \in \mathcal{X}, \quad y' = D^c f(x'). \tag{16}$$

If g is a c-differentiable c-concave function on $\mathcal{Y}$, we define the dual c-divergence $\mathbf{D}_g : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$ by

$$\mathbf{D}_g^* \left[y : y' \right] = c(x', y) - c(x', y') - (g(y) - g(y')), \quad y, y' \in \mathcal{Y}, \quad x' = D^c g(y'). \tag{17}$$

Theorem 5 (self-dual representations) *Let* $y' = D^c f(x')$. *Then*

$$\mathbf{D}_f \left[x : x' \right] = c(x, y') - f(x) - f^c(y'). \tag{18}$$

Similarly, if $x = D^c g(y)$, *then*

$$\mathbf{D}_g^* \left[y : y' \right] = c(x', y) - g(y) - g^c(x'). \tag{19}$$

Proof By the generalized Fenchel identity (14), we have

$$f(x') + f^c(y') = c(x', y') \Rightarrow f^c(y') = c(x', y') - f(x').$$

It follows that

$$\begin{aligned} \mathbf{D}_f \left[x : x' \right] &= c(x, y') - c(x', y') - (f(x) - f(x')) \\ &= c(x, y') - f(x) - f^c(y'). \end{aligned}$$

This gives (18). The proof of the second identity (19) is similar. □

Corollary 1 (Duality of c-divergence) *Suppose* f^c *is c-differentiable on the range of* $D^c f$. *Let* $y = D^c f(x)$ *and* $y' = D^c f(x')$. *Then*

$$\mathbf{D}_f \left[x' : x \right] = \mathbf{D}_{f^c}^* \left[y : y' \right]. \tag{20}$$

This result can be viewed as the "biduality"—in the language of Zhang [43] – of the c-divergence. Namely, using the c-conjugate f^c and the dual coordinates $y = D^c f(x)$ (representational duality) in the $L^{(\alpha)}$-divergence is equivalent to interchanging the order of the arguments (referential duality). See [29,44,45] for more discussions about the two dualities.

Remark 2 On the other hand, given an α-exponentially concave function φ, we don't have a natural way to associate it to a 'dual' α-exponentially convex function, say ψ, by some conjugate operation. Under our framework, the $L^{(\alpha)}$-divergence and the $L^{(-\alpha)}$-divergence are not related or dual to each other. In particular, the parameter α of our $L^{(\pm\alpha)}$-divergence does not play the same role as the α in say the α-divergence or Zhang's $\mathcal{D}_\Phi^{(\alpha)}$ divergence [43].

Example 1 Suppose $c(x, y) = \frac{1}{2}\|x - y\|^2$ is the quadratic cost on $\mathbb{R}^d$. It is easy to show that f is c-concave if and only if $\varphi(x) = \frac{1}{2}\|x\|^2 - f(x)$ is convex. The c-gradient of f is $D^c f = D\varphi$, and the c-divergence $\mathbf{D}_f$ is the Bregman divergence of φ.

2.4 A logarithmic cost function

Now we introduce a logarithmic cost which leads to the $L^{(\pm\alpha)}$-divergences. Because of the logarithm, to state the result in the framework of c-divergence we consider a specific domain, namely the positive quadrant $\mathcal{X} = \mathcal{Y} = \mathbb{R}_{++}^d = (0, \infty)^d$. (To make it a Polish space, we may use for example the metric $d(x, y) = (\sum_{i=1}^d (\log x^i - \log y^i)^2)^{1/2}$.) We keep the discussion brief as a self-contained treatment on general convex domains will be given in Sect. 3.

For $\alpha > 0$, consider the continuous cost function on $\mathbb{R}^d$ given by

$$c^{(\alpha)}(x, y) = \frac{1}{\alpha}\log(1 + \alpha x \cdot y). \tag{21}$$

Remark 3 While $c^{(\alpha)}$ may take negative values, it is equivalent (up to scaling and addition of linear terms, see Remark 1) to the non-negative cost function

$$\log\left(\frac{1}{1 + \alpha d} + \frac{\alpha}{1 + \alpha d}\sum_{i=1}^d x^i y^i\right) - \frac{\alpha}{1 + \alpha d}\sum_{i=1}^d \log(x^i y^i), \tag{22}$$

so that Theorem 4 is applicable. For $\alpha = 1$, this cost function was introduced and studied (under various parameterizations) in [32–34]. In particular, (22) (with $\alpha = 1$) is equivalent to the excess growth rate given in (3) under the change of variables $q^i = y^i/(1 + \sum_j y^j)$ and $p^i = (1/x^i)/(1 + \sum_j 1/x^j)$ for $1 \le i \le d$ (see [34] for details). The symmetric representation chosen here is more convenient in information-geometric computations.

As $\alpha \to 0^+$, the cost function $c^{(\alpha)}(x, y)$ in (21) converges to $x \cdot y$. This is equivalent to the quadratic cost

$$c(x, y) = \frac{1}{2}\|x - y\|^2 = -x \cdot y + \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2$$

after a change of variable $y \mapsto -y$ (say).

The following theorem establishes the $L^{(\alpha)}$-divergence as a c-divergence. Analogous results hold for the $L^{(-\alpha)}$-divergence if we pick $c^{(-\alpha)} = -c^{(\alpha)}$. We omit the proof as this result will not be used in the rest of the paper and most of the work (which builds on [33, Sect. 3]) is contained in Sect. 3.

Theorem 6 *Consider the cost function $c = c^{(\alpha)}$ given by (21).*

(i) *f is c-concave if and only if $e^{\alpha f}$ is concave on $\mathbb{R}^d$.*
(ii) *Suppose f is c-concave. The c-gradient of f at x, if it exists, is given by*

$$D^c f(x) = \frac{Df(x)}{1 - \alpha Df(x) \cdot x}. \tag{23}$$

(iii) *Suppose f is c-concave and differentiable. Then the c-divergence of f is the $L^{(\alpha)}$-divergence given by*

$$\mathbf{D}_f[x : x'] = \mathbf{D}^{(\alpha)}[x : x'] := \frac{1}{\alpha}\log(1 + \alpha Df(x') \cdot (x - x')) - (f(x) - f(x')).$$

3 Exponential concavity/convexity and the $L^{(\pm\alpha)}$-divergences

In this section we define exponential concavity and convexity as well as the associated $L^{(\pm\alpha)}$-divergence $\mathbf{D}^{(\pm\alpha)}$. We also derive their self-dual representations. Our treatment is motivated by, but independent from, the connection with optimal transport maps given in Sect. 2.

3.1 Exponential concavity and convexity

Definition 3 Let $\Omega \subset \mathbb{R}^d$ be an open convex set, $\varphi : \Omega \to \mathbb{R}$, and fix $\alpha > 0$. Also write $\Phi = e^{\alpha\varphi}$.

(i) We say that φ is α-exponentially concave if Φ is concave on Ω. For $\alpha = 0^+$ (that is, the limit as $\alpha \to 0^+$), we say that φ is 0^+-exponentially concave if φ is concave.
(ii) We say that φ is α-exponentially convex if Φ is convex on Ω. For $\alpha = 0^+$, we say that φ is 0^+-exponentially convex if φ is convex.

Assuming φ is twice continuously differentiable on Ω, we have

$$D^2\Phi = D^2 e^{\alpha\varphi} = \alpha e^{\alpha\varphi}\left(D^2\varphi + \alpha(D\varphi)(D\varphi)^\top\right), \tag{24}$$

where $(D\varphi)^\top$ is the transpose of the column vector $D\varphi$. From this we have the following elementary lemma which justifies the limiting cases in Definition 3.

Lemma 1 *Suppose φ is twice continuously differentiable on Ω and $\alpha > 0$.*

(i) φ is α-exponentially concave if and only if

$$- D^2\varphi - \alpha(D\varphi)(D\varphi)^\top \geq 0. \tag{25}$$

(ii) φ is α-exponentially convex if and only if

$$D^2\varphi + \alpha(D\varphi)(D\varphi)^\top \geq 0. \tag{26}$$

Here the inequalities are in the sense of positive semidefinite matrix.

Because of the product term in (24), we cannot pass between exponential concavity and convexity simply by considering $-\varphi$. This is different from the classical case $\alpha = 0$. In particular, from Lemma 1 we see that an α-exponentially concave function is always concave, but an α-exponentially convex function is not necessarily convex. In Sect. 5 we will regard (25) and (26) as the Riemannian metrics induced by $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(-\alpha)}$ respectively.

3.2 $L^{(\pm\alpha)}$-divergences

By convention we always let $\alpha > 0$ be fixed. Suppose φ is a differentiable and α-exponentially concave function. By the concavity of $\Phi = e^{\alpha\varphi}$, for any $\xi, \xi' \in \Omega$ we have

$$\Phi(\xi') + D\Phi(\xi') \cdot (\xi - \xi') \geq \Phi(\xi) \Rightarrow 1 + \alpha D\varphi(\xi') \cdot (\xi - \xi') \geq e^{\alpha(\varphi(\xi) - \varphi(\xi'))}. \tag{27}$$

This motivates the following definition of the $L^{(\alpha)}$-divergence. It is different from the Bregman divergence in that our divergence is given by the ratio rather than the difference in (27).

Definition 4 ($L^{(\alpha)}$-*divergence*) If φ is differentiable and α-exponentially concave, we define the $L^{(\alpha)}$-divergence of φ by

$$\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = \frac{1}{\alpha} \log\left(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')\right) - \left(\varphi(\xi) - \varphi(\xi')\right), \quad \xi, \xi' \in \Omega. \tag{28}$$

For $\alpha = 0^+$, we define the $L^{(0+)}$-divergence by the Bregman divergence (1).

From (27) we have that $\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] \geq 0$. If Φ is strictly concave, then $\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = 0$ only if $\xi = \xi'$.

For an α-exponentially convex function, the analog of (27) is

$$1 + \alpha D\varphi(\xi') \cdot (\xi - \xi') \leq e^{\alpha(\varphi(\xi) - \varphi(\xi'))}. \tag{29}$$

Unfortunately, now the left hand side of (29) may become negative and so its logarithm may not exist. Nevertheless, we can define the $L^{(-\alpha)}$-divergence locally when ξ and ξ' are close. This is sufficient for defining the dualistic structure (g, ∇, ∇^*).

Definition 5 (*local $L^{(-\alpha)}$-divergence*) If φ is differentiable and α-exponentially convex, we define the $L^{(-\alpha)}$-divergence of φ locally by

$$\mathbf{D}^{(-\alpha)}\left[\xi : \xi'\right] = \left(\varphi(\xi) - \varphi(\xi')\right) - \frac{1}{\alpha} \log\left(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')\right), \qquad (30)$$

for all pairs (ξ, ξ') in Ω such that the logarithm in (30) exists. For $\alpha = 0^+$, φ is convex and we define (globally) the $L^{(0-)}$-divergence as the Bregman divergence

$$\mathbf{D}^{(0-)}\left[\xi : \xi'\right] = \left(\varphi(\xi) - \varphi(\xi')\right) - D\varphi(\xi') \cdot (\xi - \xi'). \qquad (31)$$

3.3 Duality of $L^{(\pm\alpha)}$ divergence

In this subsection we develop a duality theory for α-exponentially concave and convex functions. By convention we always assume $\alpha > 0$ is fixed. We generalize the Legendre transform and conjugate which will be used to formulate the self-dual representations of the $L^{(\pm\alpha)}$-divergences. They correspond to the c-transform in Sect. 2.

To ensure that the potential function φ is well-behaved and the differential geometric objects associated with the $L^{(\pm\alpha)}$-divergences are well-defined, we will impose throughout this paper some regularity conditions.

3.3.1 Duality for α-exponentially concave functions

We first consider the case where φ is α-exponentially concave.

Condition 7 (conditions for $L^{(\alpha)}$-divergence) *We assume φ is smooth on Ω and the matrix $-D^2\varphi - \alpha(D\varphi)(D\varphi)^\top$ is strictly positive definite on Ω. By considering a translation of the variable ξ if necessary, we assume without loss of generality that for all $\xi \in \Omega$ we have*

$$1 - \alpha D\varphi(\xi) \cdot \xi = 1 - \alpha \sum_{\ell=1}^{d} \xi^\ell \frac{\partial\varphi}{\partial\xi^\ell}(\xi) > 0. \qquad (32)$$

Note that by strict concavity of $\Phi = e^{\alpha\varphi}$ and (27), the condition (32) holds whenever $0 \in \overline{\Omega}$ (for example, in Sect. 2.4 we have $\Omega = \mathbb{R}^d_{++}$ and $0 \in \overline{\Omega}$).

Observe that there does not exist any positive and strictly concave function $\Phi = e^{\alpha\varphi}$ on the real line. Thus, in order that φ is α-exponentially concave on Ω, the domain Ω must not contain any line (rays are fine). This is different from the Bregman case where the concave function may be globally defined on $\mathbb{R}^d$. Examples of feasible domains in our setting include the open simplex (considered in [33]) as well as the positive quadrant $\mathbb{R}^d_{++} = (0, \infty)^d$.

Condition (32) will be used in the following construction. For each $\xi' \in \Omega$, the supporting tangent hyperplane of $\Phi = e^{\alpha\varphi}$ at ξ' is given by

$$\xi \in \Omega \mapsto \Phi(\xi') + D\Phi(\xi') \cdot (\xi - \xi') = e^{\alpha\varphi(\xi')} \left(1 - \alpha D\varphi(\xi') \cdot \xi' + \alpha D\varphi(\xi') \cdot \xi\right).$$

Thus (32) is equivalent to positivity of the constant coefficient. By concavity of Φ, for $\xi \in \Omega$ we have

$$e^{\alpha\varphi(\xi)} = \min_{\xi' \in \Omega} e^{\alpha\varphi(\xi')} \left((1 - \alpha D\varphi(\xi') \cdot \xi') + \alpha D\varphi(\xi') \cdot \xi\right). \tag{33}$$

Since Φ is strictly concave, the minimum is attained uniquely at $\xi' = \xi$. This and the logarithm is the basis of Theorem 8 which is our duality result. To prepare for it let us introduce some notations.

Definition 6 (α-*gradient*) We define the α-gradient of φ at $\xi \in \Omega$ by

$$D^{(\alpha)}\varphi(\xi) = \frac{1}{1 - \alpha D\varphi(\xi) \cdot \xi} D\varphi(\xi), \tag{34}$$

where $D\varphi = D^{(0)}\varphi$ is the Euclidean gradient. Condition 7 ensures that it is well-defined.

Definition 7 (*dual coordinate*) We define the dual coordinate η by

$$\eta = D^{(\alpha)}\varphi(\xi). \tag{35}$$

We let $\Omega' = D^{(\alpha)}\varphi(\Omega)$ be the range of η (which may not be convex). The coordinates of η are denoted using lower-indices, i.e., $\eta = (\eta_1, \ldots, \eta_d)$.

This terminology is justified by the following result.

Proposition 1 *The α-gradient $D^{(\alpha)}\varphi$ is a diffeomorphism from Ω onto Ω' which is an open set in $\mathbb{R}^d$.*

Proof It is clear from (34) that the map $D^{(\alpha)}\varphi$ is smooth.

To show that $D^{(\alpha)}\varphi$ is injective, suppose towards a contradiction that $D^{(\alpha)}\varphi(\xi) = D^{(\alpha)}\varphi(\xi')$ for distinct $\xi, \xi' \in \Omega$. By definition, we have

$$\frac{D\varphi(\xi)}{1 - \alpha D\varphi(\xi) \cdot \xi} = \frac{D\varphi(\xi')}{1 - \alpha D\varphi(\xi') \cdot \xi'}.$$

It follows that for any $\xi'' \in \Omega$, we have

$$1 + \frac{\alpha D\varphi(\xi) \cdot \xi''}{1 - \alpha D\varphi(\xi) \cdot \xi} = 1 + \frac{\alpha D\varphi(\xi') \cdot \xi''}{1 - \alpha D\varphi(\xi') \cdot \xi'}.$$

Rearranging, we have the identity

$$\frac{1 + \alpha D\varphi(\xi) \cdot (\xi'' - \xi)}{1 - \alpha D\varphi(\xi) \cdot \xi} = \frac{1 + \alpha D\varphi(\xi') \cdot (\xi'' - \xi')}{1 - \alpha D\varphi(\xi') \cdot \xi'}.$$

Letting ξ'' be ξ and ξ' respectively, we see after some computation that

$$(1 + \alpha D\varphi(\xi) \cdot (\xi' - \xi)) \cdot \left(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')\right) = 1.$$

Observe that this is equivalent to

$$\mathbf{D}^{(\alpha)} \left[\xi' : \xi\right] + \mathbf{D}^{(\alpha)} \left[\xi : \xi'\right] = 0,$$

which is a contradiction since $\mathbf{D}^{(\alpha)} \left[\xi' : \xi\right], \mathbf{D}^{(\alpha)} \left[\xi : \xi'\right] > 0$ by Condition 7. This proves that the α-gradient is injective.

Finally, from Condition 7 again we have $-D^2\varphi - \alpha(D\varphi)(D\varphi)^\top > 0$. This can be used to show that the Jacobian $\frac{\partial \eta}{\partial \xi}$ is invertible. By the inverse function theorem, $D^{(\alpha)}\varphi : \Omega \to \Omega'$ is a diffeomorphism and Ω' is open. (Essentially the same result is proved in [33, Theorem 3.2] so we do not provide the details. Also see (72) below which gives an explicit expression of the Jacobian.) □

Definition 8 (α-*conjugate*) Consider the function $c^{(\alpha)}$ given by (21) and defined for $x, y \in \mathbb{R}^d$ with $1 + \alpha x \cdot y > 0$. We define the α-conjugate ψ of φ on the open set $\Omega' = D^{(\alpha)}\varphi(\Omega)$ by

$$\psi(\eta) = c^{(\alpha)}(\xi, \eta) - \varphi(\xi), \quad \xi = (D^{(\alpha)}\varphi)^{-1}(\eta). \tag{36}$$

By (41) below $1 + \alpha\xi \cdot \eta > 0$, hence $c^{(\alpha)}(\xi, \eta)$ and $\psi(\eta)$ are well-defined.

Theorem 8 (α-duality) *For $\xi \in \Omega$, $\eta \in \Omega'$ we have*

$$\varphi(\xi) = \min_{\eta' \in \Omega'} \left(c^{(\alpha)}(\xi, \eta') - \psi(\eta')\right), \tag{37}$$

$$\psi(\eta) = \min_{\xi' \in \Omega} \left(c^{(\alpha)}(\xi', \eta) - \varphi(\xi')\right). \tag{38}$$

Furthermore, ψ is locally α-exponentially concave on Ω' in the sense that $-D^2\psi - \alpha(D\psi)(D\psi)^\top > 0$ on Ω', and we have $1 - \alpha D\psi(\eta) \cdot \eta > 0$ on Ω' so that the α-gradient $D^{(\alpha)}\psi(\eta) = \frac{1}{1 - \alpha D\psi(\eta) \cdot \eta} D\psi(\eta)$ is well-defined. We have

$$(D^{(\alpha)}\varphi)^{-1} = D^{(\alpha)}\psi. \tag{39}$$

Proof Taking logarithm in (33) and rearranging, we have

$$
\begin{aligned}
\varphi(\xi) &= \min_{\xi' \in \Omega} \left\{ \varphi(\xi') + \frac{1}{\alpha} \log \left(1 - \alpha D\varphi(\xi') \cdot \xi' + \alpha D\varphi(\xi') \cdot \xi \right) \right\} \\
&= \min_{\xi' \in \Omega} \left\{ \frac{1}{\alpha} \log \left(1 + \alpha \frac{D\varphi(\xi')}{1 - \alpha D\varphi(\xi') \cdot \xi'} \cdot \xi \right) + \varphi(\xi') \right. \\
&\quad \left. + \frac{1}{\alpha} \log \left(1 - \alpha D\varphi(\xi') \cdot \xi' \right) \right\} \\
&= \min_{\xi' \in \Omega} \left\{ c^{(\alpha)}(\xi, \eta') + \varphi(\xi') + \frac{1}{\alpha} \log \left(1 - \alpha D\varphi(\xi') \cdot \xi' \right) \right\}.
\end{aligned}
\tag{40}
$$

Note that since $\eta' = D^{(\alpha)}\varphi(\xi')$, from (34) we have the useful identity

$$
1 + \alpha \xi' \cdot \eta' = \frac{1}{1 - \alpha D\varphi(\xi') \cdot \xi'}.
\tag{41}
$$

With this observation, we see that

$$
\varphi(\xi') + \frac{1}{\alpha} \log \left(1 - \alpha D\varphi(\xi') \cdot \xi' \right) = - \left(c^{(\alpha)}(\xi', \eta') - \varphi(\xi') \right) = -\psi(\eta').
$$

Putting this into (40) gives (37).

From (37), for any η, ξ' we have

$$
\varphi(\xi') \le c^{(\alpha)}(\xi', \eta) - \psi(\eta) \Rightarrow \psi(\eta) \le c^{(\alpha)}(\xi', \eta) - \varphi(\xi').
$$

By definition we have $\psi(\eta) = c^{(\alpha)}(\xi, \eta) - \varphi(\xi)$, so (38) holds as well. In particular, for any $\xi, \xi' \in \Omega$ we have

$$
\psi(\eta) = c^{(\alpha)}(\xi, \eta) - \varphi(\xi) \le c^{(\alpha)}(\xi', \eta) - \varphi(\xi').
\tag{42}
$$

Using (41), we compute

$$
\begin{aligned}
c^{(\alpha)}(\xi', \eta) - c^{(\alpha)}(\xi, \eta) &= \frac{1}{\alpha} \log \left(\frac{1 + \alpha \xi' \cdot \eta}{1 + \alpha \xi \cdot \eta} \right) \\
&= \frac{1}{\alpha} \log \left(\frac{1}{1 + \alpha \xi \cdot \eta} + \alpha \xi' \cdot \frac{\eta}{1 + \alpha \xi \cdot \eta} \right) \\
&= \frac{1}{\alpha} \log \left(1 + \alpha D\varphi(\xi) \cdot (\xi' - \xi) \right).
\end{aligned}
$$

It follows that

$$
c^{(\alpha)}(\xi', \eta) - c^{(\alpha)}(\xi, \eta) - \left(\varphi(\xi') - \varphi(\xi) \right) = \mathbf{D}^{(\alpha)} \left[\xi' : \xi \right]
\tag{43}
$$

is the $L^{(\alpha)}$-divergence of φ.

By (37), for any $\xi \in \Omega$, η is the unique minimizer (over Ω') of

$$\eta' \mapsto \frac{1}{\alpha} \log\left(1 + \alpha\xi \cdot \eta'\right) - \psi(\eta').$$

Since Ω' is open, we have the first order condition $\frac{\xi}{1+\alpha\xi\cdot\eta} = D\psi(\eta)$. Rearranging, we have

$$1 - \alpha D\psi(\eta) \cdot \eta = \frac{1}{1 + \alpha\xi \cdot \eta}.$$

Thus $D^{(\alpha)}\psi(\eta)$ is well-defined, $D^{(\alpha)}\psi(\eta) = \xi$, and we have (39).

It remains to show that

$$- D^2\psi(\eta) - \alpha(D\psi(\eta))(D\psi(\eta))^\top \tag{44}$$

is strictly positive definite on Ω'. Expressing (43) in terms of ψ using the definition (36) of α-conjugate, it can be shown that

$$\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = \frac{1}{\alpha} \log\left(1 + \alpha D\psi(\eta) \cdot (\eta' - \eta)\right) - \left(\psi(\eta') - \psi(\eta)\right),$$

so the $L^{(\alpha)}$-divergence of φ is the $L^{(\alpha)}$-divergence of ψ with the arguments interchanged (this is a special case of Corollary 1). It follows that (44) is the Riemannian metric of $\mathbf{D}^{(\alpha)}$ expressed in terms of the dual coordinates η. Since φ satisfies Condition 7 by assumption, the matrix (44) is strictly positive definite as well. $\square$

From (36) and (43), we immediately obtain the self-dual representation which completes the circle of ideas and the duality theory for α-exponentially concave functions.

Theorem 9 (self-dual representation) *The $L^{(\alpha)}$-divergence of φ admits the self-dual representation*

$$\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] = c^{(\alpha)}(\xi, \eta') - \varphi(\xi) - \psi(\eta'), \quad \xi, \xi' \in \Omega. \tag{45}$$

3.3.2 Duality for α-exponentially convex functions

Now we state without proof the analogous results for the $L^{(-\alpha)}$-divergence defined by an α-exponentially convex function. The only difference is that because $e^{\alpha\varphi}$ is convex, we have to take maximums instead of minimums and switch the inequalities accordingly. Also, as the $L^{(-\alpha)}$-divergence may not be globally defined, here we only consider the local geometry.

Condition 10 (conditions for $L^{(-\alpha)}$-divergence) *We assume φ is smooth and the matrix $D^2\varphi + \alpha(D\varphi)(D\varphi)^\top$ is strictly positive definite on Ω. Translating and restricting the domain if necessary, we assume that $0 \in \Omega$ and*

$$1 + \alpha D\varphi(\xi') \cdot (\xi - \xi') > 0 \tag{46}$$

for all $\xi, \xi' \in \Omega$. *This implies that the* $L^{(-\alpha)}$-*divergence is defined for all* $\xi, \xi' \in \Omega$.

Theorem 11 (duality for $L^{(-\alpha)}$-divergence) *Consider the α-gradient and $c^{(\alpha)}$ defined as in* (34) *and* (21). *Then the map* $\xi \mapsto \eta = D^{(\alpha)}\varphi(\xi) = \frac{D\varphi(\xi)}{1-\alpha\varphi(\xi)\cdot\xi}$ *is a diffeomorphism, and the $(-\alpha)$-conjugate ψ defined by* (36) *is locally α-exponentially convex. The functions φ and ψ are related by*

$$\varphi(\xi) = \max_{\eta' \in \Omega'} \left(c^{(\alpha)}(\xi, \eta') - \psi(\eta') \right), \quad \psi(\eta) = \max_{\xi' \in \Omega} \left(c^{(\alpha)}(\xi', \eta) - \varphi(\xi') \right).$$

The $L^{(-\alpha)}$-*divergence* (30) *admits the self-dual representation*

$$\mathbf{D}^{(-\alpha)} \left[\xi : \xi' \right] = \varphi(\xi) + \psi(\eta') - c^{(\alpha)}(\xi, \eta'), \quad \xi, \xi' \in \Omega. \tag{47}$$

4 Generalizations of exponential family and Rényi divergence

It is well known that Bregman divergences ($\mathbf{D}^{(0\pm)}$) arise naturally in applications involving exponential families. In fact, it was shown by Banerjee et al in [10] that there is a one-to-one correspondence between what they call regular Bregman divergences and regular exponential families. Are there parameterized families of probability distributions that correspond to the $L^{(\pm\alpha)}$ divergences?

In this section we consider generalizations of exponential family that are closely related to the α-family [2, Sect. 2.6] and the q-exponential family [3, Sect. 4.3]. We show that a suitably defined potential function is α-exponentially concave/convex, and the corresponding $L^{(\pm\alpha)}$-divergences are Rényi divergences. This gives a natural framework to study the geometry induced by the Rényi divergence.

4.1 Motivations

We first recall the main idea in [10]. Consider an exponential family of probability densities of the form

$$\log p(x, \xi) = \xi \cdot h(x) - \varphi(\xi), \tag{48}$$

where $h(x) = (h_1(x), \ldots, h_d(x))$ is a vector of sufficient statistics, and φ is the convex cumulant generating function. Let $\mathbf{D} = \mathbf{D}^{(0-)}$ be the Bregman divergence of φ. By the self-dual representation (induced by the function $c(\xi, h(x)) = -\xi \cdot h(x)$; see Sect. 2), we have

$$\log p(x, \xi) = -\mathbf{D} \left[\xi : \xi' \right] + \psi(h(x)), \tag{49}$$

where ψ is the convex conjugate of φ and $h(x)$ plays the role of the dual variable η'. Remarkably, the function $-\psi$ is the Shannon entropy (see [3, Sect. 2.1]).

Now let us consider an $L^{(\alpha)}$-divergence $\mathbf{D}^{(\alpha)}$ with $\alpha > 0$, where φ is α-exponentially concave. Its self-dual representation (see Theorem 9) is

$$\mathbf{D}^{(\alpha)} \left[\xi : \xi' \right] = \frac{1}{\alpha} \log \left(1 + \alpha\xi \cdot \eta' \right) - \varphi(\xi) - \psi(\eta').$$

Imitating (49), let us consider a parameterized density of the form

$$\log p(x, \xi) = -\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] - \psi(h(x)) = -\frac{1}{\alpha}\log(1 + \alpha\xi \cdot h(x)) + \varphi(\xi), \quad (50)$$

where $h(x) = \eta'$ as well. Rearranging gives the representation

$$p(x, \xi) = (1 + \alpha\xi \cdot h(x))^{-\frac{1}{\alpha}} e^{\varphi(\xi)}. \quad (51)$$

If μ is the dominating measure of the family, the normalizing function $\varphi(\xi)$ in (51) is given by

$$\varphi(\xi) = -\log \int (1 + \alpha\xi \cdot h(x))^{-\frac{1}{\alpha}} d\mu(x). \quad (52)$$

It is the analogue of the cumulant generating function in (48). For lack of a better name, let us call (51) an $\mathcal{F}^{(\alpha)}$-family.

Similarly, the $L^{(-\alpha)}$-divergence leads to the $\mathcal{F}^{(-\alpha)}$-family

$$p(x, \xi) = (1 + \alpha\xi \cdot h(x))^{\frac{1}{\alpha}} e^{-\varphi(\xi)}, \quad (53)$$

where

$$\varphi(\xi) = \log \int (1 + \alpha\xi \cdot h(x))^{\frac{1}{\alpha}} d\mu(x). \quad (54)$$

Remark 4 For $\alpha' \neq 1$, Amari and Nagaoka [2, Sect. 2.6] defined an α'-family of positive measures whose densities are given by

$$L^{(\alpha)}(p(x, \xi)) = C(x) + \sum_i \xi^i F_i(x), \quad L^{(\alpha)}(u) = \frac{2}{1 - \alpha'}u^{\frac{1-\alpha'}{2}},$$

$$\text{or} \quad p(x, \xi) = \left[\frac{1 - \alpha'}{2}\left(C(x) + \sum_i \xi^i F_i(x)\right)\right]^{\frac{2}{1-\alpha'}}, \quad (55)$$

where C and F_i are functions on $\mathcal{X}$. When $\alpha' \to 1$ then (55) reduces to the exponential family. Comparing this with (52) and (54), we see that our $\mathcal{F}^{(\pm\alpha)}$-families are essentially a normalized version of their α'-family with a different parameterization. We keep our formulation so as to be consistent with (50) (which is the motivation of this section) and the framework of the paper.

4.2 Exponential concavity/convexity of φ

In order that the densities (51) and (53) are well-defined, we need to check that the potential functions (52) and (54) indeed define α-exponentially concave/convex functions. We impose the following conditions which formalize the ideas in Sect. 4.1.

Condition 12 *Let μ be a probability measure on a measurable space $\mathcal{X}$. Let $h = (h_1, \ldots, h_d) : \mathcal{X} \to \mathbb{R}^n_{++}$ be a vector-valued function such that if $\xi \cdot h(x) = \xi' \cdot h(x)$ μ-almost everywhere on $\mathcal{X}$ then $\xi = \xi'$.*

For the $\mathcal{F}^{(\alpha)}$-case, we assume that $1 + \alpha \xi \cdot h > 0$ and the integral in (52) is finite for all ξ in an open convex set $\Omega \subset \mathbb{R}^d$ and we can differentiate with respect to ξ under the integral sign. Analogous conditions are imposed for the $\mathcal{F}^{(-\alpha)}$-case. In both cases we call φ the potential function of the family.

4.2.1 $\mathcal{F}^{(\alpha)}$-family

Proposition 2 *For $\alpha > 0$ fixed, the potential function $\varphi(\xi)$ of the $\mathcal{F}^{(\alpha)}$-family defined by (52) is α-exponentially concave on Ω.*

Proof We need to check that $\Phi = e^{\alpha \varphi}$ is concave on Ω. From (52), we have

$$\Phi(\xi) = \left(\int (1 + \alpha \xi \cdot h(x))^{-\frac{1}{\alpha}} d\mu(x) \right)^{-\alpha}, \tag{56}$$

which can be regarded as a generalized harmonic average.

Let X be the random element $X(x) = x$, where $x \in \mathcal{X}$ and $\mathcal{X}$ is the sample space. Given $\xi \in \Omega$, let $\mathbb{E}_\xi$ denote the expectation with respect to $p(x, \xi)$, and let Z_ξ be the random vector defined by

$$Z_\xi = \frac{h(X)}{1 + \alpha \xi \cdot h(X)}. \tag{57}$$

Differentiating (56) and simplifying, we have

$$\frac{\partial \Phi}{\partial \xi^i}(\xi) = \alpha \left(\int (1 + \alpha \xi \cdot h(x))^{-\frac{1}{\alpha}} d\mu \right)^{-\alpha - 1} \left(\int (1 + \alpha \xi \cdot h(x))^{-\frac{1}{\alpha} - 1} h_i(x) d\mu \right)$$

$$= \alpha \left(\int (1 + \alpha \xi \cdot h(x))^{-\frac{1}{\alpha}} d\mu(x) \right)^{-\alpha} \int p(x, \xi) Z_{\xi,i} d\mu(x)$$

$$= \alpha \Phi(\xi) \mathbb{E}_\xi Z_{\xi,i}.$$

In particular, we have

$$D\varphi(\xi) = \mathbb{E}_\xi[Z_\xi] \tag{58}$$

which can be interpreted as an expectation parameter (although Z_ξ is not observable if ξ is unknown). Differentiating one more time and calculating carefully, we get

$$\frac{\partial^2 \Phi}{\partial \xi^i \partial \xi^j}(\xi) = \alpha(1 + \alpha) \left(\mathbb{E}_\xi[Z_{\xi,i}] \mathbb{E}_\xi[Z_{\xi,j}] - \mathbb{E}_\xi[Z_{\xi,i} Z_{\xi,j}] \right)$$

$$= -\alpha(1 + \alpha) \mathrm{Cov}(Z_{\xi,i}, Z_{\xi,j}).$$

Since $D^2 \Phi$ is a negative multiple of the covariance matrix of Z_ξ, Φ is concave and φ is α-exponentially concave. $\qquad\square$

4.2.2 $\mathcal{F}^{(-\alpha)}$-family

Next we consider the $\mathcal{F}^{(-\alpha)}$-family. Interestingly, the result is a little different and depends on the value of α.

Proposition 3 *Consider the potential function $\varphi(\xi)$ of the $\mathcal{F}^{(-\alpha)}$-family defined by* (54).

1. *If $0 < \alpha < 1$, then φ is α-exponentially convex.*
2. *If $\alpha > 1$, then φ is α-exponentially concave.*

Proof Now we heave

$$\Phi(\xi) = e^{\alpha\varphi(\xi)} = \left(\int (1 + \alpha\xi \cdot h)^{\frac{1}{\alpha}} d\mu \right)^{\alpha}.$$

A computation similar to that in the proof of Proposition 2 shows that

$$D\varphi(\xi) = \mathbb{E}_\xi Z_\xi = \mathbb{E}_\xi \left[\frac{h(X)}{1 + \alpha\xi \cdot h(X)} \right]$$

and

$$\frac{\partial^2 \Phi}{\partial \xi^i \partial \xi^j}(\xi) = \alpha(1 - \alpha)\text{Cov}_\xi(Z_{\xi,i}, Z_{\xi,j}).$$

Now Φ is convex if $0 < \alpha < 1$ and concave if $\alpha > 1$.   □

4.3 Rényi entropy and divergence

Based on information-theoretic reasonings, Rényi entropy and divergence were first introduced by Rényi in [36] and they have numerous applications in information theory and statistical physics. See [38] for a survey of their remarkable properties. We show that these quantities arise naturally in our $\mathcal{F}^{(\pm\alpha)}$-families.

We first recall the definitions of Rényi entropy and divergence. Same as above we fix a measure space $(\mathcal{X}, \mu)$.

Definition 9 Let $\tilde{\alpha} \in (0, 1) \cup (1, \infty)$. Let P, Q be probability measures on $\mathcal{X}$ that are absolutely continuous with respect to μ. We let $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ be their densities.

1. The Rényi entropy of P of order $\tilde{\alpha}$ (and with reference measure μ) is defined by

$$\mathbf{H}_{\tilde{\alpha}}(P) = \frac{1}{1 - \tilde{\alpha}} \log \int p^{\tilde{\alpha}} d\mu. \tag{59}$$

We also denote the above by $\mathbf{H}_{\tilde{\alpha}}(p)$.

2. The Rényi divergence or order $\tilde{\alpha}$ between P and Q is defined by

$$\mathbf{D}_{\tilde{\alpha}}(P||Q) = \frac{1}{\tilde{\alpha} - 1} \log \int p^{\tilde{\alpha}} q^{1-\tilde{\alpha}} d\mu. \tag{60}$$

Note that $\mathbf{H}_{\tilde{\alpha}}(P) = -\mathbf{D}_{\tilde{\alpha}}(P||\mu)$ if μ itself is a probability measure. We use a different notation to distinguish the Rényi divergence from the $L^{(\pm\alpha)}$-divergences $\mathbf{D}^{(\pm\alpha)}[\xi : \xi']$ (and $\tilde{\alpha}$ from α). It is well known that as $\tilde{\alpha} \to 1$, the Rényi entropy and divergence converge respectively to the Shannon entropy and the Kullback–Leibler divergence.

The following is the main result of this section. The different cases correspond to the exponential concavity/convexity of φ established in Propositions 2 and 3.

Theorem 13

(i) *Consider an $\mathcal{F}^{(\alpha)}$-family with $\alpha > 0$ and potential function φ (see (52)). Then the $L^{(\alpha)}$-divergence of φ is the Rényi entropy of order $\tilde{\alpha} := 1 + \alpha$:*

$$\mathbf{D}^{(\alpha)}[\xi : \xi'] = \mathbf{D}_{\tilde{\alpha}}\left(p(\cdot, \xi')||p(\cdot, \xi)\right). \tag{61}$$

Moreover, the α-conjugate ψ of φ (see Definition 8) is the Rényi entropy of order $\tilde{\alpha}$:

$$\psi(\eta) = \mathbf{H}_{\tilde{\alpha}}(p(\cdot, \xi)), \quad \eta = D^{(\alpha)}\varphi(\xi). \tag{62}$$

(ii) *Consider an $\mathcal{F}^{(-\alpha)}$-family (see (53)). If $0 < \alpha < 1$, the $L^{(-\alpha)}$-divergence of φ is the Rényi entropy of order $\tilde{\alpha} := 1 - \alpha$:*

$$\mathbf{D}^{(-\alpha)}[\xi : \xi'] = \mathbf{D}_{\tilde{\alpha}}\left(p(\cdot, \xi')||p(\cdot, \xi)\right), \tag{63}$$

and the $(-\alpha)$-conjugate is $\psi(\eta) = -\mathbf{H}_{\tilde{\alpha}}(p(\cdot, \xi))$ (note the minus sign).
If $\alpha > 1$, the $L^{(\alpha)}$-divergence of φ is given by

$$\mathbf{D}^{(\alpha)}[\xi : \xi'] = \frac{\alpha - 1}{\alpha} \mathbf{D}_{\alpha}\left(p(\cdot, \xi)||p(\cdot, \xi')\right), \tag{64}$$

Here $\tilde{\alpha} = \alpha$ and the order of the arguments stay the same. The $(-\alpha)$-conjugate is

$$\psi(\eta) = \frac{-1}{\alpha} \log \int p(x, \xi)^{1-\alpha} d\mu.$$

Note that it is not a Rényi entropy since $1 - \alpha < 0$.

Proof (i) We want to compute

$$\mathbf{D}^{(\alpha)}[\xi : \xi'] = \frac{1}{\alpha} \log\left(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')\right) - (\varphi(\xi) - \varphi(\xi')).$$

First, using (57) and (58), we have

$$
\begin{aligned}
1 + \alpha D\varphi(\xi') \cdot (\xi - \xi') &= 1 + \alpha \mathbb{E}_{\xi'}[Z_{\xi'}] \cdot (\xi - \xi') \\
&= 1 + \alpha \mathbb{E}_{\xi'}\left[\frac{h(X)}{1 + \alpha\xi' \cdot h(X)} \cdot (\xi - \xi')\right] \\
&= \mathbb{E}_{\xi'}\frac{1 + \alpha\xi \cdot h(X)}{1 + \alpha\xi' \cdot h(X)}.
\end{aligned}
\tag{65}
$$

From (51), we have

$$
1 + \alpha\xi' \cdot h(X) = \frac{e^{\alpha\varphi(\xi')}}{p(X, \xi')^\alpha}.
$$

It follows from (65) that

$$
\begin{aligned}
\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] &= \frac{1}{\alpha}\log\left(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')\right) - (\varphi(\xi) - \varphi(\xi')) \\
&= \frac{1}{\alpha}\log \mathbb{E}_{\xi'}\frac{e^{\alpha\varphi(\xi)}/p(X, \xi)^\alpha}{e^{\alpha\varphi(\xi')}/p(X, \xi')^\alpha} - (\varphi(\xi) - \varphi(\xi')) \\
&= \frac{1}{\alpha}\log \mathbb{E}_{\xi'}\frac{p(X, \xi')^\alpha}{p(X, \xi)^\alpha}.
\end{aligned}
$$

Expressing the divergence in terms of the integral, we have

$$
\begin{aligned}
\mathbf{D}^{(\alpha)}\left[\xi : \xi'\right] &= \frac{1}{\alpha}\log \int p(x, \xi')^{1+\alpha} p(x, \xi)^{-\alpha} d\mu \\
&= \frac{1}{\tilde{\alpha} - 1}\log \int p(x, \xi')^{\tilde{\alpha}} p(x, \xi)^{1-\tilde{\alpha}} d\mu \\
&= \mathbf{D}_{\tilde{\alpha}}\left(p(\cdot, \xi') \,||\, p(\cdot, \xi)\right),
\end{aligned}
$$

where $\tilde{\alpha} = 1 + \alpha > 1$.

Next we consider the α-conjugate

$$
\psi(\eta) = \frac{1}{\alpha}\log\left(1 + \alpha\xi \cdot \eta\right) - \varphi(\xi).
$$

Using the identity (41) and following the argument in (65), we have

$$
\begin{aligned}
1 + \alpha\xi \cdot \eta &= (1 - \alpha D\varphi(\xi) \cdot \xi)^{-1} \\
&= \left(\mathbb{E}_\xi\left[(1 + \alpha\xi \cdot h(X))^{-1}\right]\right)^{-1} \\
&= \left(\mathbb{E}_\xi\left[p(X, \xi)^\alpha e^{-\alpha\varphi(\xi)}\right]\right)^{-1}.
\end{aligned}
$$

From this we get

$$\psi(\eta) = -\frac{1}{\alpha} \log \int p(x, \xi)^{1+\alpha} e^{-\alpha\varphi(\xi)} d\mu - \varphi(\xi)$$
$$= \frac{1}{1-\tilde{\alpha}} \int p(x, \xi)^{\tilde{\alpha}} d\mu = \mathbf{H}_{\tilde{\alpha}}(p(\cdot, \xi)).$$

This proves (i), and the proof of (ii) is similar.   □

Remark 5 It appears from the above results that the $\mathcal{F}^{(-\alpha)}$-family with $\alpha > 1$ is different from the other cases. Nevertheless, the $\mathcal{F}^{(\alpha)}$-family for $\alpha > 0$ and the $\mathcal{F}^{(-\alpha)}$-family for $0 < \alpha < 1$ together give all Rényi entropy/divergence of order $\tilde{\alpha} \in (0, 1) \cup (1, \infty)$.

Example 2 (*finite simplex*) Let $\mathcal{X} = \{0, 1, \ldots, d\}$ be a finite set and let μ be the counting measure on $\mathcal{X}$. Fix $\alpha > 0$. Let $h_i(x) = \delta_i(x)$ be the indicator function for point i, $i = 1, \ldots, d$. Then (51) gives, for $\alpha > 0$, the $\mathcal{F}^{(\alpha)}$-family of probability mass functions on $\mathcal{X}$ given by

$$p(i, \xi) = (1 + \alpha\xi^i)^{-1/\alpha} e^{\varphi(\xi)}, \quad i = 1, \ldots, d, \quad p(0, \xi) = e^{\varphi(\xi)}, \qquad (66)$$

where the α-exponentially concave function φ is

$$\varphi(\xi) = \log p(0, \xi) = -\log\left(1 + \sum_{i=1}^{d} (1 + \alpha\xi^i)^{-1/\alpha}\right),$$

and the domain is $\Omega = \{\xi \in \mathbb{R}^d : 1 + \alpha\xi^i > 0 \ \forall i\}$. The coordinate $\xi \in \Omega$ is a global coordinate system of the open simplex $\mathcal{S}_d = \{p = (p^0, \ldots, p^d) \in (0, 1)^{1+d} : \sum_i p^i = 1\}$. From (66), we have

$$\xi^i = \frac{1}{\alpha}\left(\left(\frac{p(0, \xi)}{p(i, \xi)}\right)^\alpha - 1\right), \quad i = 1, \ldots, d.$$

By Theorem 13, the $L^{(\alpha)}$-divergence of φ is the discrete Rényi divergence of order $\tilde{\alpha} = 1 + \alpha > 0$.

Alternatively, we may express the simplex $\mathcal{S}_d$ as an $\mathcal{F}^{(-\alpha)}$-family where

$$p(i, \xi) = (1 + \alpha\xi^i)^{\frac{1}{\alpha}} e^{-\varphi(\xi)}, \quad i = 1, \ldots, d, \quad p(0, \xi) = e^{-\varphi(\xi)},$$
$$\varphi(\xi) = -\log p(0, \xi) = \log\left(1 + \sum_{j=1}^{d} (1 + \alpha\xi^j)^{1/\alpha}\right),$$

and

$$\xi^i = \frac{1}{\alpha}\left(\left(\frac{p(i, \xi)}{p(0, \xi)}\right)^\alpha - 1\right), \quad i = 1, \ldots, d.$$

When $0 < \alpha < 1$, the $L^{(-\alpha)}$-divergence of φ is the discrete Rényi divergence of order $\tilde{\alpha} = 1 - \alpha > 0$.

The corresponding dualistic structures will be considered in Sect. 6.4.

5 Dualistic structure of $L^{(\pm\alpha)}$-divergence

In this and the next sections we study the dualistic structure induced by a given $L^{(\pm\alpha)}$-divergence. First we state the general definition of dualistic structure. For preliminaries in differential geometry see [3,9].

Definition 10 Let M be a smooth manifold. A dualistic structure on $\mathcal{M}$ is a quadruplet (M, g, ∇, ∇^*) where g (also denoted by $\langle \cdot, \cdot \rangle$) is a Riemannian metric, and (∇, ∇^*) is a pair of torsion-free affine connections that are dual with respect to g: for any vector fields X, Y and Z, the covariant derivatives satisfy the identity

$$Z\langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z^* Y \rangle. \tag{67}$$

We also call $\mathcal{M}$ equipped with a dualistic structure a statistical manifold.

For a dualistic structure, it is easy to see that the average $\overline{\nabla} = \frac{1}{2}(\nabla + \nabla^*)$ is the Levi-Civita connection of the metric g. By the general results of Eguchi (see [16,17]), any divergence (see [3, Definition 1.1]) on a manifold $\mathcal{M}$ defines a dualistic structure on $\mathcal{M}$. It is this structure induced by $\mathbf{D}^{(\pm\alpha)}$ that we study in this section.

We will derive explicit coordinate representations of this dualistic structure. The geometric consequences will be studied in Sect. 6. We stress that while any divergence induces a geometric structure, the self-dual representations motivated by optimal transport suggest the appropriate coordinate systems ξ and $\eta = D^{(\alpha)}\varphi(\xi)$.

By the results of Sect. 4, for an $\mathcal{F}^{(\pm\alpha)}$-family this gives the geometry induced by the Rényi divergence. The Rényi divergence is closely related to the α-divergence.[2] Our framework is more general as it does not depend on the probabilistic representation. We discuss the connection with the α-divergences in Sect. 6.4.

5.1 Notations

We focus on the case of $L^{(\alpha)}$-divergence. The arguments for the $L^{(-\alpha)}$-divergence are essentially the same (with some changes of signs due to the different self-dual representation) and will be left for the reader. The main results for the $L^{(-\alpha)}$-divergence will be stated in Sect. 5.3.

Henceforth we fix an α-exponentially concave function φ on Ω satisfying Condition 7, and let $\mathbf{D} = \mathbf{D}^{(\alpha)}$ be the $L^{(\alpha)}$-divergence of φ. We let $\xi = (\xi^1, \ldots, \xi^d)$ be the Euclidean coordinate system on Ω. We will use the self-dual representation derived in Theorem 9:

$$\mathbf{D}\left[\xi : \xi'\right] = \frac{1}{\alpha} \log\left(1 + \alpha\xi \cdot \eta'\right) - \varphi(\xi) - \psi(\eta'). \tag{68}$$

---

[2] The author thanks Shun-ichi Amari for pointing this out.

Here $\eta = D^{(\alpha)}\varphi(\xi) = (\eta_1, \ldots, \eta_d)$ is the dual coordinate system given by the α-gradient. To simplify the notations we let

$$\Pi(\xi, \eta') = \Pi^{(\alpha)}(\xi, \eta') = 1 + \alpha\xi \cdot \eta', \tag{69}$$

where $\xi \in \Omega$ and $\eta' \in \Omega' = D^{(\alpha)}\varphi(\Omega)$. We regard ξ and η as column vectors. The transpose of a vector or matrix A is denoted by $A^\top$. The identity matrix is denoted by I.

Consider the smooth manifold $\mathcal{M} = \Omega$ where the primal coordinate ξ (identity map) and the dual coordinate $\eta = D^{(\alpha)}\varphi(\xi)$ are global coordinate systems. Note that by symmetry of the cost function $c^{(\alpha)}$ and the associated α-duality, it suffices to consider the primal connection ∇ expressed in the primal coordinate system ξ. Consider the coordinate vector fields $\frac{\partial}{\partial\xi^1}, \ldots, \frac{\partial}{\partial\xi^d}$. By definition, the coordinate representations of (g, ∇, ∇^*) under ξ are given by

$$g_{ij}(\xi) = \left\langle \frac{\partial}{\partial\xi^i}, \frac{\partial}{\partial\xi^j} \right\rangle = - \frac{\partial}{\partial\xi^i} \frac{\partial}{\partial\xi'^j} \mathbf{D}\left[\xi : \xi'\right]\Big|_{\xi=\xi'},$$
$$\Gamma_{ijk}(\xi) = \left\langle \nabla_{\frac{\partial}{\partial\xi^i}} \frac{\partial}{\partial\xi^j}, \frac{\partial}{\partial\xi^k} \right\rangle = - \frac{\partial^2}{\partial\xi^i\partial\xi^j} \frac{\partial}{\partial\xi'^k} \mathbf{D}\left[\xi : \xi'\right]\Big|_{\xi=\xi'}, \tag{70}$$
$$\Gamma^*_{ijk}(\xi) = \left\langle \nabla^*_{\frac{\partial}{\partial\xi^i}} \frac{\partial}{\partial\xi^j}, \frac{\partial}{\partial\xi^k} \right\rangle = - \frac{\partial^2}{\partial\xi'^i\partial\xi'^j} \frac{\partial}{\partial\xi^k} \mathbf{D}\left[\xi : \xi'\right]\Big|_{\xi=\xi'}.$$

Note that by construction the duality (67) automatically holds (see for example [3, Theorem 6.2]). We also define

$$\Gamma_{ij}^{\ k}(\xi) = \Gamma_{ijm}(\xi)g^{mk}(\xi),$$
$$\Gamma_{ij}^{*k}(\xi) = \Gamma^*_{ijm}(\xi)g^{mk}(\xi), \tag{71}$$

where $(g^{ij}(\xi))$ is the inverse of $(g_{ij}(\xi))$ and the Einstein summation convention (see for example [3, p.21]) is used. This implies that

$$\nabla_{\frac{\partial}{\partial\xi^i}} \frac{\partial}{\partial\xi^j} = \Gamma_{ij}^{\ k}(\xi)\frac{\partial}{\partial\xi^k}, \quad \nabla^*_{\frac{\partial}{\partial\xi^i}} \frac{\partial}{\partial\xi^j} = \Gamma_{ij}^{*k}(\xi)\frac{\partial}{\partial\xi^k}.$$

We also let $\frac{\partial\xi}{\partial\eta} = \left(\frac{\partial\xi^i}{\partial\eta_j}\right)$ and $\frac{\partial\eta}{\partial\xi} = \left(\frac{\partial\eta_i}{\partial\xi^j}\right)$ be the Jacobian matrices of the transition maps. They are inverses of each other.

5.2 Dualistic structure of $L^{(\alpha)}$-divergence

5.2.1 The Riemannian metric

Proposition 4 *The coordinate representation $G(\xi) = \left(g_{ij}(\xi)\right)$ of the Riemannian metric g is given by*

$$G(\xi) = \frac{-1}{\Pi^{(\alpha)}(\xi, \eta)} \left(I - \frac{\alpha}{\Pi^{(\alpha)}(\xi, \eta)} \eta \xi^\top \right) \frac{\partial \eta}{\partial \xi}. \tag{72}$$

Its inverse is given by

$$G^{-1}(\xi) = \left(g^{ij}(\xi)\right) = -\Pi^{(\alpha)}(\xi, \eta) \frac{\partial \xi}{\partial \eta} \left(I + \alpha \eta \xi^\top \right). \tag{73}$$

Proof Using the self-dual representation (68), we have

$$\frac{\partial}{\partial \xi^i} \mathbf{D}\left[\xi : \xi'\right] = \frac{\eta_i'}{\Pi(\xi, \eta')} - \frac{\partial \varphi}{\partial \xi^i}(\xi),$$
$$\frac{\partial^2}{\partial \xi^i \partial \xi'^j} \mathbf{D}\left[\xi : \xi'\right] = \frac{1}{\Pi(\xi, \eta')^2} \left(\Pi(\xi, \eta') \frac{\partial \eta_i'}{\partial \xi'^j} - \alpha \eta_i' \xi^\ell \frac{\partial \eta_\ell'}{\partial \xi'^j} \right). \tag{74}$$

Setting $\xi = \xi'$ and expressing in matrix form, we obtain (72).

We obtain (73) by taking the inverse of (72). For the middle term, we may use the Sherman–Morrison formula (see (77) below) to get

$$\left(I - \frac{\alpha}{\Pi(\xi, \eta)} \eta \xi^\top \right)^{-1} = I + \alpha \eta \xi^\top.$$

$\square$

Remark 6 In Proposition 4 the Riemannian matrix is given in terms of the Jacobian matrix $\frac{\partial \eta}{\partial \xi}$. This is to emphasize the role of duality (compare (72) with [3, (1.66)]) and to enable explicit expressions of the Christoffel symbols $\Gamma_{ij}{}^k = \Gamma_{ijm} g^{mk}$ which appear in the primal geodesic equations.

Let $\xi \in \Omega$ and $v \in \mathbb{R}^d$. By direct differentiation of (28), we have

$$\left. \frac{d^2}{dt^2} \mathbf{D}\left[\xi + tv : \xi\right] \right|_{t=0} = \frac{1}{2} v^\top \left(-D^2 \varphi(\xi) - \alpha (D\varphi(\xi))(D\varphi(\xi))^\top \right) v. \tag{75}$$

Thus we also have

$$g_{ij}(\xi) = -\frac{\partial^2 \varphi}{\partial \xi^i \partial \xi^j}(\xi) - \alpha \frac{\partial \varphi}{\partial \xi^i}(\xi) \frac{\partial \varphi}{\partial \xi^j}(\xi) \tag{76}$$

which is the positive definite matrix in (25). Using the Sherman–Morrison formula

$$\left(A + uv^\top \right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}, \tag{77}$$

we can invert $G(\xi)$ to get

$$G^{-1}(\xi) = -(D^2\varphi)^{-1} + \alpha \frac{(D^2\varphi)^{-1}(D\varphi)(D\varphi)^\top (D^2\varphi)^{-1}}{1 + \alpha(D\varphi)^\top (D^2\varphi)^{-1}(D\varphi)}.$$

While this formula is explicit, it is difficult to use in differential-geometric computations. The same remark applies to the connections.

5.2.2 Affine connections

Proposition 5 *The Christoffel symbols of the primal connection are given by*

$$\Gamma_{ijk}(\xi) = \frac{\alpha}{\Pi^{(\alpha)}(\xi, \eta)^2} \eta_j \frac{\partial \eta_i}{\partial \xi^k} + \frac{\alpha}{\Pi^{(\alpha)}(\xi, \eta)^2} \eta_i \frac{\partial \eta_j}{\partial \xi^k}$$
$$- \frac{2\alpha^2}{\Pi^{(\alpha)}(\xi, \eta)^3} \eta_i \eta_j \xi^\ell \frac{\partial \eta_\ell}{\partial \xi^k}, \tag{78}$$

$$\Gamma_{ij}{}^k(\xi) = \frac{-\alpha}{\Pi^{(\alpha)}(\xi, \eta)} \left(\eta_i \delta_j^k + \eta_j \delta_i^k \right) = -\alpha \left(\frac{\partial \varphi}{\partial \xi^i} \delta_j^k + \frac{\partial \varphi}{\partial \xi^j} \delta_i^k \right). \tag{79}$$

where $\delta^{\cdot}_{\cdot}$ is the Kronecker delta.

Proof From our previous computation (74), we have

$$\frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi'^k} \mathbf{D}\left[\xi : \xi' \right] = \frac{1}{\Pi(\xi, \eta')} \frac{\partial \eta_i'}{\partial \xi'^k} - \frac{\alpha}{\Pi(\xi, \eta')^2} \eta_i' \xi^\ell \frac{\partial \eta_\ell'}{\partial \xi'^k}.$$

Differentiating one more time and writing $\Pi = \Pi(\xi, \eta')$, we get

$$\frac{\partial^2}{\partial \xi^i \partial \xi^j} \frac{\partial}{\partial \xi'^k} \mathbf{D}\left[\xi : \xi' \right] = \frac{-\alpha}{\Pi^2} \eta_j' \frac{\partial \eta_i'}{\partial \eta_k'} - \frac{\alpha}{\Pi^2} \eta_i' \frac{\partial \eta_j'}{\partial \xi'^k} + \frac{2\alpha^2}{\Pi^3} \eta_i' \eta_j' \xi^\ell \frac{\partial \eta_\ell'}{\partial \xi'^k}.$$

Setting $\xi = \xi'$, we obtain (78).

To prove (79), we first use (73) to write

$$g^{mk}(\xi) = -\Pi(\xi, \eta) \left(\frac{\partial \xi^m}{\partial \eta_k} + \alpha \frac{\partial \xi^m}{\partial \eta_\ell} \eta_\ell \xi^k \right).$$

It follows that

$$
\begin{aligned}
\Gamma_{ij}{}^k(\xi) &= \Gamma_{ijm}(\xi)g^{mk}(\xi) \\
&= -\left(\frac{\alpha}{\Pi}\eta_j\frac{\partial\eta_i}{\partial\xi^m} + \frac{\alpha}{\Pi}\eta_i\frac{\partial\eta_j}{\partial\xi^m} - \frac{2\alpha^2}{\Pi^2}\eta_i\eta_j\xi^\ell\frac{\partial\eta_\ell}{\partial\xi^m}\right)\left(\frac{\partial\xi^m}{\partial\eta_k} + \alpha\frac{\partial\xi^m}{\partial\eta_\ell}\eta_\ell\xi^k\right) \\
&= \frac{-\alpha}{\Pi}\eta_j\delta_i^k - \frac{\alpha}{\Pi}\eta_i\delta_j^k + \frac{2\alpha^2}{\Pi^2}\eta_i\eta_j\xi^k - \frac{2\alpha^2}{\Pi}\eta_i\eta_j\xi^k + \frac{2\alpha^3}{\Pi^2}\eta_i\eta_j\xi^k\xi^\ell\eta_\ell.
\end{aligned}
$$

It can be verified that the last three terms cancel out and the resulting expression gives the first equality in (78). The second equality follows from the definition of η as the α-gradient (34) as well as the identity (41). □

5.3 Dualistic structure of $L^{(-\alpha)}$-divergence

Here we state the corresponding results for the $L^{(-\alpha)}$-divergence.

Proposition 6 *For the dualistic structure generated by* $\mathbf{D}^{(-\alpha)}$ *where* φ *is* α-exponentially convex, we have

$$
g_{ij}(\xi) = \frac{\partial^2\varphi}{\partial\xi^i\partial\xi^j}(\xi) + \alpha\frac{\partial\varphi}{\partial\xi^i}(\xi)\frac{\partial\varphi}{\partial\xi^j}(\xi), \tag{80}
$$

$$
\Gamma_{ij}{}^k(\xi) = -\alpha\left(\frac{\partial\varphi}{\partial\xi^i}(\xi)\delta_j^k + \frac{\partial\varphi}{\partial\xi^j}(\xi)\delta_i^k\right). \tag{81}
$$

Note that (79) and (81) have the same form (even though here φ is α-exponentially convex). This is because $\Gamma_{ij}{}^k = \Gamma_{ijm}g^{mk}$ and the changes of signs cancel out in the product.

6 Geometric consequences

With all the coefficients available, we derive in this section the geometric properties of the dualistic structure (g, ∇, ∇^*). We state the results for any $\mathbf{D}^{(\pm\alpha)}$ with $\alpha > 0$, but the proofs will again be given for the case of $\mathbf{D}^{(\alpha)}$. In Sect. 6.4 we specialize to the unit simplex and explain the connections with the α-divergence.

6.1 Dual projective flatness

The following definition is taken from [25].

Definition 11 (*Projective flatness*) Let ∇ and $\widetilde{\nabla}$ be torsion-free affine connections on a smooth manifold. They are projectively equivalent if there exists a differential 1-form τ such that

$$
\nabla_X Y = \widetilde{\nabla}_X Y + \tau(X)Y + \tau(Y)X \tag{82}
$$

for any smooth vector fields X and Y. We say that ∇ is projectively flat if ∇ is projectively equivalent to a flat connection (a connection whose Riemann–Christoffel curvature tensor vanishes).

If we write $\tau = a_i(\xi)d\xi^i$, say, using the primal coordinate system, then (82) is equivalent to

$$\Gamma_{ij}{}^k(\xi) = \widetilde{\Gamma}_{ij}{}^k(\xi) + a_i(\xi)\delta_j^k + a_j(\xi)\delta_i^k. \tag{83}$$

Theorem 14 *For any* $\mathbf{D}^{(\pm\alpha)}$, *the induced primal connection* ∇ *and the dual connection* ∇^* *are both projectively flat.*

Proof We only consider the primal connection. Consider the 1-form defined by $\tau(X) = -\alpha X\varphi$, where $X\varphi$ means the derivative of φ in the direction of X. In primal coordinates, we have

$$\tau = -\sum_{i=1}^{d} \alpha \frac{\partial \varphi}{\partial \xi^i} d\xi^i. \tag{84}$$

Let $\widetilde{\nabla}$ be the flat Euclidean connection with respect to the coordinate system ξ such that its Christoffel symbols satisfy $\widetilde{\Gamma}_{ij}{}^k(\xi) \equiv 0$. From (83), we see that ∇ is projectively equivalent to $\widetilde{\nabla}$ and thus ∇ is projectively flat. $\square$

Motivated by Theorem 14, we say that the induced dualistic structures are dually projectively flat. Projective flatness can be related to the behaviors of the geodesics. Recall that a (smooth) curve $\gamma : [0, 1] \to \mathcal{M}$ is said to be a primal geodesic if $\nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) \equiv 0$. Equivalently, its primal coordinate representation $\xi(t) = (\xi^1(t), \dots, \xi^d(t))$ satisfies the primal geodesic equations

$$\ddot{\xi}^k(t) + \dot{\xi}^i(t)\dot{\xi}^j(t)\Gamma_{ij}{}^k(\xi(t)) = 0, \quad k = 1, \dots, d, \tag{85}$$

where the dots denote derivatives with respect to time. Similarly, the curve γ is a dual geodesic if $\nabla^*_{\dot{\gamma}(t)}\dot{\gamma}(t) \equiv 0$.

The following result is well-known in differential geometry.

Lemma 2 *If* ∇ *and* $\widetilde{\nabla}$ *are projectively equivalent, then a* ∇-*geodesic is a* $\widetilde{\nabla}$-*geodesic up to a time reparameterization, and vice versa.*

Corollary 2 *For the dualistic structure induced by* $\mathbf{D}^{(\pm\alpha)}$, *if* γ *is a primal (dual) geodesic, then its trace in the primal (dual) coordinate system is a straight line.*

Thus the primal geodesics are straight lines under the primal coordinate system but run at non-constant speeds. Let us consider the time reparameterization. Using (79), we note that the primal geodesic equation (85) (under the primal coordinate system) is equivalent to the single vector equation

$$\ddot{\xi}(t) = 2\alpha\dot{\xi}(t)\frac{d}{dt}\varphi(\xi(t)). \tag{86}$$

From this we immediately see that the trace of the primal geodesic is a straight line under the coordinate ξ.

Proposition 7 *Let $\gamma : [0, 1] \to \mathcal{M}$ be a primal geodesic. Then, in primal coordinates, we have*

$$\xi(t) = (1 - h(t))\xi(0) + h(t)\xi(1), \tag{87}$$

where $h : [0, 1] \to [0, 1]$ is given by

$$h(t) = \frac{\int_0^t e^{2\alpha\varphi(\xi(s))} ds}{\int_0^1 e^{2\alpha\varphi(\xi(s))} ds}. \tag{88}$$

Proof We already know that a primal geodesic can be written in the form (87) for some time change h. Plugging this into the geodesic equation (86), we have

$$\ddot{h}(t) = 2\alpha \dot{h}(t) \frac{d}{dt}\varphi(\xi(t)) \Rightarrow \frac{d}{dt}\log \dot{h}(t) = 2\alpha \frac{d}{dt}\varphi(\xi(t)).$$

Integrating and using the fact that $h(0) = 0$ and $h(1) = 1$, we obtain the desired formula (88).   □

A natural question is the relationship between the primal and dual coordinate vector fields. This is studied in the following proposition.

Proposition 8

(i) *For $\mathbf{D}^{(\alpha)}$, the inner product between $\frac{\partial}{\partial \xi^i}$ and $\frac{\partial}{\partial \eta_j}$ is*

$$\left\langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \eta_j} \right\rangle = \frac{-1}{\Pi(\xi, \eta)}\delta_{ij} + \frac{\alpha}{\Pi(\xi, \eta)^2}\xi^j \eta_i. \tag{89}$$

(ii) *For $\mathbf{D}^{(-\alpha)}$, the inner product between $\frac{\partial}{\partial \xi^i}$ and $\frac{\partial}{\partial \eta_j}$ is*

$$\left\langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \eta_j} \right\rangle = \frac{1}{\Pi(\xi, \eta)}\delta_{ij} - \frac{\alpha}{\Pi(\xi, \eta)^2}\xi^j \eta_i. \tag{90}$$

Proof Consider (i). Using Proposition 4, we compute

$$\left\langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \eta_j} \right\rangle = \left\langle \frac{\partial}{\partial \xi^i}, \sum_m \frac{\partial \xi^m}{\partial \eta_j}\frac{\partial}{\partial \xi^m} \right\rangle = \sum_m \frac{\partial \xi^m}{\partial \eta_j}\left\langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \xi^m} \right\rangle$$

$$= \sum_m \frac{\partial \xi^m}{\partial \eta_j}\left(\frac{-1}{\Pi(\xi, \eta)}\frac{\partial \eta_i}{\partial \xi^m} + \frac{\alpha}{\Pi(\xi, \eta)^2}\eta_i \sum_\ell \xi^\ell \frac{\partial \eta_\ell}{\partial \xi^m} \right).$$

We obtain (89) under some simplification.   □

The formulas (89) and (90) are quite interesting. In the dually flat case (i.e., $\mathbf{D}^{(0\pm)}$, also see [3, Theorem 6.6]), the two coordinate fields are orthonormal in the sense that

$\left\langle \frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \eta_j} \right\rangle = \pm \delta_{ij}$. When $\alpha > 0$, the first term of the inner product is conformal to $\pm \delta_{ij}$, but there is an extra term related to duality. This formula is the key ingredient in the proof of the generalized Pythagorean theorem.

6.2 Curvatures

Next we consider the primal Riemann–Christoffel curvature tensor R defined by

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \tag{91}$$

where X, Y, Z are vector fields and $[X, Y] = XY - YX$ is the Lie bracket. Under the primal coordinate system, we use the notation

$$R \left(\frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \xi^j} \right) \frac{\partial}{\partial \xi^k} = R_{ijk}{}^\ell \frac{\partial}{\partial \xi^\ell}.$$

It can be shown (see for example [3, Sect. 5.8]) that the coefficients are given by

$$R_{ijk}{}^\ell = \partial_i \Gamma_{jk}{}^\ell - \partial_j \Gamma_{ik}{}^\ell + \Gamma_{jk}{}^m \Gamma_{im}{}^\ell - \Gamma_{ik}{}^m \Gamma_{jm}{}^\ell. \tag{92}$$

The dual curvature tensor R^* is defined analogously.

Proposition 9

(i) *For* $\mathbf{D}^{(\alpha)}$, *the primal curvature tensor is given by*

$$R_{ijk}{}^\ell(\xi) = \alpha \left(g_{ik}(\xi) \delta_j^\ell - g_{jk}(\xi) \delta_i^\ell \right). \tag{93}$$

(ii) *For* $\mathbf{D}^{(-\alpha)}$, *the primal curvature tensor is given by*

$$R_{ijk}{}^\ell(\xi) = -\alpha \left(g_{ik}(\xi) \delta_j^\ell - g_{jk}(\xi) \delta_i^\ell \right). \tag{94}$$

Proof This is a straightforward but lengthy computation using (79), (81) and (92). We omit the details. □

Definition 12 When $\dim \mathcal{M} \geq 2$, a torsion-free connection has constant sectional curvature $k \in \mathbb{R}$ with respect to the metric g if its curvature tensor R satisfies the identity

$$R(X, Y)Z = k (\langle Y, Z \rangle X - \langle X, Z \rangle Y) \tag{95}$$

for all vector fields X, Y and Z.

Using coordinate vector fields, we see that R has constant curvature k if and only if in any coordinate system we have

$$R_{ijk}{}^\ell = k \left(g_{jk} \delta_i^\ell - g_{ik} \delta_j^\ell \right). \tag{96}$$

Comparing this with (93) and (94), we obtain

Theorem 15 *If $d = \dim \mathcal{M} \geq 2$, for $\mathbf{D}^{(\pm\alpha)}$ the primal and dual connections have constant sectional curvature $\mp\alpha$ with respect to the induced metric.*

It is known that for a general dualistic structure $(\mathcal{M}, g, \nabla, \nabla^*)$, ∇ has constant sectional curvature k with respect to g if and only if the same statement holds for ∇^*. For a proof see [12, Proposition 8.1.4]. Some general properties of dualistic structures with constant curvatures are given in [12, Theorem 9.7.2]. For example, it is known that such a manifold is conjugate symmetric, conjugate Ricci-symmetric, and the connections ∇ and ∇^* are equiaffine. In Sect. 7, we show that if we assume in addition that the manifold is dually projectively flat in the sense of Theorem 14, then the geometry can be characterized elegantly in terms of the $L^{(\pm\alpha)}$-divergence.

6.3 Generalized Pythagorean theorem

The generalized Pythagorean theorem is a fundamental result of information geometry and has numerous applications in information theory, statistics and machine learning. It was first proved for the Bregman divergence ($\mathbf{D}^{(\pm\alpha)}$ in our context) which induces a dually flat geometry. For an exposition of this beautiful result see [3, Chapter 1]. In [32,33] we introduced the L-divergence ($L^{(\alpha)}$-divergence when $\alpha = 1$) on the unit simplex and proved the generalized Pythagorean theorem. In [33] the proof is quite different from that of the Bregman case and is rather involved because there we expressed the dualistic structure in terms of the exponential coordinate system. Here we give a unified and simplified treatment covering all $L^{(\pm\alpha)}$-divergences.

Theorem 16 *Consider any $L^{(\pm\alpha)}$-divergence $\mathbf{D} = \mathbf{D}^{(\pm\alpha)}$ and the induced dualistic structure. Let $p, q, r \in \Omega$ and suppose that the dual geodesic from q to p exists (this amounts to say that the line segment between η_q and η_p lies in the dual domain Ω'. We need this assumption because Ω' may not be convex). Then the generalized Pythagorean relation*

$$\mathbf{D}[q : p] + \mathbf{D}[r : q] = \mathbf{D}[r : p] \tag{97}$$

holds if and only if the primal geodesic from q to r and the dual geodesic from q to p meet g-orthogonally at q.

Proof Again we consider the case of $\mathbf{D}^{(\alpha)}$; the proof for $\mathbf{D}^{(-\alpha)}$ is similar. By the self-dual representation (Theorem 9), we have

$$\mathbf{D}[q : p] = \frac{1}{\alpha} \log\left(1 + \alpha \xi_q \cdot \eta_p\right) - \varphi(\xi_q) - \psi(\eta_p)$$

and similarly for $\mathbf{D}[r : q]$ and $\mathbf{D}[r : p]$. Using these expressions and the identity (36), we see that the Pythagorean relation (97) holds if and only if

$$\left(1 + \alpha \xi_q \cdot \eta_p\right)\left(1 + \alpha \xi_r \cdot \eta_q\right) = \left(1 + \alpha \xi_r \cdot \eta_p\right)\left(1 + \alpha \xi_q \cdot \eta_q\right).$$

Expanding and simplifying, we have

$$(\xi_r - \xi_q) \cdot (\eta_p - \eta_q) = \alpha(\xi_q \cdot \eta_p)(\xi_r \cdot \eta_q) - \alpha(\xi_r \cdot \eta_p)(\xi_q \cdot \eta_q). \tag{98}$$

On the other hand, consider the primal geodesic from q to r. By projective flatness, in primal coordinates, the initial velocity vector (given by the inverse exponential map and expressed using the primal coordinate system) is proportional to $\xi_r - \xi_q$. Similarly, in dual coordinates, the initial velocity of the dual geodesic from q to p is proportional to $\eta_p - \eta_q$. By Proposition 8, we see that the two geodesics are orthogonal at q if and only if

$$
\begin{aligned}
0 &= \left\langle (\xi_r^i - \xi_q^i)\frac{\partial}{\partial \xi^i}, (\eta_{p,j} - \eta_{q,j})\frac{\partial}{\partial \eta_j} \right\rangle \\
&= (\xi_r^i - \xi_q^i)(\eta_{p,j} - \eta_{q,j})\left(\frac{-1}{\Pi}\delta_{ij} + \frac{\alpha}{\Pi^2}\eta_{q,i}\xi_q^j \right) \\
&= \frac{-1}{\Pi}(\xi_r - \xi_q) \cdot (\eta_p - \eta_q) + \frac{\alpha}{\Pi^2}\left((\xi_r^i - \xi_q^i)\eta_{q,i} \right)\left((\eta_{p,j} - \eta_{q,j})\xi_q^j \right),
\end{aligned}
$$

where $\Pi = \Pi(\xi_q, \eta_q) = 1 + \alpha\xi_q \cdot \eta_q$. Rearranging, we have

$$
\begin{aligned}
(1 + \alpha\xi_q \cdot \eta_q)(\xi_r \cdot \eta_p - \xi_r \cdot \eta_q - \xi_q \cdot \eta_p + \xi_q \cdot \eta_q) \\
= \alpha(\xi_r \cdot \eta_q - \xi_q \cdot \eta_q)(\xi_q \cdot \eta_p - \xi_q \cdot \eta_q).
\end{aligned}
$$

The proof is completed by verifying that this and (98) are equivalent.   □

6.4 The α-divergence

In this subsection we specialize to the unit simplex

$$\mathcal{S}_d = \{p = (p^0, p^1, \ldots, p^d) \in (0, 1)^{1+d} : p^0 + \cdots + p^d = 1\}.$$

The following definition is taken from [3, (3.39)]. Also see [9, Definition 2.9].

Definition 13 (α-*divergence*) Let $\alpha \neq \pm 1$. The α-divergence is defined by

$$\mathbf{D}_\alpha [p : q] = \frac{4}{1 - \alpha^2}\left(1 - \sum_{i=0}^d (p^i)^{\frac{1-\alpha}{2}}(q^i)^{\frac{1+\alpha}{2}} \right), \quad p, q \in \mathcal{S}_d. \tag{99}$$

We also consider the discrete Rényi divergence given by

$$\mathbf{D}_{\tilde{\alpha}} (p\|q) = \frac{1}{\tilde{\alpha} - 1}\log\left(\sum_{i=0}^d (p^i)^{\tilde{\alpha}}(q^i)^{1-\tilde{\alpha}} \right). \tag{100}$$

Note the difference in the notations to avoid confusion. The following elementary lemma (whose proof is omitted) shows that the two divergences are monotone transformations of each other.

Lemma 3 *For $\tilde{\alpha} \in (0, 1) \cup (1, \infty)$, let $\alpha = 1 - 2\tilde{\alpha} \in (-\infty, -1) \cup (-1, 1)$. Then for $p, q \in \mathcal{S}_d$ we have*

$$\mathbf{D}_{\tilde{\alpha}}(p||q) = \frac{1}{\tilde{\alpha} - 1} \log\left(1 + \tilde{\alpha}(\tilde{\alpha} - 1)\mathbf{D}_\alpha[p:q]\right). \tag{101}$$

As in Example 2, we may express $\mathbf{D}_{\tilde{\alpha}}(p||q)$ as the $L^{(\tilde{\alpha}-1)}$-divergence of the $\mathcal{F}^{(\tilde{\alpha}-1)}$-potential function when $\tilde{\alpha} \in (1, \infty)$, and as the $L^{(-(1-\tilde{\alpha}))}$-divergence of the $\mathcal{F}^{-(1-\tilde{\alpha})}$-potential function when $\tilde{\alpha} \in (0, 1)$.

Let α and $\tilde{\alpha}$ be as in Lemma 3. Let (g, ∇, ∇^*) be the dualistic structure induced on $\mathcal{S}_d$ by the α-divergence $\mathbf{D}_\alpha[p:q]$, and let $(\tilde{g}, \tilde{\nabla}, \tilde{\nabla}^*)$ be that induced by the Rényi divergence $\mathbf{D}_{\tilde{\alpha}}(p||q)$. Since the two divergences are related by a monotone transformation, from (70), (101) and the chain rule we have $\tilde{g}_{ij} = \tilde{\alpha} g_{ij}$, $\tilde{\Gamma}_{ijk} = \tilde{\alpha}\Gamma_{ijk}$ and $\tilde{\Gamma}^*_{ijk} = \tilde{\alpha}\Gamma^*_{ijk}$. (Here the multiplier $\tilde{\alpha}$ is the derivative of the transformation (101) at 0.) This implies that $\Gamma_{ij}{}^k = \tilde{\Gamma}_{ij}{}^k$ and $\Gamma^{*k}_{ij} = \tilde{\Gamma}^{*k}_{ij}$, so the two dualistic structures have the same primal and dual geodesics and the same primal and dual curvature tensors (see (92)).

Theorem 17 *For $\alpha = 1 - 2\tilde{\alpha} \in (-\infty, -1) \cup (-1, 1)$, the statistical manifold $(\mathcal{S}_d, g, \nabla, \nabla^*)$ induced by the α-divergence (99) is dually projectively flat with constant curvature $\frac{1-\alpha^2}{4}$.*

Proof We only indicate how the curvature is determined, as the rest follows immediately from other results of this section. We have $\tilde{\alpha} = \frac{1-\alpha}{2}$. Suppose $\alpha \in (-\infty, -1)$. The Rényi divergence $\mathbf{D}_{\tilde{\alpha}}(p||q)$ is then an $L^{(\alpha')}$-divergence where $\tilde{\alpha} = 1 + \alpha' \in (1, \infty)$. By Theorem 15, we have

$$\tilde{R}(X, Y)Z = -\alpha'\left(\tilde{g}(Y, Z) - \tilde{g}(X, Z)Y\right).$$

But $R = \tilde{R}$ and $\tilde{g} = \tilde{\alpha}g$. So we have

$$R(X, Y)Z = -\alpha'\tilde{\alpha}\left(g(Y, Z) - g(X, Z)\right) = \frac{1-\alpha^2}{4}\left(g(Y, Z) - g(X, Z)\right).$$

This shows that the curvature is $\frac{1-\alpha^2}{4}$. The case $\alpha \in (-1, 1)$ is similar. □

In this context, our geometry reduces to the geometry of the α-divergence, and Theorem 16 implies the corresponding Pythagorean theorem. See in particular [2, p. 179] where the curvature $\frac{1-\alpha^2}{4}$ is determined using affine differential geometry. Their (8.67) is the product form of our (97) for the Rényi divergence (101). Thus, our results provide a new approach to study the geometry (compare with [2, Sect. 3.6]). We plan to carry out a deeper study in future papers.

7 Dually projectively flat manifolds with constant curvatures

Using the expressions given in (70), any divergence $\mathbf{D}[\cdot : \cdot]$ on a manifold $\mathcal{M}$ defines a dualistic structure $(\mathcal{M}, g, \nabla, \nabla^*)$. This operation is not one-to-one; it is easy to construct examples where two different divergences induce the same dualistic structure on $\mathcal{M}$. An interesting question is whether we can single out a divergence which is in some sense most natural for a given dualistic structure. If so, we call it a canonical divergence. For dually flat manifolds, it is possible to show that there is a canonical divergence of Bregman type (see [3, Sect. 6.6]). Thus, a dualistic structure is dually flat if and only if it is (locally) induced by a Bregman divergence.

In this section we ask and solve the same question where the manifold is dually projectively flat with constant curvature. We assume that the manifold $\mathcal{M}$ has dimension $d \geq 2$, and both ∇ and ∇^* are projectively flat and have constant sectional curvature $\pm \alpha$ where $\alpha > 0$ (note that if $\alpha = 0$ the manifold is dually flat and we reduce to the classical setting). Our aim is to characterize the geometry based on these properties. The key idea is to use these properties to extend in a novel way the proofs of [3, Theorem 6.2] and [9, Theorem 4.3] which address the dually flat case. Using tools of affine differential geometry, an alternative characterization is given in [21, Corollary 1].

7.1 Dual coordinates and potential functions

Theorem 18 *Consider a statistical manifold $(\mathcal{M}, g, \nabla, \nabla^*)$ such that both ∇ and ∇^* are projectively flat and have constant negative sectional curvature $-\alpha$ where $\alpha > 0$. Then, near each point of $\mathcal{M}$, there exist local coordinate systems ξ and η such that the following statements hold:*

(i) *The primal geodesics are straight lines in the ξ coordinates up to time reparameterizations, and the dual geodesics are straight lines in the η coordinates up to time reparameterizations.*
(ii) *The coordinates satisfy the constraint $1 + \alpha \sum_{\ell=1}^{d} \xi^\ell \eta_\ell > 0$.*
(iii) *There exist local α-exponentially concave functions $\varphi(\xi)$ and $\psi(\eta)$ such that the generalized Fenchel identity*

$$\varphi(\xi) + \psi(\eta) \equiv c^{(\alpha)}(\xi, \eta) = \frac{1}{\alpha} \log(1 + \alpha \xi \cdot \eta) \tag{102}$$

holds true.
(iv) *Under the coordinate systems ξ and η respectively, the Riemannian metric is given by*

$$
\begin{aligned}
g_{ij}(\xi) &= -\frac{\partial^2 \varphi}{\partial \xi^i \partial \xi^j}(\xi) - \alpha \frac{\partial \varphi}{\partial \xi^i}(\xi) \frac{\partial \varphi}{\partial \xi^j}(\xi), \\
g_{ij}(\eta) &= -\frac{\partial^2 \psi}{\partial \eta_i \partial \eta_j}(\eta) - \alpha \frac{\partial \psi}{\partial \eta_i}(\eta) \frac{\partial \psi}{\partial \eta_j}(\eta).
\end{aligned}
\tag{103}
$$

If the constant curvature is $+\alpha > 0$, analogous statements hold where the functions φ and ψ in (iii) *are α-exponentially convex, and in* (103) *the negative signs are replaced by positive signs.*

Proof Let the curvature be $-\alpha < 0$ (the proof for positive curvature is similar). Consider the primal connection ∇ which is projectively flat. By Definition 11, there exists a flat connection $\widetilde{\nabla}$ and a 1-form τ such that (82) holds. As $\widetilde{\nabla}$ is flat, near each point there exists an affine coordinate system $\xi = (\xi^1, \ldots, \xi^d)$ under which the $\widetilde{\nabla}$-geodesics are constant velocity straight lines. By projective equivalence, the primal geodesics are straight lines under the ξ coordinates up to time reparametrizations. Furthermore, if we write $\tau = a_i(\xi)d\xi^i$ where the $a_i(\xi)$'s are smooth functions of ξ, by (83) we have

$$\Gamma_{ij}{}^k(\xi) = a_i(\xi)\delta_j^k + a_j(\xi)\delta_i^k. \tag{104}$$

Motivated by our previous result (84), we want to show that

$$a_i(\xi) = -\alpha \frac{\partial \varphi}{\partial \xi^i}(\xi) \tag{105}$$

for some α-exponentially concave function φ. To this end we will use the assumption that ∇ has constant sectional curvature $-\alpha$ with respect to g.

Using the representation (104) of the primal Christoffel symbols, we compute

$$\begin{aligned}
R_{ijk}{}^\ell(\xi) &= \partial_i \Gamma_{jk}{}^\ell - \partial_j \Gamma_{ik}{}^\ell + \Gamma_{jk}{}^m \Gamma_{im}{}^\ell - \Gamma_{ik}{}^m \Gamma_{jm}{}^\ell \\
&= (\partial_i a_j - \partial_j a_i)\delta_k^\ell + (\partial_i a_k - a_k a_i)\delta_j^\ell - (\partial_j a_k - a_k a_j)\delta_i^\ell.
\end{aligned} \tag{106}$$

On the other hand, since ∇ has constant curvature $-\alpha$, by (96) we have

$$R_{ijk}{}^\ell(\xi) = \alpha \left(g_{ik}(\xi)\delta_j^\ell - g_{jk}(\xi)\delta_i^\ell \right). \tag{107}$$

Equating (106) and (107) and using the assumption that $\dim \mathcal{M} = d \geq 2$ (so that we can pick different indices), we see that

$$\frac{\partial a_j}{\partial \xi^i}(\xi) = \frac{\partial a_i}{\partial \xi^j}(\xi) \tag{108}$$

and

$$\alpha g_{ij}(\xi) = \partial_i a_j(\xi) - a_i(\xi)a_j(\xi). \tag{109}$$

By (108), the 1-form $\tau = a_i(\xi)d\xi^i$ is closed. So, locally, it is exact. Thus, there exists a locally defined function $\varphi(\xi)$ such that our claim (105) holds. In particular, the 1-form τ is locally given by $\tau(X) = -\alpha X \varphi$.

Now we may write the metric (109) in the form

$$g_{ij}(\xi) = -\frac{\partial^2 \varphi}{\partial \xi^i \partial \xi^j}(\xi) - \alpha \frac{\partial \varphi}{\partial \xi^i}(\xi) \frac{\partial \varphi}{\partial \xi^j}(\xi). \tag{110}$$

Since the matrix $(g_{ij}(\xi))$ is strictly positive definite, from (24) we have that $D^2 e^{\alpha\varphi} < 0$. In particular, φ is locally α-exponentially concave.

Without loss of generality we may assume 0 belongs to the domain of φ. This implies (see (27)) that

$$1 - \alpha D\varphi(\xi) \cdot \xi > 0$$

and the α-gradient

$$\eta := \nabla^{(\alpha)}\varphi(\xi) = \frac{1}{1 - \alpha D\varphi(\xi) \cdot \xi} \nabla\varphi(\xi) \tag{111}$$

is well-defined. As in (41) we have $1 + \alpha\xi \cdot \eta > 0$.

With the dual coordinate η defined by (111), we define the dual function $\psi(\eta)$ by the generalized Fenchel identity:

$$\psi(\eta) = c^{(\alpha)}(\xi, \eta) - \varphi(\xi). \tag{112}$$

By α-duality ψ is locally α-exponentially concave, and the Fenchel identity (102) holds by construction.

Since the dual connection ∇^* is uniquely determined given g and ∇ (see [3, (6.6)]), with the results in Sect. 5 it is a now routine exercise to verify that η is an affine coordinate system for the projectively flat dual connection ∇^*, and the metric is also given by the second formula in (103). This completes the proof of the theorem. □

7.2 Canonical divergence

Using Theorem 18, we can define locally a canonical divergence which generalizes the one for a dually flat manifold. It is interesting to know whether this canonical divergence coincides with the one defined by Ay and Amari in [8] for a generic dualistic structure.

Theorem 19 (canonical divergence) *Consider the setting of Theorem 18 where $\alpha > 0$. Let ξ, η, φ and ψ be given as in the theorem.*

(i) *If the curvature is $-\alpha < 0$, we define the local $L^{(\alpha)}$-divergence*

$$\mathbf{D}[q : p] = \frac{1}{\alpha} \log\left(1 + \alpha\xi_q \cdot \eta_p\right) - \varphi(\xi_q) - \psi(\eta_p). \tag{113}$$

(ii) *If the curvature is $\alpha > 0$, we define the local $L^{(-\alpha)}$-divergence*

$$\mathbf{D}[q : p] = \varphi(\xi_q) + \psi(\eta_p) - \frac{1}{\alpha} \log\left(1 + \alpha\xi_q \cdot \eta_p\right). \tag{114}$$

Then these divergences are independent of the choices of ξ, η, φ and ψ. Furthermore, the canonical divergence induces the given Riemannian metric and the primal and dual connections.

Proof Note that the primal affine coordinate system ξ is unique up to an affine transformation. The rest are then determined by (111) and (112). By a direct computation, we can verify that (113) and (114) remain invariant. $\square$

References

1. Adams, S., Dirr, N., Peletier, M.A., Zimmer, J.: From a large-deviations principle to the Wasserstein gradient flow: a new micro-macro passage. Commun. Math. Phys. **307**(3), 791 (2011)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. Translations of Mathematical Monographs, vol. 191. Am. Math. Soc./Oxford University Press, Providence/London (2000)
3. Amari, S.I.: Information Geometry and Its Applications. Springer, Berlin (2016)
4. Amari, S.I., Karakida, R., Oizumi, M.: Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem (2018) **(to appear in Information Geometry)**
5. Amari, S.I., Ohara, A.: Geometry of q-exponential family of probability distributions. Entropy **13**(6), 1170–1185 (2011)
6. Ambrosio, L., Gigli, N.: A User's Guide to Optimal Transport. In: Modelling and Optimisation of Flows on Networks. Lecture Notes in Mathematics, vol 2062. Springer, Berlin, Heidelberg (2013). https://link.springer.com/chapter/10.1007/978-3-642-32160-3_1
7. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Spaces of Probability Measures, Second Edition. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag (2008). https://www.springer.com/us/book/9783764387211
8. Ay, N., Amari, S.I.: A novel approach to canonical divergences within information geometry. Entropy **17**(12), 8111–8129 (2015)
9. Ay, N., Jost, J., Lê, H., Schwachhöfer, L.: Information Geometry. Springer, Berlin (2017)
10. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with bregman divergences. J. Mach. Learn. Res. **6**(Oct), 1705–1749 (2005)
11. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**(4), 375–417 (1991)
12. Calin, O., Udrişte, C.: Geometric modeling in probability and statistics. Springer, Berlin (2014)
13. Chen, Y., Li, W.: Natural gradient in Wasserstein statistical manifold (2018). arXiv preprint arXiv:1805.08380
14. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Unbalanced optimal transport: geometry and kantorovich formulation (2015). arXiv preprint arXiv:1508.05216
15. Dillen, F., Nomizu, K., Vranken, L.: Conjugate connections and radon's theorem in affine differential geometry. Monatshefte für Mathematik **109**(3), 221–235 (1990)
16. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**(3), 793–803 (1983)
17. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**(3), 631–647 (1992)
18. Erbar, M., Kuwada, K., Sturm, K.T.: On the equivalence of the entropic curvature–dimension condition and Bochner's inequality on metric measure spaces. Invent. Math. **201**(3), 993–1071 (2015)
19. Felice, D., Ay, N.: Towards a canonical divergence within information geometry (2018). arXiv preprint arXiv:1806.11363
20. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
21. Kurose, T.: Dual connections and affine geometry. Mathematische Zeitschrift **203**(1), 115–121 (1990)
22. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tohoku Math. J. Second Ser. **46**(3), 427–433 (1994)

23. Léonard, C.: From the Schrödinger problem to the Monge–Kantorovich problem. J. Funct. Anal. **262**(4), 1879–1920 (2012)
24. Li, W., Montufar, G.: Ricci curvature for parametric statistics via optimal transport (2018). arXiv preprint arXiv:1807.07095
25. Matsuzoe, H.: Geometry of contrast functions and conformal geometry. Hiroshima Math. J. **29**(1), 175–191 (1999)
26. Moreau, J.J.: Inf-convolutions, sous-additive, convexite des fonctions numeriques. J. Math. Anal. Appl. **49**, 109–154 (1970)
27. Nagaoka, H., Amari, S.I.: Differential geometry of smooth families of probability distributions. Tech. Rep. METR 82–7. University of Tokyo, Tokyo (1982)
28. Naudts, J.: The q-exponential family in statistical physics. In: Journal of Physics: Conference Series, vol. 201, p. 012003. IOP Publishing (2010). http://iopscience.iop.org/article/10.1088/1742-6596/201/1/012003/meta
29. Naudts, J., Zhang, J.: Rho-tau embedding and gauge freedom in information geometry (2018). https://doi.org/10.1007/s41884-018-0004-6
30. Pal, S.: Embedding optimal transports in statistical manifolds. Indian J. Pure Appl. Math. **48**(4), 541–550 (2017)
31. Pal, S., Wong, T.K.L.: Energy, entropy, and arbitrage (2013). arXiv preprint arXiv:1308.5376
32. Pal, S., Wong, T.K.L.: The geometry of relative arbitrage. Math. Fin. Econ. **10**, 263–293 (2016)
33. Pal, S., Wong, T.K.L.: Exponentially concave functions and a new information geometry. Ann. Probab. **46**(2), 1070–1113 (2018)
34. Pal, S., Wong, T.K.L.: Multiplicative Schrödinger problem and the Dirichlet transport (2018). arXiv preprint arXiv:1807.05649
35. Rachev, S.T., Rüschendorf, L.: Mass Transportation Problems: Volume I: Theory, vol. 1. Springer, Berlin (1998)
36. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California (1961)
37. Santambrogio, F.: Optimal Transport for Applied Mathematicians. Springer, Berlin (2015)
38. Van Erven, T., Harremos, P.: Rényi divergence and Kullback–Leibler divergence. IEEE Trans. Inf. Theory **60**(7), 3797–3820 (2014)
39. Villani, C.: Topics in Optimal Transportation. American Mathematical Society, Rhode Island (2003)
40. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2008)
41. Wong, T.K.L.: Optimization of relative arbitrage. Ann. Financ. **11**(3–4), 345–382 (2015)
42. Wong, T.K.L.: On portfolios generated by optimal transport (2017). arXiv preprint arXiv:1709.03169
43. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**(1), 159–195 (2004)
44. Zhang, J.: Referential duality and representational duality on statistical manifolds. In: Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan, vol. 1216, p. 5867 (2005)
45. Zhang, J.: Reference duality and representation duality in information geometry. In: AIP Conference Proceedings, vol. 1641, pp. 130–146. AIP (2015)