Tsallis and Rényi Deformations Linked via a New λ-Duality

Ting-Kam Leonard Wong and Jun Zhang

***Abstract*—Tsallis and Rényi entropies, which are monotone transformations of each other, are deformations of the celebrated Shannon entropy. Maximization of these deformed entropies, under suitable constraints, leads to the q-exponential family which has applications in non-extensive statistical physics, information theory and statistics. In previous information-geometric studies, the q-exponential family was analyzed using classical convex duality and Bregman divergence. In this paper, we show that a generalized λ-duality, where $\lambda = 1 - q$ is to be interpreted as the constant information-geometric curvature, leads to a generalized exponential family which is essentially equivalent to the q-exponential family and has deep connections with Rényi entropy and optimal transport. Using this generalized convex duality and its associated logarithmic divergence, we show that our λ-exponential family satisfies properties that parallel and generalize those of the exponential family. Under our framework, the Rényi entropy and divergence arise naturally, and we give a new proof of the Tsallis/Rényi entropy maximizing property of the q-exponential family. We also introduce a λ-mixture family which may be regarded as the dual of the λ-exponential family, and connect it with other mixture-type families. Finally, we discuss a duality between the λ-exponential family and the λ-logarithmic divergence, and study its statistical consequences.**

***Index Terms*—Tsallis entropy, Rényi entropy, exponential family, q-exponential family, mixture family, logarithmic divergence, convex duality, information geometry, optimal transport.**

I. Introduction

EPONENTIAL families of probability distributions play important roles in probability, statistics, information theory, statistical physics, among other fields [1], [2]. An *exponential family* is a parameterized probability density (with respect to a given reference measure ν) of the form

$$p(x;\theta) = e^{\theta \cdot F(x) - \phi(\theta)}, \tag{I.1}$$

where $F = (F_1, \dots, F_d)$ is a vector of statistics, $\theta \cdot F = \sum_{i=1}^d \theta_i F_i$ is the dot product, and $\phi(\theta)$ is the cumulant generating function. In statistical physics, the divisive normalization $Z(\theta) = e^{\phi(\theta)}$ is called the partition function [3]. Densities of the form (I.1) maximize the Shannon entropy subject to constraints on the expected value of F, thus generalizing the Boltzmann-Gibbs distribution.

Ting-Kam Leonard Wong (tkl.wong@utoronto.ca) is with the Department of Statistical Sciences, University of Toronto, Canada. Jun Zhang (junz@umich.edu) is with the Department of Psychology and Department of Statistics, University of Michigan, USA.

Manuscript received July 20, 2021; revised January 12, 2022.

Another very useful family of probability distributions is the *mixture family*. Let $p_0(x), p_1(x), \dots, p_d(x)$ be given densities. The mixture family consists of the convex combinations

$$p(x;\eta) = \sum_{i=0}^d \eta_i p_i(x), \tag{I.2}$$

where the mixture parameters satisfy $\eta_i \geq 0$ and $\sum_{i=0}^d \eta_i = 1$. The exponential and mixture families are closely related, as both families are dually flat in the information-geometric sense and can be analyzed via convex duality, Shannon entropy and the Kullback-Leibler (KL) divergence; see Section II-A.

To account for non-exponential – especially power-law – behaviours in statistical physics, Tsallis [4] introduced[1] a generalized entropy, now called the *Tsallis entropy*, which for a given parameter $q \in \mathbb{R} \setminus \{1\}$ (the entropic index) and a given reference measure ν, is defined by

$$\mathbf{H}_q^{\text{Tsallis}}(p) = \frac{1}{q-1}\left(1 - \int p^q(x) \mathrm{d}\nu(x)\right), \tag{I.3}$$

where p is a probability density with respect to ν. The Tsallis entropy is closely related to the *Rényi entropy* [6], which for $q \in (0,\infty) \setminus \{1\}$ is given by

$$\mathbf{H}_q^{\text{Rényi}}(p) = \frac{1}{1-q} \log\left(\int p^q(x) \mathrm{d}\nu(x)\right). \tag{I.4}$$

When $q > 0$, the Tsallis and Rényi entropies are monotonic functions of each other:

$$\mathbf{H}_q^{\text{Tsallis}}(p) = \frac{1}{1-q}\left(e^{(1-q)\mathbf{H}_q^{\text{Rényi}}(p)} - 1\right). \tag{I.5}$$

So maximizing the Tsallis entropy is equivalent to maximizing the Rényi entropy. Note that letting $q \to 1$ in (I.3) and (I.4) recovers the classical *Shannon entropy*

$$\mathbf{H}(p) = -\int p(x) \log p(x) d\nu(x). \tag{I.6}$$

Thus, both the Tsallis and Rényi entropies are *deformations* of the Shannon entropy. Some arguments that favour the Rényi entropy as a physical concept are given in [3, Section 9.3]; for example, the Rényi entropy satisfies the additive property under independence but the Tsallis entropy does not.

Maximization of Tsallis entropy, subject to constraints on the *escort expectation* (see Section II-B), leads to a generalized exponential family called the *q-exponential family*. The idea is to consider a *deformation* of the exponential function [7].

[1] See [5, Remark 4.1.4] for a more complete historical discussion.

For $q \in \mathbb{R}\setminus\{1\}$, define the *q-exponential function* $\exp_q : \mathbb{R} \to [0,\infty]$ by

$$\exp_q(t) = [1+(1-q)t]_+^{1/(1-q)}, \tag{I.7}$$

where $t_+ = \max\{t, 0\}$ and by convention $0^t = \infty$ for $t < 0$. We have $\frac{d}{dt}\exp_q(t) = [\exp_q(t)]^q$ when $1+(1-q)t > 0$; so $\exp_q$ is convex if and only if $q > 0$. For this reason, it is natural to restrict q to be strictly positive as in e.g. [3], [8] and later in this paper. The q-exponential family, which is an instance of the more general ϕ-exponential family to be mentioned below, is then defined as the following parameterized density with *subtractive* normalization:

$$p(x;\theta) = \exp_q\left(\theta \cdot F(x) - \phi_q(\theta)\right). \tag{I.8}$$

The normalization $\int p_\theta d\nu = 1$ determines the function $\phi_q(\theta)$ which we call the *(subtractive) q-potential function*. Note that when $q \to 1$ we recover the exponential family (I.1). Contrary to the classical exponential family, the support of the density (I.8) may depend on the parameter θ; a specific example, namely the *q-Gaussian distribution*, is discussed in Example III.17. In the literature the q-exponential function is also used to define parameterized densities under *divisive* normalization, i.e.,

$$p(x;\vartheta) = \frac{1}{Z_q(\vartheta)}\exp_q(\vartheta \cdot F(x)), \tag{I.9}$$

where ϑ is an alternative parameter; see [9, Section 4] for examples. In Propositions III.7 and III.8, we give precise relationships between the parameters θ and ϑ. So far, systematic information-geometric studies of the q-exponential families as well as other deformed exponential families typically apply convex duality to the *subtractive* potential function which can be shown to be convex under suitable conditions [8], [10]. By *Tsallis deformation* we mean the classical framework under which (I.8) is analyzed using standard convex duality; see Section II-B.

Distributions of the form (I.8) or (I.9) have found numerous applications in statistical physics [11]; a specific example is the momentum distribution of cold atoms in dissipative optical lattices [12]. Recently, the q-exponential family has also been applied to statistics and machine learning; see for example [13], [14], [15], [16], [17], [18]. In the literature one can find more general formulations of deformed exponential families such as the ϕ-exponential family [19], the conjugate (ρ,τ)-embedding [20], [21], [22] by the second author, and the U-model [23]. Note that all of these families are studied under subtractive normalization as in (I.8). In this paper, we focus on the q-exponential family which is the simplest deformation to the exponential function and has many applications.

On the other hand, motivated by optimal transport [24] and the duality between Bregman divergence and exponential family [25], the first author considered in [26] some deformed exponential families called the $\mathcal{F}^{(\pm\alpha)}$*-families*, where $\alpha > 0$ and

$$p(x;\vartheta) = \begin{cases} (1+\alpha\vartheta\cdot F(x))^{\frac{-1}{\alpha}}e^{\varphi(\vartheta)} & (\mathcal{F}^{(\alpha)}\text{-family}) \\ (1+\alpha\vartheta\cdot F(x))^{\frac{1}{\alpha}}e^{-\varphi(\vartheta)} & (\mathcal{F}^{(-\alpha)}\text{-family}) \end{cases} \tag{I.10}$$

The similarity between this and (I.9) will be explained below; see in particular (I.12). While the use of divisive normalization is not new, the novelty of this approach is to analyze the *divisive* potential $\varphi(\vartheta)$ using a *generalized convex duality* motivated by optimal transport. For example, in the $\mathcal{F}^{(\alpha)}$ case it can be shown under suitable conditions that $e^{\alpha\varphi}$ is concave; following [26], we say that this φ is α*-exponentially concave*. Under this framework, which adopts a divisive normalization and a generalized duality, the Rényi entropy and divergence arise naturally. Hence, we call this approach *Rényi deformation*. In this paper, we present a novel framework, consisting of the λ-duality and the λ-exponential family, that unifies the $\mathcal{F}^{(\pm\alpha)}$-families and links them with the Tsallis deformation.

A. Outline

The main idea of this paper is a λ*-duality* which is a deformation of the usual convex duality reviewed in Section II-A. The λ parameter is related to the classical q parameter by $\lambda = 1-q$ and this relation is always assumed in the paper. In a nutshell, instead of convex functions, for $\lambda \neq 0$ fixed we work with functions f, defined on a given open convex set of $\mathbb{R}^d$, such that $\frac{1}{\lambda}(e^{\lambda f}-1)$ is convex. These functions are related to exponentially convex and exponentially concave functions in the literature [26], [27] (see Remark II.4). Also, instead of the convex conjugate we use the λ*-conjugate* given by

$$\begin{aligned} f^{c_\lambda}(v) &= \sup_u\{-c_\lambda(u,v) - f(u)\} \\ &= \sup_u\left\{\frac{1}{\lambda}\log(1+\lambda u\cdot v) - f(u)\right\}, \end{aligned} \tag{I.11}$$

where $c_\lambda(u,v) = -\frac{1}{\lambda}\log(1+\lambda u\cdot v)$ is the cost function in the sense of optimal transport. It deforms the usual convex duality and, as we will see below, is naturally compatible with Rényi entropy and divergence. The λ-duality also leads to the following generalization of Bregman divergence:

$$\mathbf{L}_{\lambda,f}[u:u'] = f(u) - f(u') - \frac{1}{\lambda}\log(1+\lambda\nabla f(u')\cdot(u-u')).$$

We call this the λ*-logarithmic divergence*. The details of this duality, which was motivated by optimal transport and previous works of Pal and the first author [26], [27], [28], [29], [30], [31], are given in Section II.

Using the λ-duality we study the q-exponential family from a new perspective. In Section III we introduce the λ*-exponential family* (Definition III.1) which is essentially the q-exponential family under divisive normalization (see (I.9)):

$$p(x;\vartheta) = e^{-c_\lambda(\vartheta,F(x))-\varphi_\lambda(\vartheta)} = \exp_q(\vartheta\cdot F(x))e^{-\varphi_\lambda(\vartheta)}, \tag{I.12}$$

where $q = 1-\lambda$ and ϑ is another natural parameter related to θ via $\theta = \vartheta e^{-\lambda\varphi_\lambda(\vartheta)}$. The precise relationships between (I.12) and (I.8) are given in Propositions III.7 and III.8. This λ-exponential family unifies the $\mathcal{F}^{(\pm\alpha)}$-families (I.10). For an exponential family ($\lambda \to 0$), the subtractive and divisive normalizations are the same because the exponential function satisfies the functional equation $\exp(s+t) = \exp(s)\exp(t)$. We show that the divisive representation (I.12) of the q-exponential family is naturally compatible with the λ-duality,

in the same way that convex duality describes the pairing of the log-partition function with the negative Shannon entropy for the exponential family. In particular, when $q = 1 - \lambda > 0$ and under suitable regularity conditions, the *divisive λ-potential* φ_λ defined by (I.12) is c_λ-convex and its λ-logarithmic divergence is the Rényi divergence of order q. Using the λ-duality, we give a new proof of the Rényi entropy maxmizing property under constraints on the escort expectation.

Note that we have not mentioned the "q-analogue" of the mixture family. While generalized mixture families have been considered in the literature (see [1, Section 4.2]), they are not usually studied together with deformed exponential families under a unified framework. In Section IV we study mixture-type families under the λ-duality. We first show that a λ-exponential family is closed under the α-mixture of Amari [32] (also see Example III.21). Then we introduce a new *λ-mixture family* (Definition IV.6) which is in some sense dual to the λ-exponential family.

In Section V we describe the information geometry of λ-exponential and λ-mixture families induced by the associated λ-logarithmic divergences or equivalently the Rényi divergences. Inheriting the results from [26], [28], this geometry is dually projectively flat with constant sectional curvature λ, and the divergence satisfies a generalized Pythagorean theorem. Section VI explores further the relationship between λ-exponential family and λ-logarithmic divergence, and discusses some statistical implications. Finally, in Section VII we conclude and point out several directions for further research.

II. λ-DUALITY AS A DEFORMATION OF CONVEX DUALITY

A. Review of convex duality

The key mathematical concept which underlies the exponential family, as well as previous treatments of the q-exponential family (and other deformed exponential families under subtractive normalization, as in e.g. [10]), is the Legendre duality of convex functions. To wit, the cumulant generating function $\phi(\theta)$ in (I.1), and the q-potential function $\phi_q(\theta)$ in (I.8) for $q > 0$, are convex functions of the parameter θ [8, Theorem 2] under suitable regularity conditions.

For readability we do not spell out all technical conditions in this brief review, and refer the reader to [33], [34] for a comprehensive treatment of convex analysis on Euclidean space and its application to exponential family. Given a function f on $\mathbb{R}^d$, its convex conjugate is defined by

$$f^*(v) = \sup_u \left(u \cdot v - f(u) \right), \quad v \in \mathbb{R}^d. \tag{II.1}$$

Then f is convex and lower-semicontinuous if and only if $f^{**} = f$. When f is strictly convex and differentiable, the Legendre transformation

$$v = \nabla f(u), \tag{II.2}$$

which can be motivated by the first order condition in (II.1), defines a "dual coordinate" v, and its inverse is given by $v = \nabla f^*(u)$. Brenier's theorem in optimal transport theory [24] states that the Legendre transformation is an optimal transport map under the quadratic cost $c(u, v) = \frac{1}{2}|u - v|^2$. The convex function f also induces a *Bregman divergence*, which is widely applied in statistics and machine learning, by

$$\mathbf{B}_f[u : u'] = f(u) - f(u') - \nabla f(u') \cdot (u - u') \geq 0. \tag{II.3}$$

Now consider the cumulant generating function $\phi(\theta)$ of an exponential family (I.1), where θ, the primal variable, is the natural parameter. Then ϕ is convex and the dual variable $\eta = \nabla\phi(\theta)$, under the Legendre duality, is the *expectation parameter* given by

$$\eta = \mathbb{E}_\theta[F(X)] = \int F(x) p(x; \theta) \mathrm{d}\nu(x), \tag{II.4}$$

where under $\mathbb{E}_\theta$ the random variable X is distributed according to the density $p(\cdot; \theta)$. As a function of η, the Legendre conjugate $\psi = \phi^*$ is the negative Shannon entropy, namely $\psi(\eta) = -\mathbf{H}(p_\theta) = -\int p_\theta \log p_\theta \mathrm{d}\nu$, where for notational simplicity we write $p_\theta = p(\cdot; \theta)$. Furthermore, the Bregman divergences of ϕ and ψ can be expressed as *Kullback-Leilber (KL) divergences*:

$$\mathbf{B}_\phi[\theta : \theta'] = \mathbf{B}_\psi[\eta' : \eta] = \mathbf{H}(p_{\theta'} || p_\theta), \tag{II.5}$$

where

$$\mathbf{H}(p||p') = \int p \log \frac{p}{p'} \mathrm{d}\nu. \tag{II.6}$$

Consequently, the local second order approximation of $\mathbf{B}_\phi$, which defines a Riemannian metric on the exponential family regarded as a statistical manifold, is given by the *Fisher information metric*. Explicitly, we have

$$\mathbf{B}_\phi[\theta + \Delta\theta : \theta] = \frac{1}{2} (\Delta\theta)^\top g(\theta) (\Delta\theta) + O(|\Delta\theta|^3), \tag{II.7}$$

where

$$g_{ij}(\theta) = \int \frac{\partial \log p_\theta}{\partial \theta_i} \frac{\partial \log p_\theta}{\partial \theta_j} p_\theta \mathrm{d}\nu \tag{II.8}$$

defines the Fisher information metric. Note that the same metric is obtained if we expand $\mathbf{B}_\phi[\theta : \theta + \Delta\theta]$ instead. The natural parameter θ and the expectation parameter η can be regarded as two sets of affine coordinate systems that are "dual" with respect to the metric g. The corresponding *dually flat geometry* is well-studied in information geometry [1].

Similarly, for a mixture family (I.2), it can be shown that the negative Shannon entropy $\psi(\eta) = -\mathbf{H}(p_\eta)$, where $p_\eta = p(\cdot; \eta)$, is a convex function of the mixture parameter, and its Bregman divergence is again a KL-divergence:

$$\mathbf{B}_\psi[\eta : \eta'] = \mathbf{H}(p_\eta || p_{\eta'}). \tag{II.9}$$

The induced Riemannian metric is again the Fisher metric. Thus, convex duality and Bregman divergence underlie both the exponential and mixture families.

B. Classical deformation theory

Consider now a q-exponential family (I.8) which is a deformation of the exponential family. Here, the exponential function $\exp$ is replaced by the deformed exponential $\exp_q$ given by (I.7). When $q > 0$, the q-potential function ϕ_q in (I.8) can be shown to be convex (this requires differentiability under the integral sign), and hence defines a dual variable $\eta = \nabla\phi_q(\theta)$ via the Legendre transformation. To

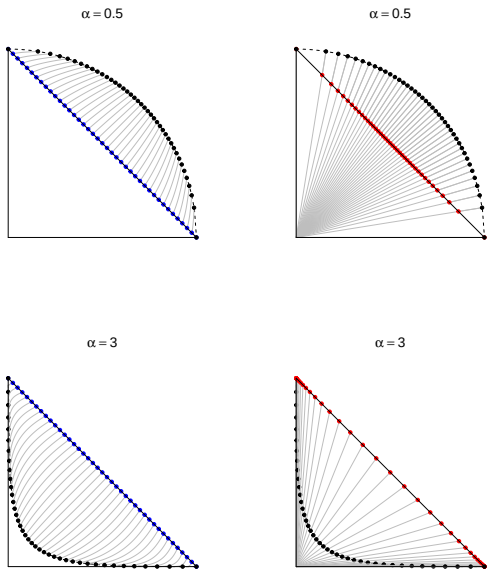


Fig. 1: Illustration of the escort transformation $p \mapsto \mathcal{E}_\alpha[p]$, where $p = (p_1, p_2)$ is a probability vector with length 2. Left: p (in blue) on the unit simplex is first mapped to $p^\alpha = (p_1^\alpha, p_2^\alpha)$ (in black). The curve (in grey) shows the trajectory $t \mapsto p^{(1-t)+t\alpha}$ for $0 \le t \le 1$. Right: Normalize p^α along the straight line (in grey) passing through the origin to obtain the escort distribution $\widetilde{p} = \mathcal{E}_\alpha[p] = p^\alpha / \int p^\alpha \mathrm{d}\nu$ (in red). (Top row: $\alpha = 0.5$. Bottom row: $\alpha = 3$.) Note that when $\alpha < 0$ the image of $p \mapsto p^\alpha$ becomes unbounded.

interpret η probabilistically we recall the concept of *escort distribution* which arises naturally in the study of generalized exponential families and their applications in non-extensive statistical physics [3]. Given a probability density $p(x)$ (or more generally a non-negative function which is not ν-almost everywhere zero) with respect to the reference measure ν, and an exponent $\alpha \in \mathbb{R} \setminus \{0\}$, we define the *escort distribution with exponent* α be the density

$$\widetilde{p} = \mathcal{E}_\alpha[p] := \frac{p^\alpha}{\int p^\alpha \mathrm{d}\nu}, \tag{II.10}$$

provided that the integral $\int p^\alpha \mathrm{d}\nu$ is finite. Some physical and information-theoretic interpretations of the escort distribution are given in [35], [36]. Now we may interpret the dual parameter η as an *escort expectation*:

$$\eta = \widetilde{\mathbb{E}}_\theta[F(X)] := \int F(x) \widetilde{p}_\theta(x) \mathrm{d}\nu(x), \tag{II.11}$$

where $\widetilde{p}_\theta = \mathcal{E}_q[p_\theta]$ is the escort distribution with exponent q. It is helpful to think of the escort transformation as the composition of two operations, namely $p \mapsto p^\alpha$ and the normalization $p^\alpha \mapsto p^\alpha / \int p^\alpha \mathrm{d}\nu$. In Figure 1 we illustrate these operations where p is a density function on a two-point set with respect to the counting measure, so p can be identified with a probability vector $p = (p_1, p_2)$. For an exponential family the escort transformation is equivalent to a dilation with respect to the natural parameter, i.e., $\mathcal{E}_\alpha[p_\theta] = p_{\alpha\theta}$, whenever $\alpha\theta$ belongs to the parameter set. This corresponds to the fact that the escort transformation is the scalar multiplication under the *Aitchison geometry* in compositional data analysis [37].

It can be shown (see e.g. [8, Section 4]) that densities of the form (I.8) maximize the Tsallis entropy under constraints on the escort expectation; in Theorem III.15, we give a new proof of this result using our λ-duality. The Tsallis (or equivalently Rényi) entropy maximization property gives a theoretical justification of the q-exponential family. Nevertheless, other fundamental properties of exponential families described in Section II-A do not have "exact" analogues in previous treatments of the q-exponential family. For example, as first shown in [8], the Bregman divergence of ϕ_q is not the Tsallis relative entropy, and the associated Riemannian metric is not the Fisher metric but is a conformal transformation of it. Also see [38] for a more recent attempt. In this paper, we show that our framework using λ-duality provides natural and elegant statements which nicely parallel the case of exponential and mixture families.

C. Generalized λ-duality

The key idea of this paper is the following. Instead of deforming the exponential and logarithm functions, we deform the notion of conjugation, and hence the convex duality, by replacing the pairing $u \cdot v$ in (II.1) by a nonlinear function $-c_\lambda$ of it. This generalized duality, which can be defined for a generic cost function $c(u, v)$, is well-known in the optimal transport literature [24], [39], [40] in characterizations of optimal transport plans. Classical Legendre duality corresponds to $c(u, v) = -u \cdot v$ which is the cross term of the quadratic cost $\frac{1}{2}|u - v|^2$ when expanded. In what follows we let a constant $\lambda \in \mathbb{R} \setminus \{0\}$ be given and, for reasons that will become clear in Section V, call it the *curvature parameter*. It is related to the q parameter via $\lambda = 1 - q$. Thus the usual exponential family (the limit as $\lambda \to 0$) corresponds to zero curvature (dual flatness). When applying the λ-duality to the λ-exponential family, we typically assume (as in Section II-B) that $q > 0$ (so that $\exp_q$ is convex), or equivalently $\lambda < 1$. The specific functional form of the logarithmic cost function c_λ is motivated by previous works of Pal and the first author [26], [27], [28], [29] which led to tractable results; see in particular [29] where we obtained an analogue of Brenier's theorem for the Dirichlet transport problem on the unit simplex. Here we streamline the treatment by introducing the unifying parameter λ (rather than $\pm\alpha$ as in (I.10)), and work instead with c-convex functions (rather than c-concave functions which are more common in optimal transport theory) so that the notations parallel those in Section II-A.

By continuity, we let $\log t = -\infty$ for $t \le 0$ and $e^{-\infty} = 0$. In particular, we have

$$e^{\log t} = t_+, \quad t \in \mathbb{R}. \tag{II.12}$$

Definition II.1 (λ-duality). *Fix $\lambda \in \mathbb{R} \setminus \{0\}$.*

(i) We define $c_\lambda : \mathbb{R}^d \times \mathbb{R}^d \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ by

$$c_\lambda(u, v) = \frac{-1}{\lambda} \log(1 + \lambda u \cdot v). \tag{II.13}$$

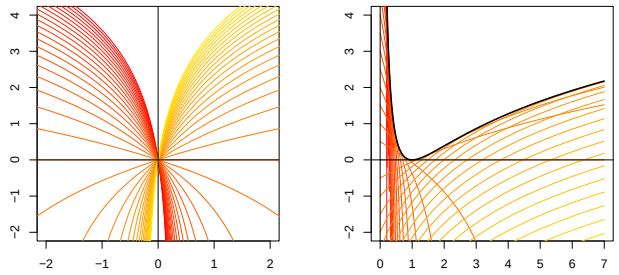


Fig. 2: Illustration of λ-duality on the real line where $\lambda = 0.5$. Left: Graphs of the mappings $u \mapsto \frac{1}{\lambda}\log(1+\lambda uv)$ for several values of v in the interval $[-5,5]$. Right: The function $f(u) = \frac{1}{\lambda}\left(\frac{1}{u} - 1 + \log u\right)$, which is c_λ-convex on $\Omega = (0,\infty)$, shown as the upper envelope of functions of the form $\frac{1}{\lambda}\log(1+\lambda uv) - g(v)$, where $g(v) = v$. Note that f is not convex in the usual sense.

(ii) If $f : \Omega \subset \mathbb{R}^d \to \bar{\mathbb{R}}$, we define $f^{c_\lambda} : \mathbb{R}^d \to \bar{\mathbb{R}}$ by

$$f^{c_\lambda}(v) = \sup_{u\in\Omega} \left\{-c_\lambda(u,v) - f(u)\right\}, \quad v \in \mathbb{R}^d, \quad \text{(II.14)}$$

where by convention we set $+\infty - (+\infty) = -\infty - (-\infty) = -\infty$. We call f^{c_λ} the c_λ-conjugate of f.

(iii) A function $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is c_λ-convex on Ω if $f \not\equiv +\infty$ and $f = g^{c_\lambda}$ on Ω for some $\Omega' \subset \mathbb{R}^d$ and $g : \Omega' \to \mathbb{R} \cup \{+\infty\}$.

This generalized duality is illustrated in Figure 2 where $\lambda = 0.5$. Note that for $u, v \in \mathbb{R}^d$ we have $\lim_{\lambda\to 0} c_\lambda(u,v) = u \cdot v$. Thus, when $\lambda \to 0$ we recover the usual convex duality. Observe also that

$$e^{-c_\lambda(u,v)} = (1 + \lambda u \cdot v)_+^{1/\lambda} = \exp_q(u \cdot v), \quad \text{(II.15)}$$

where $q = 1 - \lambda$. As will be seen in Section III, this and the λ-duality allow us to give an alternative treatment of the q-exponential family without deforming the exponential function.

In optimal transport (see e.g. [41]), the operation (II.14) for a generic cost function $c(x,y)$ is called the c_--transform (the c_+-transform, which involves an infimum rather than a supremum, leads to c-concave functions). In general, it is not possible to characterize c-concave and c-convex functions explicitly in terms of familiar convexity concepts. The cost function c_λ is quite special as shown by the following result which specifies a class of "nice" c_λ-functions useful for our applications. Note that in Definition II.1(iii) f and g are defined on respective domains $\Omega, \Omega' \subset \mathbb{R}^d$. In the context of Theorem II.2 below, this allows us to deduce that $1+\lambda u\cdot v > 0$ for $(u,v) \in \Omega \times \Omega'$ and avoid infinity values. Developing the duality and differential theory in full generality along the lines of Rockafellar's classic treatise [34] is of independent mathematical interest, and is left for future research. Also see the discussion in Section VI.

Theorem II.2 (Main results of λ-duality)**.** *Let $\lambda \neq 0$ and let $\Omega \subset \mathbb{R}^d$ be an open convex set. Consider a smooth function $f : \Omega \to \mathbb{R}$ such that the Hessian of $F_\lambda = \frac{1}{\lambda}(e^{\lambda f} - 1)$ is strictly positive definite (hence F_λ is strictly convex) and $1 - \lambda\nabla f(u)\cdot u > 0$ on Ω. Then:*

(i) f is c_λ-convex function on Ω.

(ii) Define for $u \in \Omega$ the mapping

$$v = \nabla^{c_\lambda} f(u) := \frac{1}{1 - \lambda\nabla f(u)\cdot u}\nabla f(u). \quad \text{(II.16)}$$

Then $\nabla^{c_\lambda} f$ is a diffeomorphism from Ω onto its range Ω' which is an open set. We call $\nabla^{c_\lambda} f$ the λ-gradient.

(iii) Consider f^{c_λ} as a function on the range $\Omega' = \nabla^{c_\lambda} f(\Omega)$. Then $(f^{c_\lambda})^{c_\lambda} = f$ on Ω.

(iv) For $u \in \Omega$ and $v = \nabla^{c_\lambda} f(u) \in \Omega'$ we have $1+\lambda u\cdot v > 0$ and the following identity holds:

$$f(u) + f^{c_\lambda}(v) \equiv -c_\lambda(u,v) = \frac{1}{\lambda}\log(1+\lambda u\cdot v). \quad \text{(II.17)}$$

In particular, f^{c_λ} is smooth on Ω'.

(v) The inverse of $\nabla^{c_\lambda} f$ is given by

$$\nabla^{c_\lambda} f^{c_\lambda}(v) := \frac{1}{1 - \lambda\nabla f^{c_\lambda}(v)\cdot v}\nabla f^{c_\lambda}(v),$$

which is well-defined on Ω'.

Proof. See [26, Section 3.3]. Here we rephrased the results in terms of the λ-duality. To illustrate some of the ideas involved, we provide the proof of (i) in the Appendix. □

The definition of the λ-gradient $\nabla^{c_\lambda} f(u)$ can be motivated by the optimality condition in (II.14) (compare with (II.1)). The λ-gradient (analogous to the Brenier map) can be interpreted as an optimal transport map under the logarithmic cost c_λ. For the geometric meaning of the condition $1 - \lambda\nabla f(u)\cdot u > 0$ see the proof of Theorem II.2(i) in the Appendix. Analytically, it allows us to apply convex/concave duality to $e^{\lambda f}$ and then take logarithm to obtain a generalized convex duality based on the logarithmic cost c_λ. This is essentially a normalization which makes 0 a reference point (see the left panel of Figure 2). By Theorem III.9, it holds for the divisive λ-potential of the λ-exponential family under suitable regularity conditions. Geometrically, $f = \frac{1}{\lambda}\log(1+\lambda F_\lambda)$ is simply a multiple of the logarithm of a positive convex/concave function (depending on the sign of λ). By Theorem II.2(ii), it is given on Ω as the supremum of a collection of vertically translated logarithmic functions (right panel of Figure 2).

It is convenient to introduce a terminology for the functions that satisfy the hypotheses of Theorem II.2.

Definition II.3 (Regular c_λ-convex function)**.** *By a regular c_λ-convex function we mean a function f which satisfies the hypotheses of Theorem II.2.*

In the context of Theorem II.2, the usual convexity is replaced by convexity of the transformation $F_\lambda = \frac{1}{\lambda}(e^{\lambda f} - 1)$, and the λ-gradient $\nabla^{c_\lambda} f$ defines a new dual variable. Note that the additive term $-1/\lambda$ in F_λ is not necessary and is included so that as $\lambda \to 0$ we have $F_\lambda \to f$.

If f is C^2 (twice continuously differentiable), then F_λ is convex if and only if the matrix

$$e^{-\lambda f(u)}\nabla^2 F_\lambda(u) = \nabla^2 f(u) + \lambda(\nabla f(u))(\nabla f(u))^\top \quad \text{(II.18)}$$

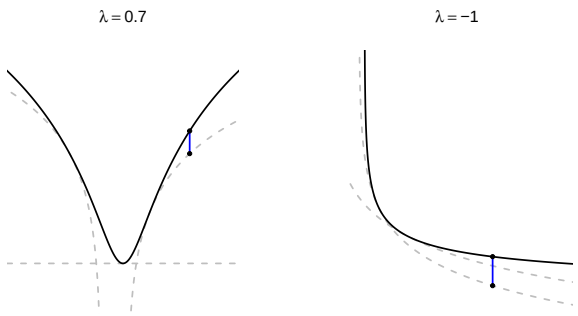


Fig. 3: Illustration of λ-logarithmic divergence where f is defined on an interval of the real line. Left: $\lambda = 0.7$. Right: $\lambda = -1$. The graph of f is shown as a solid curve (in black). The logarithmic first order approximation, which by construction supports the graph of f from below, is shown as a dashed curve (in grey). Its deviation, shown as the vertical line segment in blue, gives the value of $\mathbf{L}_{\lambda,f}[u : u']$.

is positive semidefinite on Ω (here and throughout $\nabla f(u)$ is regarded as a column vector and $\cdot^\top$ denotes transposition). Note that when $\lambda < 0$ then

$$\nabla^2 f(u) \succeq (-\lambda)(\nabla f(u))(\nabla f(u))^\top \succeq 0,$$

so that f itself is convex (here $\succeq$ is the Loewner order). In Section V, we will use (II.18) to define a Riemannian metric on Ω.

Remark II.4 (Exponential convexity and concavity). Let f be regular c_λ-convex. If $\lambda > 0$, then $e^{\lambda f}$ is convex on Ω; following [27], [26], [27], [28], we say that f is α-*exponentially convex* on Ω with $\alpha = \lambda$. If $\lambda < 0$, then $e^{|\lambda|(-f)}$ is a positive concave function, and we say that $-f$ is α-*exponentially concave* with $\alpha = |\lambda|$. The present framework unifies the two cases in [26, Section 3].

We close this subsection by making the observation (not used in the rest of the paper) that the transformation $t \mapsto \frac{1}{\lambda}\log(1+\lambda t)$, which characterizes the deformation in c_λ, as well as its inverse $s \mapsto \frac{1}{\lambda}(e^{\lambda s} - 1)$, which defines F_λ in Theorem II.2, are closely related to the *Box–Cox power transformation* $s^{(\lambda)} = \frac{1}{\lambda}(s^\lambda - 1)$ (and its inverse) introduced in [42], where the same λ parameter is used. Specifically, we have $F_\lambda = (\exp f)^{(\lambda)}$.

D. λ-logarithmic divergence

Recall the derivation of the Bregman divergence (II.3). If f is convex and differentiable, for any u, u' we have

$$f(u') + \nabla f(u') \cdot (u - u') \le f(u). \tag{II.19}$$

The Bregman divergence $\mathbf{B}_f[u : u']$ is defined by taking the difference.

Our λ-duality leads to a different divergence. Consider a function f on a convex set Ω such that $F_\lambda = \frac{1}{\lambda}(e^{\lambda f} - 1)$ is convex. If f is differentiable at $u' \in \Omega$, a convexity argument (write (II.19) for F_λ then rearrange) gives the inequality

$$\frac{1}{\lambda} + \nabla f(u') \cdot (u - u') \le \frac{1}{\lambda} e^{\lambda(f(u) - f(u'))}, \quad u \in \Omega. \tag{II.20}$$

We can define a divergence by taking logarithm on both sides and rearranging. There are two cases depending on the sign of λ, but the resulting expression is the same. The following definition unifies the $L^{(\pm\alpha)}$-divergences introduced in [26, Section 3]. A graphical illustration is shown in Figure 3. Also see [31] for the general framework of *c-divergence* which defines divergences based on optimal transport maps.

Definition II.5 (λ-logarithmic divergence)**.** *Let $\Omega \subset \mathbb{R}^d$ be convex and let $f : \Omega \to \mathbb{R}$ be a function such that $\frac{1}{\lambda}(e^{\lambda f} - 1)$ is convex. If $u, u' \in \Omega$ and f is differentiable at u', we define the λ-logarithmic divergence by*

$$\mathbf{L}_{\lambda,f}[u : u'] = f(u) - f(u') - \frac{1}{\lambda}\log(1 + \lambda \nabla f(u') \cdot (u - u')). \tag{II.21}$$

From (II.20), we have $\mathbf{L}_{\lambda,f}[u : u'] \ge 0$; also $\mathbf{L}_{\lambda,f}[u' : u'] = 0$. When $\lambda > 0$, it is possible that $1 + \lambda \nabla f(u') \cdot (u - u') \le 0$. When this happens, the definition implies that $\mathbf{L}_{\lambda,f}[u : u'] = \infty$. On the other hand, when $\lambda < 0$ we have $\mathbf{L}_{\lambda,f}[u : u'] < \infty$ for all $u, u' \in \Omega$. Clearly, if $\frac{1}{\lambda}(e^{\lambda f} - 1)$ is strictly convex, then $\mathbf{L}_{\lambda,f}[u : u']$ is strictly positive for $u \ne u'$. For later use, observe that if for a given function f the right hand side of (II.21) is non-negative for $u, u' \in \Omega$, then by reversing the argument in (II.20) we see that $\frac{1}{\lambda}(e^{\lambda f} - 1)$ is convex.

We end this section with two basic examples of our framework. The second example will be revisited in Section IV-B where we introduce the λ-mixture family.

Example II.6 (Excess growth rate). The following example, taken from [28], is the original motivation of the theory of logarithmic divergences in mathematical finance. Consider $d \ge 2$ stocks. Over a holding period, suppose that the price of stock i moves from u_i' to u_i. We have $u, u' \in \Omega = (0, \infty)^d$. Consider a portfolio with weights w_i, where $w_i \ge 0$ and $\sum_{i=1}^d w_i = 1$. Then the log return of the portfolio is given by

$$\log\left(\sum_{i=1}^d w_i \frac{u_i}{u_i'}\right).$$

By Jensen's inequality, this is greater than or equal to the weighted average log return of the stocks:

$$\mathbf{D}[u : u'] = \log\left(\sum_{i=1}^d w_i \frac{u_i}{u_i'}\right) - \sum_{i=1}^d w_i \log \frac{u_i}{u_i'} \ge 0.$$

We call this non-negative quantity the *excess growth rate* of the portfolio. Then $\mathbf{D}[u : u'] = \mathbf{L}_{-1,\varphi}[u : u']$ is the (-1)-logarithmic divergence of the function $\varphi(u) = -\sum_{i=1}^d w_i \log u_i$ which is regular c_{-1}-convex ($\lambda = -1$) on Ω. This can be generalized to other portfolios and $\lambda < 0$; see [28] and [30], where the $L^{(\alpha)}$-divergence there is equivalent to our $(-\alpha)$-logarithmic divergence.

Example II.7 (Rényi entropy and divergence). Consider the open unit simplex in $\mathbb{R}^{1+d}$ given by

$$\Delta^d = \{u = (u_0, u_1, \ldots, u_d) : u_i > 0, \sum_{i=0}^d u_i = 1\}. \tag{II.22}$$

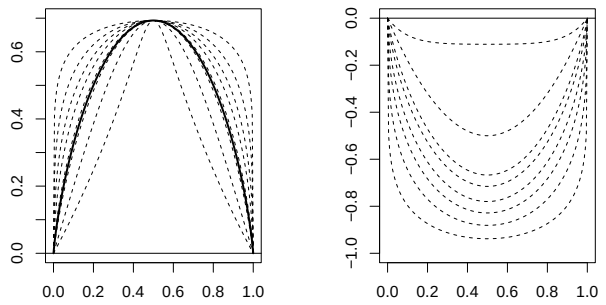


Fig. 4: Left: The Rényi entropy $\mathbf{H}_q^{\text{Rényi}}(\widetilde{p})$ of the escort distribution $\widetilde{p}$, as a function of $p \in \Delta^1$ (identified with the unit interval), for $\lambda \in \{-5, -2, -1, -0.5, -0.1, 0.1, 0.5, 0.9\}$ and $q = 1 - \lambda$. The solid line is the Shannon entropy ($\lambda = 0$ or $q = 1$). Right: Graphs of the functions $\frac{1}{\lambda}(e^{\lambda\varphi_\lambda} - 1)$, which are convex, for the same values of λ. See Example II.7 which is a special case of the λ-mixture family introduced in Section IV-B.

For $\lambda < 1$ and $\lambda \neq 0$, consider the negative (discrete) Rényi entropy of order $q = 1 - \lambda > 0$ given by

$$\varphi_\lambda(p) = -\mathbf{H}_q^{\text{Rényi}}(\widetilde{p}) = \frac{-1}{1-q} \log \sum_{i=0}^{d} \widetilde{p}_i^q, \quad p \in \Delta^d,$$

where $\widetilde{p} = \mathcal{E}_{1/q}[p]$ is the escort transformation with exponent $1/q$ (here ν is the counting measure). Then it is not difficult to verify that φ_λ is a regular c_λ-convex function on Δ^d (see Figure 4 for an illustration where $d = 1$).[2] The corresponding λ-logarithmic divergence is the *Rényi divergence* (see (III.12)) of the same order:

$$\mathbf{L}_{\lambda,\varphi_\lambda}[p : p'] = \mathbf{H}_q^{\text{Rényi}}(\widetilde{p}|\widetilde{p}') = \frac{1}{q-1} \log \sum_{i=0}^{d} (\widetilde{p}_i)^q (\widetilde{p}'_i)^{1-q}.$$

This extends the familiar fact, which we recover in the limit $\lambda \to 0$, that the (discrete) KL-divergence is the Bregman divergence associated with the negative Shannon entropy as the convex potential function.

III. λ-EXPONENTIAL FAMILY

In this section we introduce the λ-exponential family as an alternative representation of the q-exponential family tailored to the λ-duality studied in Section II. It unifies the $\mathcal{F}^{(\pm\alpha)}$-families introduced in [26, Section 4]. We also explain how the subtractive and divisive frameworks are related. Concrete examples will be given in Section III-D.

[2]We may think of φ_λ as a function of $(p_1, \ldots, p_d)$ which takes values in an open convex set of $\mathbb{R}^d$.

A. Definition and two parameterizations

Definition III.1 (λ-exponential family)**.** *Let $\lambda \neq 0$ and $q = 1 - \lambda$. The λ-exponential family is the parameterized density, with respect to a given reference measure ν, defined by*

$$\begin{aligned} p_\vartheta(x) &= p(x; \vartheta) \\ &= e^{-c_\lambda(\vartheta, F(x)) - \varphi_\lambda(\vartheta)} \\ &= \exp_q(\vartheta \cdot F(x)) e^{-\varphi_\lambda(\vartheta)}, \end{aligned} \tag{III.1}$$

where $F = (F_1, \ldots, F_d)$ is a vector of statistics. The potential function $\varphi_\lambda(\vartheta)$ is defined by the normalization $\int p(x; \vartheta) d\nu(x) = 1$. Explicitly, we have

$$\varphi_\lambda(\vartheta) = \log \int e^{-c_\lambda(\vartheta, F(x))} \mathrm{d}\nu(x). \tag{III.2}$$

The natural parameter set Ω is given by $\Omega = \{\vartheta \in \mathbb{R}^d : \varphi_\lambda(\vartheta) < \infty\}$.

Note that we use the symbol ϑ for the "canonical parameter" of the λ-exponential family, and distinguish it from θ to be introduced in (III.5) below.

Lemma III.2. *If $\lambda < 1$ (or $q > 0$), then the natural parameter set Ω is convex.*

Proof. Note that Ω is the set of $\vartheta \in \mathbb{R}^d$ such that

$$\int \exp_q(\vartheta \cdot F(x)) \mathrm{d}\nu(x) < \infty.$$

Assume $q > 0$. Then $\exp_q(\vartheta \cdot F(x))$ is convex in ϑ (see the paragraph below (I.7)). Hence Ω is a convex set. □

For the subsequent analysis we require some regularity conditions on the family. We believe some of these conditions can possibly be deduced under appropriate assumptions on the family (in the spirit of the [33]). It is also of interest to study the λ-exponential family and its regularity properties from the viewpoint of nonparametric information geometry [43]. We leave this as well as a complete treatment of λ-duality to future research.

Assumption III.3 (Regularity conditions)**.** *Let $\lambda < 1$ (or $q = 1 - \lambda > 0$). We assume that the natural parameter set Ω is a non-empty open convex set, and we may differentiate $\varphi_\lambda(\vartheta)$ as many times as needed under the integral sign.*

Remark III.4 (Interpretation of λ)*.* We have chosen to parameterize the function c_λ in terms of λ rather than $q = 1 - \lambda$. As will be explained in Section V, if we equip the λ-exponential family, regarded as a smooth manifold (with global coordinate system ϑ), with the logarithmic divergence $\mathbf{L}_{\lambda,\varphi_\lambda}$, we obtain a statistical manifold (in the sense of information geometry [1], [44]) with constant sectional curvature λ. Thus λ can be thought of as the curvature parameter. On the other hand, the constant q (when positive) is the order of the corresponding Rényi entropy and divergence. Thus both λ and q are meaningful.

Lemma III.5. *Let $\mathbb{P}_\vartheta$ be a probability measure under which the random variable X has density $p(\cdot; \vartheta)$, where $\vartheta \in \Omega$ is fixed. Then $\mathbb{P}_\vartheta(1 + \lambda\vartheta \cdot F(X) > 0) = 1$.*

Proof. We only need to note that if $1+\lambda\vartheta\cdot F(x)\le 0$ then $\exp_q(\vartheta\cdot F(x))$ is either 0 or $+\infty$. □

Note that when $1+\lambda\vartheta\cdot F(x)>0$, as in Lemma III.5, then the density (III.1) is given by

$$p_\vartheta(x)=(1+\lambda\vartheta\cdot F(x))^{\frac{1}{\lambda}}e^{-\varphi_\lambda(\vartheta)}. \tag{III.3}$$

Comparing (III.3) and (I.10), we see that the λ-exponential family coincides with the $\mathcal{F}^{(-\alpha)}$-family when $\lambda=\alpha>0$, and is equivalent to the $\mathcal{F}^{(\alpha)}$-family when $\lambda=-\alpha<0$ (and F is replaced by $-F$).

Remark III.6 (Support condition). In contrast to the exponential family, it is possible that the support of the density depends on the parameter ϑ; an explicit example is the q-Gaussian distribution discussed in Example III.17. Some derivations of this paper require that the support $\{x : p(x;\vartheta)>0\}$ is independent of ϑ. When this holds, we say that the family satisfies the *support condition*.[3]

Note that both (III.1) and (I.8) involve the q-exponential function. The difference lies in the way the density is normalized. In the q-exponential family, the q-potential function $\phi_q(\theta)$ is said to be a *subtractive* normalization (since it is defined within $\exp_q$), while for the λ-exponential family $\varphi_\lambda(\vartheta)$ is given as a *divisive* normalization. Let us call $\phi_q(\theta)$ the *substractive* q-potential, and $\varphi_\lambda(\vartheta)$ the *divisive* λ-potential.

We show that the λ-exponential family (III.1) and q-exponential family (I.8) are essentially equivalent up to some changes of parameterization. This involves some subtleties regarding the domains of the parameters. We first start with a λ-exponential family and rewrite it as a q-exponential family.

Proposition III.7. *Let $\mathcal{M}=\{p_\vartheta(x)=\exp_q(\vartheta\cdot F(x))e^{-\varphi_\lambda(\vartheta)}:\vartheta\in\Omega\}$ be a λ-exponential family satisfying the support condition (Remark III.6). On the common support, say $\mathcal{X}_0$, and with respect to the same F and dominating measure ν (restricted to $\mathcal{X}_0$), let $\mathcal{N}$ be the q-exponential family $\mathcal{N}=\{p_\theta(x)=\exp_q(\theta\cdot F(x)-\phi_q(\theta)):\theta\in\Theta\}$, where Θ is the natural (maximal) parameter set.*

For $\vartheta\in\Omega$, let $\theta=\vartheta e^{-\lambda\varphi_\lambda(\vartheta)}$. Then $\theta\in\Theta$ and $p_\vartheta(\cdot)=p_\theta(\cdot)$ on $\mathcal{X}_0$. Thus $\mathcal{M}$ may be considered as a subspace of $\mathcal{N}$. If $F_1,\dots,F_d$ are linearly independent with respect to ν on $\mathcal{X}_0$, then the mapping $\vartheta\in\Omega\mapsto\theta=\vartheta e^{-\lambda\varphi_\lambda(\vartheta)}\in\Theta$ is one-to-one.

Proof. Start with (III.1) and write, for $x\in\mathcal{X}_0$,

$$\begin{aligned}&p_\vartheta(x)\\&=\exp_q(\vartheta\cdot F(x))e^{-\varphi_\lambda(\vartheta)}\\&=[1+(1-q)\vartheta\cdot F(x)]_+^{1/(1-q)}e^{-\varphi_\lambda(\vartheta)}\\&=\left[1+(1-q)\left(\vartheta e^{-\lambda\varphi_\lambda(\vartheta)}\cdot F(x)-\frac{e^{-\lambda\varphi_\lambda(\vartheta)}-1}{-\lambda}\right)\right]_+^{\frac{1}{1-q}}.\end{aligned} \tag{III.4}$$

Introduce the parameter

$$\theta=\vartheta e^{-\lambda\varphi_\lambda(\vartheta)}, \tag{III.5}$$

and define

$$\widetilde{\phi}_q(\theta)=\frac{1}{-\lambda}(e^{-\lambda\varphi_\lambda(\vartheta)}-1). \tag{III.6}$$

Then, we have

$$\begin{aligned}p_\vartheta(x)&=\exp_q(\vartheta\cdot F(x))e^{-\varphi_\lambda(\vartheta)}\\&=\exp_q(\theta\cdot F(x)-\widetilde{\phi}_q(\theta)),\quad x\in\mathcal{X}_0.\end{aligned} \tag{III.7}$$

Since $\int_{\mathcal{X}_0}\exp_q(\theta\cdot F(x)-\widetilde{\phi}_q(\theta))\mathrm{d}\nu(x)=1$ by construction, we have $\theta\in\Theta$, $\tilde{\phi}_q=\phi_q$ and $p_\vartheta=p_\theta$. Note that the support condition allows us to integrate on the fixed set $\mathcal{X}_0$.

Suppose that $F_1,\dots,F_d$ are linearly independent (with respect to ν on $\mathcal{X}_0$). Then the mappings $\vartheta\in\Omega\mapsto p_\vartheta\in\mathcal{M}$ are $\theta\in\Theta\mapsto p_\theta\in\mathcal{N}$ are one-to-one, i.e., the two models are uniquely identifiable. Given distinct $\vartheta,\vartheta'\in\Omega$, the densities p_ϑ and $p_{\vartheta'}$ define distinct probability distributions. But $p_\vartheta=p_\theta$ and $p_{\vartheta'}=p_{\theta'}$, where $\theta=\vartheta e^{-\lambda\varphi_\lambda(\vartheta)}$ and $\theta'=\vartheta' e^{-\lambda\varphi_\lambda(\vartheta')}$. Thus θ and θ' are also distinct elements of Θ. □

Next we want to express a given q-exponential family as a λ-exponential family. Note that reversing in a direct manner the argument in (III.4) requires that $1-\lambda\phi_q(\theta)>0$. So, unfortunately, the converse of Proposition III.7 does not hold in general. Nevertheless, we show that a reparameterization can be done *locally* by possibly redefining F. Here is a precise statement.

Proposition III.8. *Consider a q-exponential family $\{p_\theta(x)=\exp_q(\theta\cdot F(x)-\phi_q(\theta)):\theta\in\Theta\}$. Let θ_0 be in the interior of Θ. If $1-\lambda\phi_q(\theta_0)>0$ or $\theta_0\neq 0$, there exists a neighborhood U of θ_0 such that each p_θ for $\theta\in U$ can be written in the form*

$$p_\theta(x)=\exp_q(\vartheta\cdot\widetilde{F}(x))e^{-\varphi_\lambda(\vartheta)}, \tag{III.8}$$

where ϑ is a function of θ, $\widetilde{F}(x)=F(x)-c\theta_0$ and $c\in\mathbb{R}$ is a constant.

Proof. We first consider the case that $1-\lambda\phi_q(\theta_0)>0$. Then there exists a neighborhood U such that $1-\lambda\phi_q(\theta)>0$ for $\theta\in U$. Reversing the computation in (III.4), we have, for $\theta\in U$,

$$\begin{aligned}p_\theta(x)&=[1+\lambda(\theta\cdot F(x)-\phi_q(\theta))]_+^{\frac{1}{1-q}}\\&=(1-\lambda\phi_q(\theta))^{\frac{1}{\lambda}}\left[1+\lambda\frac{\theta}{1-\lambda\phi_q(\theta)}\cdot F(x)\right]_+^{\frac{1}{\lambda}}.\end{aligned}$$

Writing

$$\vartheta=\frac{\theta}{1-\lambda\phi_q(\theta)}\quad\text{and}\quad\varphi_\lambda(\vartheta)=\frac{1}{-\lambda}\log(1-\lambda\phi_q(\theta)), \tag{III.9}$$

we have $p_\theta(x)=\exp_q(\vartheta\cdot F(x))e^{-\varphi_\lambda(\vartheta)}$ which is in the form of a λ-exponential family (where we may pick $c=0$).

Next suppose $1-\lambda\phi_q(\theta_0)\le 0$ and $\theta_0\neq 0$. Define $\widetilde{F}(x)=F(x)-c\theta_0$, where $c\in\mathbb{R}$ is a constant to be chosen. Fix $\epsilon>0$. Let U be a neighborhood of θ_0 on which $\theta\cdot\theta_0>\frac{1}{2}|\theta_0|^2$ and $1-\lambda\phi(\theta)>1-\lambda\phi_q(\theta_0)-\varepsilon$. For $\theta\in U$, write

$$1-\lambda\phi_q(\theta)+\lambda\theta\cdot F(x)=1-\lambda\phi(\theta)+\lambda c\theta\cdot\theta_0+\lambda\theta\cdot\widetilde{F}(x).$$

[3]The support condition is a useful technical condition which simplifies many proofs. It may be possible to remove this assumption in some results.

Choose $c \in \mathbb{R}$ such that $c\lambda > 0$ and $1-\lambda\phi(\theta_0)-\epsilon+c\frac{\lambda}{2}|\theta_0|^2 > 0$ (this requires $\theta_0 \neq 0$). Then, for $\theta \in U$, we have

$$1 - \lambda\phi(\theta) + \lambda c\theta \cdot \theta_0 > 0.$$

Now we may write

$$p_\theta(x) = (1 - \lambda\phi(\theta) + \lambda c\theta \cdot \theta_0)^{\frac{1}{\lambda}} \cdot \left[1 + \lambda \frac{\theta}{1 - \lambda\phi(\theta) + \lambda c\theta \cdot \theta_0} \cdot \widetilde{F}(x)\right]_+^{\frac{1}{\lambda}},$$

which has the form (III.8) if we let

$$\vartheta = \frac{\theta}{1 - \lambda\phi(\theta) + \lambda c\theta \cdot \theta_0}$$

and

$$\varphi_\lambda(\vartheta) = \frac{1}{-\lambda}\log(1 - \lambda\phi(\theta) + \lambda c\theta \cdot \theta_0).$$

□

Roughly speaking, the above results say that when studying local properties of the family it does not matter whether we use the subtractive (q-exponential family) or divisive (λ-exponential family) formulations.

B. Linking to Rényi entropy and Rényi divergence

While divisive normalizations of the q-exponential family had been considered before (see for example (7.13) in [3, Section 7.3]; a similar generalized exponential family is considered in [45]), its theoretical significance was not recognized because it was not paired with the λ-duality. Here, we show that the λ-duality, made possible by the following result, offers fresh insights into this family and leads naturally to Rényi entropy and divergence.

Theorem III.9. *Consider a λ-exponential family satisfying Assumption III.3. Let φ_λ be the divisive λ-potential. Then the function $\frac{1}{\lambda}(e^{\lambda\varphi_\lambda(\vartheta)} - 1)$ is convex on Ω. Moreover, we have that $1 - \lambda\nabla\varphi(\vartheta) \cdot \vartheta > 0$.*

Proof. The proof of the first statement can be found in Propositions 2 and 3 of [26] where the results are stated in terms of the $\mathcal{F}^{(\pm\alpha)}$-families. To prove the second statement, consider

$$e^{\varphi(\vartheta)} = \int [1 + \lambda\vartheta \cdot F]_+^{1/\lambda} \mathrm{d}\nu.$$

Differentiating under the integral sign, which is possible by assumption, we have

$$e^{\varphi(\vartheta)}\nabla\varphi(\vartheta) = \int [1 + \lambda\vartheta \cdot F]_+^{1/\lambda - 1} F \mathrm{d}\nu$$
$$\Rightarrow \nabla\varphi(\vartheta) = \int p(x;\theta)\frac{F}{1 + \lambda\vartheta \cdot F}\mathrm{d}\nu,$$

where the second line follows from Lemma III.5. We get

$$\begin{aligned} 1 - \lambda\nabla\varphi(\vartheta) \cdot \vartheta &= \int p(x;\theta)\left(1 - \frac{\lambda\vartheta \cdot F}{1 + \lambda\vartheta \cdot F}\right)\mathrm{d}\nu \\ &= \int p(x;\theta)\frac{1}{1 + \lambda\vartheta \cdot F}\mathrm{d}\nu \\ &= \int p(x;\theta)\frac{1}{[1 + \lambda\vartheta \cdot F]_+}\mathrm{d}\nu > 0. \end{aligned}$$

□

Condition III.10. *In the remainder of Section III we assume that Assumption III.3 holds and that the Hessian of $\frac{1}{\lambda}(e^{\lambda\varphi_\lambda(\vartheta)} - 1)$ is strictly positive definite. This implies that φ_λ is a regular c_λ-convex function and Theorem II.2 applies. We also assume that the support condition (see Remark III.6) holds and that $(F_1, \ldots, F_d)$ are linearly independent.*

By Proposition III.7, there are two potential functions, namely $\phi_q(\theta)$ and $\varphi_\lambda(\vartheta)$, associated respectively to the two representations of the density. The two potential functions and their respective dualities define apparently two dual variables, namely

$$\eta_{\text{subtractive}} = \nabla_\theta\phi_q(\theta) \text{ and } \eta_{\text{divisive}} = \nabla_\vartheta^{c_\lambda}\varphi_\lambda(\vartheta).$$

For clarity, we sometimes use ∇_u to denote the gradient with respect to the variable u. We show that the two dual variables are actually the same.

Lemma III.11. *The mapping $\vartheta \in \Omega \mapsto \theta = \vartheta e^{-\lambda\varphi_\lambda(\vartheta)}$ is a diffeomorphism from Ω onto its range.*

Proof. Note that Condition III.10 is in force. Clearly the mapping $\vartheta \mapsto \theta = \vartheta e^{-\lambda\varphi_\lambda(\vartheta)}$ is differentiable. By Proposition III.7, it is also one-to-one. Recall that the gradient is regarded as a column vector. By a direct differentiation, we see that the Jacobian is given by

$$\frac{\partial\theta}{\partial\vartheta}(\vartheta) = e^{-\lambda\varphi_\lambda(\theta)}\left(\mathbf{I}_d - \lambda\vartheta(\nabla_\vartheta\varphi_\lambda(\vartheta))^\top\right),$$

where $\mathbf{I}_d$ is the identity matrix. Since $1 - \lambda\nabla_\vartheta\varphi(\vartheta) \cdot \vartheta > 0$ by assumption, by the Shermon-Morrison formula we can invert the Jacobian as follows:

$$\left(\frac{\partial\theta}{\partial\vartheta}(\vartheta)\right)^{-1} = e^{\lambda\varphi_\lambda(\vartheta)}\left(\mathbf{I}_d + \frac{\lambda\vartheta(\nabla_\vartheta\varphi_\lambda(\vartheta))^\top}{1 - \lambda\nabla_\vartheta\varphi_\lambda(\vartheta) \cdot \vartheta}\right). \tag{III.10}$$

By the inverse function theorem, the mapping $\vartheta \mapsto \theta$ is a diffeomorphism. □

Theorem III.12. *We have*

$$\nabla_\vartheta^{c_\lambda}\varphi_\lambda(\vartheta) = \nabla_\theta\phi_q(\theta) =: \eta. \tag{III.11}$$

Thus, under both the subtractive and divisive representations, the dual variable η is the escort expectation (II.11).

Proof. Recall that the gradient is regarded as a column vector. By the lemma above, θ is a function of ϑ. Applying the chain rule to (III.6), we have

$$(\nabla_\theta\phi_q(\theta))^\top = e^{-\lambda\varphi_\lambda(\vartheta)}(\nabla_\vartheta\varphi_\lambda(\vartheta))^\top\frac{\partial\vartheta}{\partial\theta}(\theta).$$

Note that $\frac{\partial\vartheta}{\partial\theta}(\theta)$ is given by (III.10). Plugging this into the above and using (II.16), we compute

$$\begin{aligned} (\nabla_\theta\phi_q(\theta))^\top &= (\nabla_\vartheta\varphi_\lambda(\vartheta))^\top\left(\mathbf{I}_d + \frac{\lambda\vartheta(\nabla_\vartheta\varphi_\lambda(\vartheta))^\top}{1 - \lambda\nabla_\vartheta\varphi_\lambda(\vartheta) \cdot \vartheta}\right) \\ &= (\nabla_\vartheta\varphi_\lambda(\vartheta))^\top\left(1 + \frac{\lambda\nabla_\vartheta\varphi(\vartheta) \cdot \vartheta}{1 - \lambda\nabla_\vartheta\varphi_\lambda(\vartheta) \cdot \vartheta}\right) \\ &= \frac{(\nabla_\vartheta\varphi_\lambda(\vartheta))^\top}{1 - \lambda\nabla_\vartheta\varphi_\lambda(\vartheta) \cdot \vartheta} \\ &= (\nabla_\vartheta^{c_\lambda}\varphi_\lambda(\vartheta))^\top. \end{aligned}$$

Thus $\nabla_{\vartheta}^{c_\lambda}\varphi_\lambda(\vartheta)=\nabla_\theta\phi_q(\theta)$ and the theorem is proved. □

From (II.11), the dual variable η is the expected value of $F(X)$ under the escort distribution $\widetilde{p}_\vartheta$ (with exponent q). In Theorem VI.4 below, we give under suitable conditions a new probabilistic interpretation of η in terms of the *original* distribution p_ϑ without undergoing the escort transformation. Note that although the dual parameter η remains the same, the primal variables θ (subtractive case) and ϑ (divisive case) are different and are related by (III.5).

Remark III.13 (Geometric meanings of the coordinate systems ϑ and η). It is helpful to think of the primal coordinate system ϑ of a λ-exponential family as a projective affine coordinate system, in the sense that a ϑ-straight line is the trajectory of a primal geodesic (up to a time reparameterization). Similarly, the dual coordinate system is the dual projective affine coordinate system. These notions are justified by the information geometry of λ-logarithmic divergence described in Section V.

As explained in Section II-A, under the classical Legendre duality, the exponential family leads naturally to the Shannon entropy $\mathbf{H}(\cdot)$ and the KL-divergence $\mathbf{H}(\cdot||\cdot)$. For a λ-exponentialy family under the λ-duality, the natural objects are the Rényi entropy and Rényi divergence. The Rényi entropy was defined in (I.4). Recall that the *Rényi divergence* of order $q>0$ between probability densities p_1,p_2 with respect to ν is defined by

$$\mathbf{H}_q^{\text{Rényi}}(p_1||p_2)=\frac{1}{q-1}\log\int p_1(x)^q p_2(x)^{1-q}\mathrm{d}\nu(x). \tag{III.12}$$

See [46] for a useful overview of the properties of the Rényi entropy and divergence.

The following theorem illustrates the theoretical elegance of the λ-duality. Here, the Shannon entropy (the negative conjugate function for the exponential family) is replaced by the Rényi entropy, and the KL-divergence (Bregman divergence of the potential function) is replaced by the Rényi entropy. Recall that Condition III.10 is imposed.

Theorem III.14. *The λ-logarithmic divergence of the divisive λ-potential φ_λ is the Rényi divergence:*

$$\mathbf{L}_{\lambda,\varphi_\lambda}[\vartheta:\vartheta']=\mathbf{H}_q^{\text{Rényi}}(p_{\vartheta'}||p_\vartheta). \tag{III.13}$$

Moreover, the λ-conjugate of φ_λ is given on the range of $\nabla^{c_\lambda}\varphi_\lambda$ by the negative Rényi entropy:

$$\psi_\lambda(\eta)=-\mathbf{H}_q^{\text{Rényi}}(p_\vartheta),\quad \eta=\nabla^{c_\lambda}\varphi_\lambda(\vartheta). \tag{III.14}$$

Proof. See [26, Theorem 13]. □

C. Rényi entropy maximization

Let $q>0$ and a let a reference measure ν be given. Let $F=(F_1,\dots,F_d)$ be a vector of statistics. Consider the Rényi entropy maximization problem

$$\max_{P\sim\nu}\mathbf{H}_q^{\text{Rényi}}(p)\text{ subject to }\widetilde{\mathbb{E}}_P[F(X)]=y, \tag{III.15}$$

where P is a probability measure with density $p=\frac{\mathrm{d}P}{\mathrm{d}\nu}$, and $\widetilde{\mathbb{E}}_P$ is the expectation with respect to the escort distribution $\widetilde{p}=\mathcal{E}_q[p]$. The constraint $P\sim\nu$ means that the P and ν are equivalent (thus $p>0$ ν-a.e.); intuitively, we assume that the support of the distribution P is known to be that of ν. Since the Tsallis and Rényi entropies are monotonic transformations of each other (see (I.5)), problem (III.15) remains the same if we maximize instead the Tsallis entropy of order q. It is well-known that the solution to (III.15) can be written in the form of a q-exponential family; see for example the q-Max-Ent Theorem in [8]. The usual proof of this result uses Lagrange multipliers. Here we give a new proof which utilizes the λ-exponential family and the associated λ-logarithmic divergence. By letting $\lambda\to 0$ we recover the entropy maximizing property of the exponential family.

Theorem III.15. *With the given F and ν, consider the λ-exponential family* (III.1) *and suppose for some parameter $\vartheta^*\in\Omega$ we have $\widetilde{\mathbb{E}}_{\vartheta^*}[F(X)]=y$. Then the distribution P^* with $\frac{\mathrm{d}P^*}{\mathrm{d}\nu}=p_{\vartheta^*}$ is the unique solution to the Rényi entropy maximization problem* (III.15).

Proof. Let $P\sim\nu$ be a distribution which satisfies the constraint on the escort expectation, and let $p=\frac{dP}{d\nu}>0$ be its density.

Consider the Rényi divergence

$$\mathbf{H}_q^{\text{Rényi}}(p||p_{\vartheta^*})=\frac{1}{q-1}\log\int p^q p_{\vartheta^*}^{1-q}\mathrm{d}\nu.$$

Since $\lambda=1-q$, by (III.1) and Lemma III.5, we have, ν-a.e.,

$$\begin{aligned}p_{\vartheta^*}^{1-q}(x)&=[1+\lambda\vartheta^*\cdot F(x)]_+\,e^{-\lambda\varphi_\lambda(\vartheta^*)}\\&=[1+\lambda\vartheta^*\cdot F(x)]\,e^{-\lambda\varphi_\lambda(\vartheta^*)}.\end{aligned}$$

Now we compute

$$\begin{aligned}&\mathbf{H}_q^{\text{Rényi}}(p||p_{\vartheta^*})\\&=\frac{-1}{\lambda}\log\left(\int p^q\,(1+\lambda\vartheta^*\cdot F)\,\mathrm{d}\nu\right)+\varphi_\lambda(\vartheta^*)\\&=\frac{-1}{\lambda}\log\int p^q\mathrm{d}\nu-\\&\qquad\frac{1}{\lambda}\log\left(1+\lambda\vartheta^*\cdot\int\frac{p^q}{\int p^q\mathrm{d}\nu}F\mathrm{d}\nu\right)+\varphi_\lambda(\vartheta^*)\\&=-\mathbf{H}_q^{\text{Rényi}}(p)-\frac{1}{\lambda}\log(1+\alpha\vartheta^*\cdot y)+\varphi_\lambda(\vartheta^*).\end{aligned} \tag{III.16}$$

Note that in the last equality we used the assumption $\widetilde{\mathbb{E}}_P[F(X)]=y$.

On the other hand, by construction we have

$$\begin{aligned}&c=\int F\frac{p_{\vartheta^*}^q}{\int p_{\vartheta^*}^q\mathrm{d}\nu}\mathrm{d}\nu\\&\Rightarrow 1+\lambda\vartheta^*\cdot y=\int(1+\lambda\vartheta^*\cdot F)\frac{p_{\vartheta^*}^q}{\int p_{\vartheta^*}^q\mathrm{d}\mu}\mathrm{d}\mu.\end{aligned}$$

Taking logarithm and rearranging, we have

$$\begin{aligned}&\log(1+\lambda\vartheta^*\cdot y)\\&=\log\int(1+\lambda\vartheta^*\cdot F)p_{\vartheta^*}^q\mathrm{d}\nu-\log\int p_{\vartheta^*}^q\mathrm{d}\nu\\&=\log\int p_{\theta^*}e^{\lambda\varphi_\lambda(\vartheta^*)}\mathrm{d}\nu-\log\int p_\vartheta^{1+\alpha}\mathrm{d}\nu\\&=\lambda\varphi_\lambda(\vartheta^*)-\log\int p_{\vartheta^*}^q\mathrm{d}\nu.\end{aligned}$$

It follows that

$$\mathbf{H}_q^{\text{Rényi}}(p_{\vartheta^*}) = \frac{-1}{\lambda}\log(1+\lambda\vartheta^*\cdot y) + \varphi_\lambda(\vartheta^*).$$

Plugging into (III.16), we have

$$\mathbf{H}_q^{\text{Rényi}}(p_{\vartheta^*}) - \mathbf{H}_q^{\text{Rényi}}(p) = \mathbf{H}_q^{\text{Rényi}}(p||p_{\vartheta^*}) \geq 0.$$

Thus $\mathbf{H}_q^{\text{Rényi}}(p_{\vartheta^*}) \geq \mathbf{H}_q^{\text{Rényi}}(p)$ and equality holds only if $P = P_{\vartheta^*}$. This completes the proof of the theorem. □

Observe that the proof of the theorem reveals something more. Not only do we get the Rényi entropy maximizing property of the λ-exponential family, we also obtain an *explicit* expression of the optimality gap in terms of the Rényi divergence:

$$\mathbf{H}_q^{\text{Rényi}}(p_{\vartheta^*}) = \mathbf{H}_q^{\text{Rényi}}(p) + \mathbf{H}_q^{\text{Rényi}}(p||p_{\vartheta^*}). \tag{III.17}$$

Thus the Rényi divergence equals the difference of two Rényi entropies. To the best of our knowledge, this result is new. We also recall that the Rényi divergence is additive under independent sampling: Given two product probability measures $P_1 \otimes P_2$ and $Q_1 \otimes Q_2$ (with densities $(p_1 \otimes p_2)(x_1, x_2) = p_1(x_1)p_2(x_2)$ and similarly for $q_1 \otimes q_2$), we have

$$\begin{aligned}&\mathbf{H}_q^{\text{Rényi}}(p_1 \otimes p_2||q_1 \otimes q_2)\\ &= \mathbf{H}_q^{\text{Rényi}}(p_1||q_1) + \mathbf{H}_q^{\text{Rényi}}(p_2||q_2).\end{aligned} \tag{III.18}$$

Equations (III.17) and (III.18) show the advantage of using Rényi entropy and divergence (which correspond to the λ-duality) over the Tsallis entropy and divergence. We also mention the recent paper [47] which studies Rényi entropy maximization using the $L^{(\alpha)}$-divergence of [26] (which is the same as the $(-\alpha)$-logarithmic divergence) and orthogonal foliations of statistical manifolds. The foliation is studied independently in [48] and applied to nonlinear principal component analysis.

D. Examples

In this subsection we illustrate the framework with some representative probability distributions. In particular, we observe that the α-*family* studied by Amari and Nagaoka [44] can be viewed as a special case of the λ-exponential family (and hence the q-exponential family).

Example III.16 (Cauchy location-scale family). Consider the Cauchy location-scale family whose density on $\mathbb{R}$ (with respect to the Lebesgue measure) is given by

$$p(x;\mu,\sigma) = \frac{1}{\pi}\frac{\sigma}{\sigma^2+(x-\mu)^2}, \tag{III.19}$$

where $(\mu,\sigma) \in \mathbb{R}\times(0,\infty)$. Define $F(x) = (x, x^2)$. Rearranging (III.19), we may write

$$p(x;\mu,\sigma) = p(x;\vartheta) := \frac{1}{1-\vartheta\cdot F(x)}\frac{1}{\pi}\sqrt{-\vartheta_2 - \frac{\vartheta_1^2}{4}}, \tag{III.20}$$

where

$$\vartheta = (\vartheta_1,\vartheta_2) = \left(\frac{2\mu}{\mu^2+\sigma^2}, \frac{-1}{\mu^2+\sigma^2}\right).$$

Equivalently, we have $\mu = \frac{-\vartheta_1}{2\vartheta_2}$ and $\sigma^2 = \frac{-4\vartheta_2-\vartheta_1^2}{4\vartheta_2^2}$. This expresses the Cauchy family as a λ-exponential family with $\lambda = -1$. The divisive potential function is given by

$$\varphi_{-1}(\vartheta) = \frac{-1}{2}\log\left(-\vartheta_2 - \frac{\vartheta_1^2}{4}\right) + \log\pi, \tag{III.21}$$

and the natural parameter set is $\Omega = \{\vartheta\in\mathbb{R}^2 : \vartheta_1^2+4\vartheta_2 < 0\}$ whose boundary is a parabola. By a direct computation using (III.21), it can be shown that the dual variable is given by

$$\eta = \nabla^{c_{-1}}\varphi_{-1}(\vartheta) = \left(\frac{-\vartheta_1}{2\vartheta_2}, \frac{-1}{\vartheta_2}\right).$$

Similarly, for any fixed degree of freedom, Student's t-distribution parameterized by location and scale can be expressed as a λ-exponential family; see Example IV.5.

Example III.17 (q-Gaussian distribution). For $q < 3$, the (centered) q-Gaussian distribution on the real line is defined by the parameterized density

$$f(x;\vartheta) = C_q\sqrt{\vartheta}\exp_q(-\vartheta x^2), \quad x\in\mathbb{R}, \tag{III.22}$$

where ν is the Lebesgue measure, $\vartheta\in\Omega = (0,\infty)$ and $C_q > 0$ is an explicit constant depending only on the value of q (see for example [3, (7.41)]). When $q < 1$ the density is supported on the finite interval $|x| < \sqrt{\vartheta/(1-q)}$ which depends on ϑ, so the support condition is violated. For $1 < q < 3$, the density is a scaled and reparameterized version of the t-distribution with $\frac{3-q}{q-1}$ degrees of freedom. When $q = 1$, it reduces to the normal distribution $N(0,\frac{1}{2})$. When $q \geq 3$ the density cannot be normalized. It is helpful to consider $\sigma^2 = 1/\vartheta$ where $\sigma > 0$ is the scale parameter.

Comparing (III.22) with (III.1), we see that (III.22) is a λ-exponential family with $\lambda = 1-q$, $\vartheta\in(0,\infty)$ and $F(x) = -x^2$. The divisive λ-potential is given by

$$\varphi_\lambda(\vartheta) = \frac{-1}{2}\log\vartheta + C'_\lambda, \tag{III.23}$$

where $C'_\lambda = \log C_q$. It is easy to verify that φ_λ is regular c_λ-convex for all $q < 3$ (or $\lambda > -2$). In particular, for $\lambda > -2$ we have $1-\lambda\varphi'_\lambda(\vartheta)\vartheta = 1+\frac{\lambda}{2} > 0$.

The λ-logarithmic divergence is given for $\vartheta,\vartheta'\in(0,\infty)$ by

$$\mathbf{L}_{\lambda,\varphi_\lambda}[\vartheta:\vartheta'] = \frac{1}{2}\log\frac{\vartheta'}{\vartheta} - \frac{1}{\lambda}\log\left(1-\frac{\lambda}{2}\left(\frac{\vartheta}{\vartheta'}-1\right)\right). \tag{III.24}$$

See Figure 5 for a graphical illustration of this divergence for several values of λ. Note that the divergence may take value $+\infty$ when $\lambda > 0$.

The dual parameter is given by

$$\eta = \nabla^{c_\lambda}\varphi_\lambda(\vartheta) = \frac{-1}{2+\lambda}\frac{1}{\vartheta} = \frac{-1}{3-q}\sigma^2.$$

Note that we have a minus sign because $F(x) = -x^2$. This is consistent with the known value $\widetilde{\mathbb{E}}_\vartheta[X^2] = \frac{1}{3-q}\sigma^2$ (see for example [3, Problem 7.6]). Note that here this formula holds even when the support condition is violated.

Next consider the λ-conjugate of φ_λ given by

$$\psi_\lambda(\eta) = \sup_{\vartheta'>0}\left\{\frac{1}{\lambda}\log(1+\lambda\vartheta'\eta) - \left(\frac{-1}{2}\log\vartheta' + C'_\lambda\right)\right\}.$$

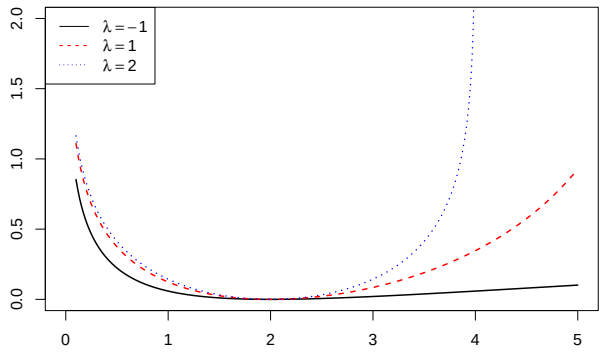


Fig. 5: The λ-logarithmic divergence $\vartheta \mapsto \mathbf{L}_{\lambda,\varphi_\lambda}[\vartheta : \vartheta_0]$ of the q-Gaussian distribution (see (III.24)), as a function of the first variable $\vartheta \in (0,\infty)$, for several values of λ. Here $\vartheta_0 = 2$.

We may optimize over $\vartheta > 0$ because $\frac{-1}{2}\log\vartheta = \infty$ for $\vartheta \le 0$. It can be verified that the λ-conjugate, which is the Rényi entropy by Theorem III.14, is given by

$$\psi_\lambda(\eta) = \frac{-1}{2}\log(-\eta) + C''_\lambda, \tag{III.25}$$

where C''_λ is a constant depending on λ. Note that (III.23) and (III.25) have the same form other than a change of sign.

Another fundamental example of λ-exponential family is the *Dirichlet perturbation model*. It was used in [29] to study the Dirichlet optimal transport problem whose solution utilizes the logarithmic duality used in this paper. According to [49], this distribution – also called the *shifted Dirichlet distribution* – was considered by Savage as early as 1966. To the best of our knowledge, it has not been considered in the context of q-exponential family (also see [50]).

Example III.18 (Dirichlet perturbation). Consider the open unit simplex Δ^d defined by (II.22). We define a commutative operation, called perturbation, on Δ^d by

$$p \oplus q = \left(\frac{p_0 q_0}{\sum_{k=0}^d p_k q_k}, \dots, \frac{p_d q_d}{\sum_{k=0}^d p_k q_k}\right). \tag{III.26}$$

Under the Aitchison geometry in compositional data analysis [51], [52], the perturbation is a vector addition. As mentioned in Section II-B, the escort transformation plays the role of scalar multiplication.

Fix $\sigma > 0$ to be interpreted as a noise parameter, and let $\lambda = -\sigma < 0$. Let $D = (D_0, D_1, \dots, D_d)$ be a Dirichlet random vector with parameters $(\sigma^{-1}/(1+d), \dots, \sigma^{-1}/(1+d))$. Note that as $\sigma \downarrow 0$ the distribution of D concentrates at the barycenter $\overline{e} = (1/(1+d), \dots, 1/(1+d))$ of Δ^d. For $p \in \Delta^d$ fixed, consider the random vector $Q = p \oplus D$. It is helpful to think of this probabilistic model as a multiplicative analogue of the additive Gaussian model $Y = \theta + Z$ where $Z \sim N(0, \sigma^2 I)$. In [29] we used this model to construct a probabilistic solution to the Dirichlet transport.

The distribution of Q is given in [29, Lemma 8]:

Lemma III.19. *The density of Q with respect to the Lebesgue measure $dq_1 dq_2 \cdots dq_d$ on the domain $\{q_1, \dots, q_d > 0, q_1 + \cdots + q_d < 1\}$ is given by*

$$f_\lambda(q \mid p) = \frac{C_{d,\lambda}}{\prod_{i=0}^d q_i} \prod_{i=0}^d \left(\frac{q_i}{p_i}\right)^{\frac{-1}{\lambda(1+d)}} \left(\sum_{i=0}^d \frac{q_i}{p_i}\right)^{1/\lambda}, \tag{III.27}$$

where $C_{d,\lambda} > 0$ is a constant depending only on d and λ.

Consider on Δ^d the Dirichlet cost function given by

$$c(p,q) = \log\left(\frac{1}{1+d}\sum_{i=0}^d \frac{q_i}{p_i}\right) - \sum_{i=0}^d \frac{1}{1+d}\log\frac{q_i}{p_i}. \tag{III.28}$$

Following [29], we observe that

$$f_\lambda(q \mid p) \propto \frac{1}{\prod_{i=0}^d q_i} e^{-\sigma c(p,q)}. \tag{III.29}$$

This justifies the name of the optimal transport problem. We note in passing that $\frac{1}{\prod_{i=0}^d q_i} dq_1 \cdots dq_d$ is a Haar measure on the Aitchison simplex, and in this context it is commonly called the *Aitchison measure* (see for example [53]).

Now we observe that the distribution of $Q = p \oplus D$ (parameterized by p) on Δ^d can be expressed as a λ-exponential family after a suitable reparameterization.

Proposition III.20 (Dirichlet perturbation as an λ-exponential family). *Let $\lambda = -\sigma < 0$. Consider the state space Δ^d be the state space. For $i = 1, \dots, d$, let $F_i(q) = q_i/q_0$ and $\vartheta_i = p_0/(\lambda p_i)$. Note that ϑ takes values in $\Omega = (-\infty, 0)^d$.*

Let the reference measure be

$$\mathrm{d}\nu(q) = C'_{n,\lambda} \frac{\prod_{i=1}^d F_i(q)^{-1/\lambda(1+d)}}{\prod_{i=0}^d q_i} \mathrm{d}q_1 \cdots \mathrm{d}q_d,$$

where $C'_{d,\lambda} > 0$ is an appropriate constant, and let $\rho(q;\theta)$ be the density of Q with respect to ν. Then

$$\rho(q;\vartheta) = (1 + \lambda\vartheta \cdot F(q))^{1/\lambda}\, e^{\sum_{i=1}^d \frac{-1}{\lambda(1+d)}\log(-\vartheta_i)}. \tag{III.30}$$

Thus $\rho(\cdot;\vartheta)$ is a λ-exponential family with

$$\varphi_\lambda(\vartheta) = \sum_{i=1}^d \frac{1}{\lambda(1+d)}\log(-\vartheta_i), \quad \vartheta \in (-\infty, 0)^d. \tag{III.31}$$

We omit the proof as it is mostly direct computation. Statistical applications of the Dirichlet perturbation model and its logarithmic divergence will be investigated in a future paper. See Section VI-C for more discussion.

Our final example in this subsection is the α-*family* studied by Amari and Nagaoka [44]. In Section IV-A we will specialize it to study mixture-type families.

Example III.21 (α-family). For $\alpha \in \mathbb{R} \setminus \{0\}$ fixed, define the α-embedding function

$$L_\alpha(t) = \frac{2}{1-\alpha} t^{\frac{1-\alpha}{2}}, \quad t > 0. \tag{III.32}$$

A family $\mathcal{I}$ of positive functions on a state space $\mathcal{X}$ is said to be α-*affine* if the family $\{L_\alpha \circ f : f \in \mathcal{I}\}$ is convex. A

parameterized density $p(\cdot;\xi)$, $\xi \in \Xi \subset \mathbb{R}^d$, is an α-*family* if its denormalization

$$\{\tau p(\cdot;\xi) : \tau > 0, \xi \in \Xi\}$$

is α-affine (i.e., autoparallel with respect to the α-connection). By [44, (2.76)], an α-family has the representation

$$p(x;\xi) = \left(\sum_{i=0}^{d} \theta_i(\xi) F_i(x)\right)^{2/(1-\alpha)}. \tag{III.33}$$

See [20] for a more general ρ-affine family and its dual affine geometry.

Let us observe that the α-family can be reparameterized as a λ-exponential family, and hence a q-exponential family. To see this, we assume, without loss of generality, that $F_0 > 0$ and $\theta_0(\xi) > 0$. Then, we have

$$\begin{aligned} p(x;\xi) &= \left(\theta_0(\xi)F_0(x) + \sum_{i=1}^{d} \theta_i(\xi)F_i(x)\right)^{2/(1-\alpha)} \\ &= (\theta_0(\xi)F_0(x))^{1/\lambda} \left(1 + \lambda \sum_{i=1}^{d} \frac{\theta_i(\xi)}{\lambda\theta_0(\xi)} \frac{F_i(x)}{F_0(x)}\right)^{1/\lambda} \\ &= c(\xi)(1 + \lambda\vartheta(\xi)\cdot \widetilde{F}(x))^{1/\lambda}, \end{aligned}$$

where $\lambda = \frac{1-\alpha}{2}$, $\vartheta_i(\xi) = \theta_i(\xi)/\lambda\theta_0(\xi)$, $\widetilde{F}_i = F_i/F_0$, and $c(\xi)$ is a normalizing constant. Assuming the mapping $\xi \mapsto \vartheta$ is invertible, we see that the α-family can be expressed as a λ-exponential family, where $c(\xi) = e^{-\varphi_\lambda(\vartheta(\xi))}$.

IV. Mixture-type families under λ-duality

The exponential family (I.1) is not the only family which is naturally associated to the KL-divergence via the Legendre duality. The mixture family plays, in some sense, a dual role [1, Section 2.3]. For motivations, we recall that the probability simplex Δ^d is at the same time an exponential family and a mixture family, and in this case the expectation parameter coincides with the mixture parameter. Nevertheless, the exponential and mixture families are in general distinct objects. In this section we consider mixture-type families under the framework of λ-duality. We first specialize the α-family (Example III.21) to the mixture case and show that its interpolations are compatible with the geometry of the corresponding λ-exponential family. Next, we introduce a new λ-*mixture family* which can be regarded as the dual of the λ-exponential family. We also discuss some relationships between the λ-exponential and λ-mixture families.

A. α-mixture family

Let $p_0(x), p_1(x), \ldots, p_d(x)$ be given (affinely independent) probability densities, with respect to a dominating measure ν, on a given state space. A *mixture-type family* may be defined abstractly as a parameterized density $\{p(\cdot;w)\}_w$, where the mixture parameter $w = (w_0, \ldots, w_d)$ satisfies $w_i \geq 0$, $\sum_i w_i = 1$, such that when $w = e_i = (0, \ldots, 1, \ldots, 0)$ is a vertex of the simplex then $p(\cdot;w) = p_i$. Geometrically, it is an embedding of the closed simplex $\overline{\Delta^d}$ into the space of all densities, such that the vertices are fixed.

We first consider the α-family which can be used to define a mixture-type family. To distinguish it from other families, we call it the α-*mixture family*. Let $\alpha \neq 1$ be fixed and recall the α-embedding function L_α defined by (III.32).

Definition IV.1 (α-mixture family [32])**.** *The α-mixture family with respect to the densities $p_0, \ldots, p_d$ is defined by the α-mean, namely*

$$\begin{aligned} p_\alpha(x;w) &= c(w) L_\alpha^{-1}\left(\sum_{i=0}^{d} w_i L_\alpha(p_i(x))\right) \\ &= c(w)\left(\sum_{i=0}^{d} w_i p_i(x)^{(1-\alpha)/2}\right)^{2/(1-\alpha)}, \end{aligned} \tag{IV.1}$$

where $c(w)$ is a normalizing constant. Here we assume that the integral which defines c converges.

Thus an α-mixture family is an α-family (III.33) where each F_i takes the form of a density function p_i and $\theta_0, \ldots, \theta_d$ are the mixture weights. As seen in Example III.21, the α-mixture family can be reparameterized as a λ-exponential family, and hence a q-exponential family, where $\lambda = \frac{1-\alpha}{2}$ and $q = 1 - \lambda$. In this case there is a more natural way to reparameterize the density. Here we assume that p_0 is strictly positive. Let $\vartheta = (\vartheta_1, \ldots, \vartheta_d) = (w_1, \ldots, w_d)$ which takes values in the open convex set $\Omega = \{\vartheta : \vartheta_i > 0, \vartheta_1 + \cdots + \vartheta_d < 1\}$. We have

$$\begin{aligned} &p_\alpha(x;w) \\ &= c(w)\left(\sum_{i=0}^{d} w_i p_i(x)^{(1-\alpha)/2}\right)^{2/(1-\alpha)} \\ &= c(w)\left(w_0(p_0(x))^\lambda + \sum_{i=1}^{d} w_i (p_i(x))^\lambda\right)^{1/\lambda} \\ &= c(w)p_0(x)\left(1 + \lambda\sum_{i=1}^{d} \vartheta_i \frac{1}{\lambda}\left[\left(\frac{p_i(x)}{p_0(x)}\right)^\lambda - 1\right]\right)^{1/\lambda} \\ &= c(w)p_0(x)(1 + \lambda\vartheta\cdot F(x))^{1/\lambda}, \end{aligned} \tag{IV.2}$$

where $F_i(x) = \frac{1}{\lambda}\left[\left(\frac{p_i(x)}{p_0(x)}\right)^\lambda - 1\right]$ is the Box-Cox transformation of the likelihood ratio. So again we get a λ-exponential family. As a λ-exponential family, the α-mixture family automatically enjoys the properties established in Sections III and V as long as $\lambda < 1$ and the required regularity conditions hold.

Amari [32] showed that the α-mixture family can be interpreted as the weighted barycenter with respect to the α-*divergence*, which we now recall. For $\alpha \neq \pm 1$, the α-divergence is defined by

$$\mathbf{D}_\alpha[p_1 : p_2] = \frac{4}{1-\alpha^2}\left(1 - \int p_1^{\frac{1-\alpha}{2}} p_2^{\frac{1+\alpha}{2}} d\nu\right). \tag{IV.3}$$

When $\alpha \to 1$ (respectively -1) the α-divergence converges to the KL-divergence $\mathbf{H}(p_1||p_2)$ (respectively $\mathbf{H}(p_2||p_1)$). Note that the α-divergence can be expressed as a monotonic transformation of the Rényi divergence. Then, by [32, Theorem 2],

the α-mixture $p_\alpha(x;w)$ is the *right barycenter* of the densities $p_0,\dots,p_d$ with respect to the α-divergence. Namely, we have

$$p_\alpha(\cdot;w)=\arg\min_p\left\{\sum_{i=0}^d w_i\mathbf{D}_\alpha[p_i:p]\right\}. \tag{IV.4}$$

The following result shows that if each density p_i is taken from a base λ-exponential family $\mathcal{M}$, then the resulting α-family (where $\alpha=1-2\lambda$) forms a submanifold (with boundary) of $\mathcal{M}$. We also show that linear interpolations (with respect to w) are compatible with the primal coordinates ϑ of $\mathcal{M}$. This extends and clarifies the analysis of α-mixtures between two distributions (called *q-paths* in [16]) which adopts subtractive normalization and considers only mixtures of two distributions.

Proposition IV.2. *Consider a λ-exponential family $\mathcal{M}$ given by*

$$p(x;\vartheta)=(1+\lambda\vartheta\cdot F(x))^{1/\lambda}e^{-\varphi_\lambda(\vartheta)},$$

where $1+\lambda\vartheta\cdot F(x)>0$ for all ϑ and x. Let $p_i=p(\cdot;\vartheta^{(i)})$, for $i=0,1,\dots,d$, be $d+1$ members of the family $\mathcal{M}$, each specified by a parameter $\vartheta^{(i)}$. Let $\alpha=1-2\lambda$. Then the corresponding α-mixture family is a subset of $\mathcal{M}$. Specifically, we have $p_\alpha(x;w)=p(x;\vartheta)$, where

$$\vartheta=\vartheta(w)=\sum_{i=0}^d\frac{w_ie^{-\lambda\varphi_\lambda(\vartheta^{(i)})}}{\sum_j w_je^{-\lambda\varphi_\lambda(\vartheta^{(j)})}}\theta^{(i)}. \tag{IV.5}$$

In words, given the $d+1$ "inducing" elements of $\mathcal{M}$, each α-mixture exists and is an element of $\mathcal{M}$.

Proof. We have

$$\begin{aligned}
&\sum_{i=0}^d w_ip_i^{(1-\alpha)/2}=\sum_{i=0}^d w_ip_i^\lambda\\
&=\sum_{i=0}^d w_i(1+\lambda\vartheta^{(i)}\cdot F)e^{-\lambda\varphi_\lambda(\vartheta^{(i)})}\\
&=c'\left(1+\lambda\sum_{i=0}^d\frac{w_ie^{-\lambda\varphi_\lambda(\vartheta^{(i)})}}{\sum_j w_je^{-\lambda\varphi_\lambda(\vartheta^{(j)})}}\theta^{(i)}\cdot F\right)\\
&=c'(1+\lambda\vartheta\cdot F),
\end{aligned}$$

where $c'=c'(w)$ is a constant and ϑ is given by (IV.5). Note that the $(1/\lambda)$-th power of the last expression is integrable since the parameter set of $\mathcal{M}$ is convex by Lemma III.2. It follows that $p_\alpha(x;w)=p(x;\vartheta)$. □

Remark IV.3. This property is related to the fact that the λ-exponential family is α-affine in the sense of [44]. Here, we identified the parameter system ϑ such that the mapping $w\mapsto\vartheta$ is *projective*, i.e., taking straight lines to straight lines. The last property is shown in the following corollary.

Corollary IV.4. *Under the setting of Proposition IV.2, consider a straight line $w(t)=(1-t)w^{(0)}+tw^{(1)}$ under the mixture parameters of the α-family. Then the corresponding probability density functions $\{p_\alpha(\cdot;w(t)):0\le t\le1\}$, which are all members of the λ-exponential family $\mathcal{M}$, trace out a straight line under the primal coordinate system ϑ.*

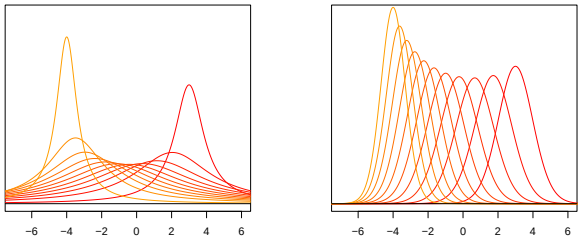


Fig. 6: The α-family formed by two t-distributions (with different location and scale parameters) on the real line (see Example IV.5). Left: df (degree of freedom) $=3$, so that $\alpha=0.5$. Right: df $=30$, so that $\alpha\approx1.13$. In both bases we use the location and scale parameters $(\mu_0,\sigma_0)=(-4,0.7)$, $(\mu_1,\sigma_1)=(3,1)$ for p_0 and p_1. The interpolation corresponds to a straight line under the ϑ-coordinate system of the associated λ-exponential family (see Corollary IV.4), but is a nonlinear curve under the θ-coordinate system of the corresponding q-exponential family. As df $\to\infty$ (so $\alpha\to1$) the α-family converges to the *exponential mixture* which is analogous to (but different from) McCann's displacement interpolation [54] between two Gaussian distributions.

Proof. Let $\vartheta(t)$, given by (IV.5), be the ϑ-coordinate of $p_\alpha(\cdot;w(t))$ as an element of $\mathcal{M}$. We have

$$\vartheta(t)=\sum_{i=0}^d\frac{((1-t)w_i^{(0)}+tw_i^{(1)})a_i}{\sum_j((1-t)w_j^{(0)}+tw_j^{(1)})a_j}\theta^{(i)},$$

where $a_i=e^{-\lambda\varphi_\lambda(\theta^{(i)})}$.

Consider the weight of the i-term. Write

$$\begin{aligned}
&\frac{((1-t)w_i^{(0)}+tw_i^{(1)})a_i}{\sum_j((1-t)w_j^{(0)}+tw_j^{(1)})a_j}\\
&=\frac{(1-t)\sum_j w_j^{(0)}a_j}{\sum_j[\cdots]a_j}\frac{w_i^{(0)}a_i}{\sum_j w_j^{(0)}a_j}+\frac{t\sum_j w_j^{(1)}a_j}{\sum_j[\cdots]a_j}\frac{w_i^{(1)}a_i}{\sum_j w_j^{(1)}a_j}\\
&=:(1-s(t))\frac{w_i^{(0)}a_i}{\sum_j w_j^{(0)}a_j}+s(t)\frac{w_i^{(1)}a_i}{\sum_j w_j^{(1)}a_j},
\end{aligned} \tag{IV.6}$$

where $[\cdots]=(1-t)w_j^{(0)}+tw_j^{(1)}$. Note that $s(t)$ does not depend on i and increases from 0 to 1. Thus we have

$$\vartheta(t)=(1-s(t))\vartheta(0)+s(t)\vartheta(1),$$

which is a time change of a constant-speed straight line under the ϑ coordinates. □

According to the geometric language explained in Section V (also see Remark III.13), the path $\{p_\alpha(\cdot;w(t)):0\le t\le1\}$ is a *primal pre-geodesic* under the information geometry induced by the λ-logarithmic divergence $\mathbf{L}_{\lambda,\varphi_\lambda}$ on $\mathcal{M}$. Indeed, equations analogous to (IV.6) also appear naturally in [27], [26], and the time change is related to the fact that the underlying geometry is *projectively flat*.

Example IV.5 (Multivariate t-distribution). Let the state space be $\mathbb{R}^n$. The multivariate t-distribution with k degrees of freedom, and with parameters (μ, Σ) (where $\mu \in \mathbb{R}^n$ and Σ is a $n \times n$ positive definite matrix) has density (with respect to the Lebesgue measure) given by

$$p(x; k, \mu, \Sigma) = c \left[1 + \frac{1}{k}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]^{-(k+n)/2}, \tag{IV.7}$$

where $c = c(\mu, \Sigma) > 0$ is a normalization constant. It can be shown that for k fixed this can be reparameterized as a λ-exponential family where $\lambda = \frac{-2}{k+n} < 0$ (this generalizes Examples III.16 and III.17). Thus we may write

$$p(x; k, \mu, \Sigma) = p(x; \vartheta) = (1 + \lambda \vartheta \cdot F(x))^{\frac{1}{\lambda}} e^{-\varphi_\lambda(\vartheta)}, \tag{IV.8}$$

where $F(x) = ((x_i)_i, (x_i x_j)_{i \le j})$. Note that as $k \to \infty$ it converges to the Gaussian distribution $N(\mu, \Sigma)$.

Let $p_i = p(\cdot; k, \mu^{(i)}, \Sigma^{(i)})$ for $i = 0, 1, \dots, d$ and consider the α-mixture family where $\alpha = 1 - 2\lambda$. By Proposition IV.2, each α-mixture belongs to the same family, and hence is a t-distribution. This fact may appear strange at first sight as the ordinary mixture (of unimodal distributions) is typically multimodal. However, as mentioned above, the α-mixture (III.33) may also be interpreted as the barycenter with respect to the α-divergence [32]. Here, the α-mixture is analogous to the Wasserstein barycenter in optimal transport theory [55], where it can be shown that the Wasserstein barycenter of Gaussian distributions is also Gaussian. In Figure 6 (inspired by [16]) we give a graphical illustration of the α-mixture of two t-distributions on the real line. We remark that here α is chosen as a function of the degrees of freedom k.

B. λ-mixture family

In (IV.2), where we express the α-mixture family as a λ-exponential family, the mixture parameter w plays the role of the primal variable ϑ. However, as mentioned in Section II-A, the negative Shannon entropy of a (conventional) mixture family is convex in the mixture parameter which plays the role of the dual variable, and its Bregman divergence is the KL-divergence. In this subsection we introduce a new λ-mixture family which preserves this analogy. To motivate the definition, recall Example II.7 where the negative Rényi entropy of the escort distribution is c_λ-convex on the simplex. Recall also that $\mathcal{E}_q[\cdot]$ is the escort transformation with exponent q (see (II.10)):

Definition IV.6 (λ-mixture family)**.** *Let $\lambda \neq 1$ be given and let $q = 1 - \lambda$. We define the λ-mixture family with respect to the densities $p_0, \dots, p_d$ by*

$$p_\eta(x) = p(x; \eta) = \mathcal{E}_{1/q}\left[\sum_{i=0}^d \eta_i \mathcal{E}_q[p_i]\right], \tag{IV.9}$$

where $\eta_i \ge 0$, $\sum_{i=0}^d \eta_i = 1$ are the λ-mixture parameters, provided that the integrals involved all converge. Explicitly, we have

$$p_\eta(x) = \frac{1}{\int (\sum_{j=0}^d \eta_j \widetilde{p}_j)^{1/q} \mathrm{d}\nu} \left(\sum_{i=0}^d \eta_i \widetilde{p}_i\right)^{1/q}, \tag{IV.10}$$

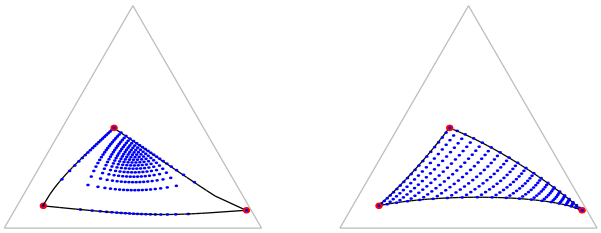


Fig. 7: Graphical illustration of the λ-mixture family in the context of Example IV.7. Left: $\lambda = -2$. Right: $\lambda = 0.7$. The red points show the densities p_i, $i = 0, 1, 2$. Each $p(\cdot; \eta)$, where η ranges over a uniform grid on the closed simplex $\overline{\Delta^2}$, is shown as a blue dot in Δ^2 (where the superscript denotes the dimension of the simplex Δ.

where $\widetilde{p}_i = \mathcal{E}_q[p_i] = p_i^q / \int p_i^q \mathrm{d}\nu$.

From (IV.9), if $\widetilde{p}_\eta = \mathcal{E}_q[p_\eta]$ is the escort transformation of p_η, then

$$\widetilde{p}_\eta = \sum_{i=0}^d \eta_i \widetilde{p}_i$$

is an ordinary mixture family with respect to the escort distributions $\widetilde{p}_0, \dots, \widetilde{p}_d$. Thus the λ-mixture family is nothing but the escort transformation (with exponent $1/q$) of an ordinary mixture family. See Section IV-C for the relationship between the λ-mixture and α-mixture families.

Before proceeding we illustrate the λ-mixture family with a concrete example.

Example IV.7. Consider the finite state space $\mathcal{X} = \{0, 1, 2\}$ and let ν be the counting measure. Consider for the sake of illustration the three densities

$$p_0 = (0.8, 0.1, 0.1),$$
$$p_1 = (0.02, 0.9, 0.08),$$
$$p_2 = (0.35, 0.2, 0.45).$$

In Figure 7 we visualize the corresponding λ-mixture family where $\lambda = -2$ (left) and $\lambda = 0.7$ (right). Each density $p(\cdot; \eta)$, where η ranges over a uniform grid on the unit simplex, is shown as a blue dot on the unit simplex Δ^2.

Observe that the λ-mixture family leads to nonlinear interpolating curves on the simplex. For example, when $\lambda < 0$ the blue dots are denser when η is near the center of the simplex, and are far apart when η is close to the boundary. When $\lambda \to 0$ we cover the ordinary mixture family (I.2).

The following result shows that the λ-mixture family is naturally compatible with the λ-duality, and preserves the analogy with the mixture family (see (II.9)).

Theorem IV.8. *Consider a λ-mixture family where $\lambda < 1$ (or $q = 1 - \lambda > 0$) (this assumes implicitly that all integrals involved in* (IV.9) *converge for all η), such that the integral $\int (\sum_{i=0}^d \eta_i \widetilde{p}_i)^{1/q} \mathrm{d}\nu$ can be differentiated with respect to η under the integral sign when $\eta \in \Delta^d$. Consider the negative Rényi entropy*

$$\psi(\eta) = -\mathbf{H}_q^{\text{Rényi}}(p_\eta), \quad q = 1 - \lambda. \tag{IV.11}$$

Then $\frac{1}{\lambda}(e^{\lambda\psi}-1)$ *is convex. Moreover, the* λ*-logarithmic divergence of* ψ *is the Rényi divergence with order* q*:*

$$\mathbf{L}_{\lambda,\psi}[\eta:\eta'] = \mathbf{H}_q^{\text{Rényi}}(p_\eta||p_{\eta'}), \quad \eta' \in \Delta^d. \tag{IV.12}$$

Proof. Let $Z(\eta) = \int(\sum_j \eta_j \widetilde{p}_j)^{1/q}\mathrm{d}\nu$ be the partition function in (IV.10), so that

$$p(x;\eta) = \frac{1}{Z(\eta)}\left(\sum_{j=0}^d \eta_j \widetilde{p}_j\right)^{1/q}.$$

By (IV.9) and the definition of Rényi entropy (I.4), we have

$$\psi(\eta) = \frac{q}{1-q}\log\int\left(\sum_{i=0}^d \eta_i\widetilde{p}_i\right)^{1/q}\mathrm{d}\nu = \frac{q}{1-q}\log Z(\eta). \tag{IV.13}$$

Let η, η' be given. By a direct computation, we have

$$\begin{aligned}1+\lambda\nabla\psi(\eta')\cdot(\eta-\eta') &= \frac{\int(\sum_i \eta_i'\widetilde{p}_i)^{1/q-1}(\sum_i \eta_i\widetilde{p}_i)\mathrm{d}\nu}{\int(\sum_i \eta_i'\widetilde{p}_i)^{1/q}\mathrm{d}\nu}\\ &= \frac{Z(\eta)^q}{Z(\eta')^q}\int p_\eta(x)^q p_{\eta'}(x)^{1-q}\mathrm{d}\nu.\end{aligned}$$

It follows that

$$\begin{aligned}&\psi(\eta)-\psi(\eta')-\frac{1}{\lambda}\log(1+\lambda\nabla\psi(\eta')\cdot(\eta-\eta'))\\ &= \frac{1}{q-1}\log\int p_\eta(x)^q p_{\eta'}(x)^{1-q}\mathrm{d}\nu(x)\\ &= \mathbf{H}_q^{\text{Rényi}}(p_\eta||p_{\eta'}) \geq 0.\end{aligned}$$

This argument shows that $\frac{1}{\lambda}(e^{\lambda\psi}-1)$ is convex whenever $q = 1-\lambda > 0$, and the λ-logarithmic divergence of ψ is the Rényi divergence of order q. □

Remark IV.9. It is easy to verify by examples that Theorem IV.8 does not hold for the α-mixture family if we define instead $\psi(w) = -\mathbf{H}_q^{\text{Rényi}}(p_\alpha(\cdot;w))$. For an α-mixture family (which can be expressed as a λ-exponential family), it is the c_λ-conjugate ψ_λ of the divisive λ-potential φ_λ which can be expressed as a Rényi entropy (see Theorem III.14).

Now we place Example II.7 in the proper context.

Example IV.10 (Finite simplex). Consider the unit simplex Δ^d as in (II.22). Each $u \in \Delta^d$ can be regarded as the density p of a probability measure on a finite set $\mathcal{X} = \{0,1,\dots,d\}$ with respect to the counting measure. So we may view $\Delta^d \cong \mathcal{P}_+(\mathcal{X})$ as the set of strictly positive probability measures on $\mathcal{X}$. We write $u_x = p(x)$, $x \in \mathcal{X}$.

For $\lambda \neq 0$, Δ^d can be expressed as a λ-exponential family. To see this, let $F_i(x) = \delta_i(x)$ be the indicator of $i \in \mathcal{X}$, $i = 1,\dots,d$. Then we may write

$$u_x = p_\vartheta(x) = \left(1+\lambda\sum_{i=1}^d \vartheta_i\delta_i(x)\right)^{1/\lambda} e^{-\varphi_\lambda(\vartheta)}, \tag{IV.14}$$

where

$$\vartheta_i = \frac{1}{\lambda}\left[\left(\frac{u_i}{u_0}\right)^\lambda - 1\right] \tag{IV.15}$$

is the Box-Cox transformation of u_i/u_0, and the parameter set is $\Omega = \{\vartheta \in \mathbb{R}^d : 1+\lambda\vartheta^i > 0\ \forall i\}$. Note that when $\lambda \to 0$, ϑ reduces to the usual exponential coordinates $\theta_i = \log(u_i/u_0)$. The potential function, given by

$$\varphi_\lambda(\vartheta) = -\log u_0 = \log\left(1+\sum_{i=1}^d(1+\lambda\vartheta_i)^{1/\lambda}\right),$$

is c_λ-convex for $\lambda < 1$. By Theorem III.12, the dual variable $\eta = \nabla^{c_\lambda}\varphi_\lambda(\vartheta)$ gives the escort distribution with exponent q:

$$\eta_i = \frac{(p_\vartheta(i))^q}{\sum_{j=0}^d (p_\vartheta(j))^q}, \quad i = 1,\dots,d. \tag{IV.16}$$

We define $\eta_0 = 1-\sum_{i=1}^d \eta_i$. Dually, for $i = 0,1,\dots,d$, let $p_i(x) = \delta_i(x)$ be the density of the point mass at point i and note that $\widetilde{p}_i = \mathcal{E}_q[p_i] = p_i$. Consider the corresponding λ-mixture family p_η (not to be confused with p_ϑ). By (IV.9), we have

$$p_\eta(i) = \frac{\eta_i^{1/q}}{\sum_{j=0}^d \eta_j^{1/q}}. \tag{IV.17}$$

Note that if η is given by (IV.16), then we have $p_\vartheta = p_\eta$. Thus the simplex is at the same time a λ-exponential family and a λ-mixture family, and the mixture parameter of the λ-mixture family is the dual parameter of the λ-exponential family. That the unit simplex can be expressed as a generalized exponential family (including the q-exponential family) was observed in earlier works such as [10]. Here, we use the λ-representation and connect it with the λ-mixture family.

Formula (IV.15), which gives the "primal" variable (if we regard η as the dual one), can be extended to a general λ-mixture family. Since the set $\{\eta \in (0,1)^{1+d} : \eta_0 + \cdots + \eta_d = 1\}$ is not open, we consider instead the open domain

$$\Omega' = \{\bar\eta = (\bar\eta_1,\dots,\bar\eta_d) \in (0,1)^d : \bar\eta_1 + \cdots + \bar\eta_d < 1\},$$

where the notation $\bar\eta$ signifies that the 0-th coordinate is dropped. Consider the negative Rényi entropy ψ, given by (IV.11), as a function of $\bar\eta$. Note that

$$\frac{\partial\psi}{\partial\bar\eta_i} = \frac{\partial\psi}{\partial\eta_i} - \frac{\partial\psi}{\partial\eta_0}, \quad i = 1,\dots,d. \tag{IV.18}$$

Note that $\frac{1}{\lambda}(e^{\lambda\psi}-1)$ is convex in $\bar\eta$ (which is a linear transformation of η) and so we may consider the λ-duality. One can verify that in the context of Example IV.10 the formula (IV.19) below reduces to (IV.15).

Proposition IV.11. *With the above notations, consider the primal variable* $\bar\vartheta = \nabla^{c_\lambda}_{\bar\eta}\psi(\bar\eta)$*. We have*

$$\bar\vartheta^i = \frac{1}{\lambda}\left[\frac{\int p(x;\eta)^\lambda\widetilde{p}_i(x)\mathrm{d}\nu}{\int p(x;\eta)^\lambda\widetilde{p}_0(x)\mathrm{d}\nu} - 1\right], \quad i = 1,\dots,d. \tag{IV.19}$$

Proof. Given $\bar\eta \in \Omega$, consider η where $\eta_i = \bar\eta_i$ for $1 \leq i \leq d$ and $\eta_0 = 1-\sum_i \bar\eta_i$. Using (IV.18), we have

$$\frac{\partial\psi}{\partial\bar\eta_i} = \frac{1}{\lambda Z(\eta)}\int\left(\sum_{j=0}^d \eta_j\widetilde{p}_j\right)^{1/q-1}(p_i - p_0)\mathrm{d}\nu.$$

Now (IV.19) follows from the definition of the λ-deformed Legendre transformation. □

C. Further relations

We have seen in (IV.2) that a α-mixture family can be reparameterized as a λ-exponential family with $\lambda = \frac{1-\alpha}{2}$. In this subsection we consider other relations between the α-mixture family, λ-mixture family and λ-exponential family.

The α-mixture family and the λ-mixture family differ in the way the densities are normalized before taking the average: in (III.33) one considers $\sum_i w_i p_i^{(1-\alpha)/2}$, while in (IV.9) we have instead $\sum_i \eta_i \mathcal{E}_q[p_i]$. The α-mixture family is more general since the first sum always exists while the escort distributions $\mathcal{E}_q[p_i]$ may not be well-defined (since the integral $\int p_i^q \mathrm{d}\nu$ may diverge). Nevertheless, when the state space is a finite or bounded and $p_i > 0$ for all i, the λ-mixture family is always well-defined. Again we note that the main motivation for the λ-mixture family is Theorem IV.8 which is analogous to the Legendre duality of the ordinary mixture family.

Now we show that a λ-mixture family, when exists, is a reparameterization of the α-mixture family. So the two families give the same collection of distributions.

Proposition IV.12. *A λ-mixture family is a reparameterization of the α-mixture family (with the same base densities $p_0, \dots, p_d$) where $\lambda = \frac{\alpha+1}{2}$.*

Proof. Consider a λ-mixture family $p(\cdot;\eta)$ of $p_0, \dots, p_d$. Let $q = 1 - \lambda$. Write $\mathcal{E}_q[p_i] = p_i^q / Z_i$, where $Z_i = \int p_i^q d\nu$ is a constant which is finite by assumption. In what follows c, c' are positive constants. Then we have

$$p(\cdot;\eta) = c \left(\sum_{i=0}^d \eta_i \frac{p_i^q}{Z_i} \right)^{1/q} = c' \left(\sum_{i=0}^d \frac{\eta_i/Z_i}{\sum_j \eta_j/Z_j} p_i^q \right)^{1/q}$$
$$= c' \left(\sum_{i=0}^d w_i p_i^{(1-\alpha)/2} \right)^{2/(1-\alpha)} = p_\alpha(\cdot; w),$$

where the two sets of mixture parameters are related by $w_i = \frac{\eta_i/Z_i}{\sum_j \eta_j/Z_j}$ and $\eta_i = \frac{w_i Z_i}{\sum_j w_j Z_j}$. Here the parameter transformations are also projective. □

Because of this result, the α-mixture family and λ-mixture family (when $\lambda = \frac{\alpha+1}{2}$) are qualitatively similar. It is interesting that the reparameterization $w \mapsto \eta$ enables us to link the family with the λ-duality. In view of this result we have the following

Corollary IV.13. *Suppose $p_0 > 0$. Then a λ-mixture family is, after a reparameterization, a λ'-exponential family, and hence a q'-exponential family, where $\lambda' = 1-\lambda$ and $q' = 1-\lambda' = \lambda$.*

Thus both the α-mixture family and the λ-mixture family can be regarded as special cases of the λ-exponential family (for different values of λ). This is a special feature of our framework as it is well-known that in general a mixture family cannot be expressed as an exponential family. We also remark that while the λ-exponential family and the q-exponential family give algebraically the most general expressions, the λ-mixture family satisfies the duality in Theorem IV.8 (which adopts the Rényi entropy) when $\lambda < 1$ (or $q > 0$), while the results of λ'-exponential family (which adopts another potential function) requires $\lambda' < 1$.

V. Information geometry of λ-logarithmic divergence

Both the λ-exponential and λ-mixture families can be regarded as manifolds of probability distributions. By Theorem III.14 and Theorem IV.8, each family is associated to a λ-logarithmic divergence (Definition II.5) which is shown to be a Rényi divergence. Again, this parallels the classical exponential and mixture families which are connected to the Kullback-Leibler divergence, see (II.5) and (II.9). In information geometry [1], a divergence induces a geometry on the underlying manifold of probability families that carries a dualistic structure. Formally, it consists of a Riemannian metric as well as a pair of mutually dual and torsion-free affine connections; in differential geometry, this structure is also known as a "statistical manifold". Using the results of [26], in this section we summarize the information geometry induced by a λ-logarithmic divergence.

Let Ω be an open convex set of $\mathbb{R}^d$ and let $\mathbf{L}_{\lambda,\varphi}[\cdot : \cdot]$ be a λ-logarithmic divergence on Ω induced by a regular c_λ-convex function φ. For concreteness, the reader may keep in mind the situation of Theorem III.14 where Ω is the natural parameter set of a λ-exponential family and $\varphi = \varphi_\lambda$ is the divisive potential function.

A. Riemannian metric

Consider two nearly points ϑ and $\vartheta + \Delta\vartheta$ in Ω and consider the divergence $\mathbf{L}_{\lambda,\varphi}[\vartheta + \Delta\vartheta : \vartheta]$. Applying a Taylor approximation to (II.21), we have the local quadratic approximation

$$\mathbf{L}_{\lambda,\varphi}[\vartheta + \Delta\vartheta : \vartheta] = \frac{1}{2} (\Delta\vartheta)^\top g(\vartheta) (\Delta\vartheta) + O(|\Delta\vartheta|^3), \quad \text{(V.1)}$$

where the matrix

$$g(\vartheta) = \nabla^2 \varphi(\vartheta) + \lambda (\nabla\varphi(\vartheta))(\nabla\varphi(\vartheta))^\top \quad \text{(V.2)}$$

is positive definite as can be seen from (II.18). We regard $g(\vartheta)$ as a metric tensor on the manifold Ω:

$$\mathrm{d}s^2 = \sum_{i,j} g_{ij}(\vartheta) \mathrm{d}\vartheta^i \mathrm{d}\vartheta^j.$$

Let us compare (V.2) with the metric induced by a Bregman divergence. If $\phi(\theta)$ is a convex potential which defines a Bregman divergence, the corresponding metric tensor is the Hessian $\nabla^2\phi(\theta)$ [1, Chapter 1]. From (V.2), we see that when φ is itself convex, then the metric induced by the λ-logarithmic divergence is a rank-1 correction of a Hessian metric. Alternatively, from (II.18), we may also cast (V.2) as

$$g(\vartheta) = e^{-\lambda\varphi(\vartheta)} \nabla^2 \Phi(\vartheta),$$

where $\Phi = \frac{1}{\lambda}(e^{\lambda\varphi} - 1)$ is convex. Thus, the metric can also be regarded as a conformal transformation of the Hessian metric induced by Φ; hence, we call it a *conformal Hessian metric*. In fact, as shown in [56], the λ-logarithmic divergence may be regarded as a monotone transformation of a conformal Bregman divergence.

For the λ-exponential and λ-mixture families, where the divergence is a Rényi divergence, the induced Riemannian metric is q times the Fisher information metric (see [46,

Section III-H]). This is different from classical deformation theory (see for example [8]) where the induced metric is a conformal transformation of the Fisher metric. By using the λ-duality we recover exactly the Fisher metric up to a multiplicative constant.

Remark V.1. In [57], it was shown that deformed exponential families (under subtractive normalization) admit in general both a Hessian metric and a conformal Hessian metric, where the Fisher metric is deformed as a (ρ, τ)-metric.

B. Geodesics and generalized Pythagorean theorem

Remarkably, the λ-logarithmic divergence (which contains the Bregman divergence in the limit) satisfies a *generalized Pythagorean theorem*. To state the result it we need the notion of primal and dual geodesics. We regard $\vartheta \in \Omega$ as the primal coordinate system, and $\eta = \nabla^{c_\lambda}\varphi(\vartheta)$ as the dual coordinate system. Let $\gamma(t)$ be a curve in the statistical manifold.

- We say that γ is a *primal pre-geodesic* if its image under the primal coordinate system is a straight line.
- Similarly, γ is a *dual pre-geodesic* if its image is a straight line under the dual coordinate system.

Note that we call the curves pre-geodesics because the actual geodesics (defined by the primal and dual geodesic equations) run in non-constant speed in the respective coordinate systems. We refer the reader to [26] for the associated primal and dual affine connections which define the geodesic equations. Nevertheless, the trajectories of these geodesics are straight lines under the respective coordinate systems. Because of this feature, the geometry is said to be *dually projectively flat*. We are now ready to state the theorem.

Theorem V.2 (Generalized Pythagorean theorem)**.** *Let $\vartheta_P, \vartheta_Q, \vartheta_R$ be three points in the statistical manifold. Then*

$$\mathbf{L}_{\lambda,\varphi}[\vartheta_Q : \vartheta_P] + \mathbf{L}_{\lambda,\varphi}[\vartheta_R : \vartheta_Q] = \mathbf{L}_{\lambda,\varphi}[\vartheta_R : \vartheta_P] \qquad \text{(V.3)}$$

if and only if the primal pre-geodesic from ϑ_Q to ϑ_R and the dual pre-geodesic from ϑ_Q to ϑ_P are g-orthogonal at ϑ_Q. (When $\lambda > 0$, we assume that the divergences involved are all finite.)

Proof. See [26, Theorem 16]. □

Furthermore, by [26, Theorem 15], the geometric structure induced by a λ-logarithmic divergence has *constant* sectional curvature (in the information-geometric sense) equal to λ. (For a converse see [26, Theorem 19].) This justifies the interpretation of the constant λ as the curvature. We refer the reader to [56], [31] for some geometric interpretations of the curvature in terms of the logarithmic divergence between a primal-dual pair of geodesics. Some aspects of projections with respect to the $L^{(\alpha)}$-divergence (equivalent to a λ-logarithmic divergence where $\alpha = -\lambda > 0$) are studied in [47], [48] where a dual foliation is constructed.

VI. Duality between λ-exponential family and λ-logarithmic divergence

Our previous results show that the λ-exponential and λ-mixture families have a rich analytic and geometric structure when considered from the viewpoint of λ-duality and λ-logarithmic divergence. In this section we study some statistical implications of our approach.

A. Motivation

An important property of exponential family which is perhaps less well known is that the log likelihood of an exponential family can be associated to a Bregman divergence. To give a concrete example, consider the normal location family $\{N(\theta, I)\}_{\theta\in\mathbb{R}^d}$. Let the reference measure ν be the standard normal distribution $N(0, I)$. Then the density of $N(\theta, I)$ is given by

$$p(x;\theta) = e^{x\cdot\theta - \frac{1}{2}|\theta|^2},$$

which is an exponential family with $F(x) = x$, $\phi(\theta) = \frac{1}{2}|\theta|^2$ and $\psi(\eta) = \phi^*(\eta) = \frac{1}{2}|\eta|^2$. We have $\eta = \nabla\phi(\theta) = \theta$, so the normal location family (with unit variance) is *self-dual*: the natural parameter θ is the same as the expectation parameter η. Using the identity $\frac{1}{2}|x-\eta|^2 = \frac{1}{2}|x|^2 - \theta\cdot x + \frac{1}{2}|\theta|^2$, we may write

$$\log p(x;\theta) = -\frac{1}{2}|x-\eta|^2 + \frac{1}{2}|x|^2 = -\mathbf{B}_\psi[x:\eta] + \psi(x). \qquad \text{(VI.1)}$$

Thus the log-likelihood can be expressed in terms of the Bregman divergence of ψ, which in this case is $1/2$ times the squared distance. This maybe regarded as a probabilistic basis of the least squares criterion. Next consider a general exponential family. By considering the distribution of $Y = F(X)$, we may consider exponential families on $\mathbb{R}^d$ of the form

$$p(y;\theta) = e^{\theta\cdot y - \phi(\theta)}.$$

Let $\psi = \phi^*$ be the convex conjugate of ϕ. In a seminal paper [25], it was proved that when the family is absolutely continuous with respect to either the counting or Lebesgue measure, and satisfies natural regularity conditions, then one has the representation

$$p(y;\theta) = e^{-\mathbf{B}_\psi[y:\eta]+\psi(y)}. \qquad \text{(VI.2)}$$

Note that in general it is the dual parameter $\eta = \nabla\phi(\theta) = \mathbb{E}_\theta[F(X)]$, not θ, which has a direct interpretation on the "data space" where y lives. This representation implies that maximization of the likelihood is equivalent to minimization of the corresponding Bregman divergence. We also mention that

$$\eta = \arg\min_{\eta'} \mathbb{E}_\theta[\mathbf{B}_\psi[Y:\eta']], \quad Y \sim p(\cdot;\theta). \qquad \text{(VI.3)}$$

So the expectation parameter η is also the right barycenter of the distribution with respect to the Bregman divergence $\mathbf{B}_\psi$.

B. λ-exponential family and λ-logarithmic divergence

In [26, Section 4], the $\mathcal{F}^{(\pm\alpha)}$-families were derived by replacing the Bregman divergence in (VI.2) by a logarithmic divergence. For example, the density (III.29) of the Dirichlet perturbation model can be expressed in the form $p(\cdot) \propto e^{-\mathbf{D}}$, where $\mathbf{D}$, given by (III.28), is a logarithmic divergence (see Example II.6). Here, it is natural to adopt divisive rather

than subtractive normalization. In this subsection, we derive heuristically the analogue of (VI.2) for the λ-exponential family, and leave the rigorous treatment to future research.

Following the approach of [25], consider a λ-exponential family on $\mathbb{R}^d$ where $F(y) = y$. The dominating measure is assumed to be absolutely continuous with respect to either the counting or Lebesgue measure. Under the support condition (so that we may assume $1 + \lambda\vartheta \cdot y > 0$ on a common domain independent of ϑ), we may write the density as

$$\begin{aligned} p(y;\vartheta) &= (1+\lambda\vartheta\cdot y)^{1/\lambda} e^{-\varphi_\lambda(\vartheta)} \\ &= e^{\frac{1}{\lambda}\log(1+\lambda\vartheta\cdot y)-\varphi_\lambda(\vartheta)}. \end{aligned} \tag{VI.4}$$

Here $\varphi_\lambda(\vartheta)$ is the divisive λ-potential which is c_λ-convex. Let ψ_λ be the λ-conjugate of φ_λ which is also c_λ-convex.

Let $\eta = \nabla^{c_\lambda}\varphi_\lambda(\vartheta)$ be the dual parameter corresponding to ϑ. By the c_λ-duality, we have

$$\frac{1}{\lambda}\log(1+\lambda\vartheta\cdot\eta) = \varphi_\lambda(\vartheta) + \psi_\lambda(\eta). \tag{VI.5}$$

Also, we have the identity (see [26, (41)])

$$1 + \lambda\vartheta\cdot\eta = \frac{1}{1-\lambda\nabla\psi_\lambda(\eta)\cdot\eta}. \tag{VI.6}$$

Substituting (VI.5) into (VI.4), we see that the exponent can be written as

$$\frac{1}{\lambda}\log(1+\lambda\vartheta\cdot y) - \frac{1}{\lambda}\log(1+\lambda\vartheta\cdot\eta) + \psi_\lambda(\eta). \tag{VI.7}$$

Write $\vartheta = \nabla^{c_\lambda}\psi_\lambda(\eta) = \frac{\nabla\psi_\lambda(\eta)}{1-\lambda\nabla\psi_\lambda(\eta)\cdot\eta}$. Using this and (VI.6), we may rearrange (VI.7) to get

$$\frac{1}{\lambda}\log(1+\lambda\nabla\psi(\eta)\cdot(y-\eta)) + \psi_\lambda(\eta) = -\mathbf{L}_{\lambda,\psi}[y:\eta] + \psi_\lambda(y),$$

where in the last equality it is implicitly assumed that y belongs to the domain of ψ_λ so that the divergence $\mathbf{L}_{\lambda,\psi}[y:\eta]$ is well-defined (this is the main subtlety; also see the proof of [25, Theorem 4]). Thus, we have shown heuristically that under suitable conditions, the density of the λ-exponential family can be expressed in the form

$$p(y;\vartheta) = e^{-\mathbf{L}_{\lambda,\psi}[y:\eta]+\psi_\lambda(y)}. \tag{VI.8}$$

Remark VI.1. For an explicit λ-exponential family such as the Dirichlet perturbation model (Example III.18), the representation (VI.8) can be verified directly (see (III.29)). Note that (VI.8) does not hold for the Cauchy location-scale family (Example III.16); this is because the distribution of $Y = F(X) = (X, X^2)$ is supported on a parabola and hence is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. The same issue also occurs in the case of exponential family.

Remark VI.2. Using results from classical convex analysis [34] and deeper mathematical properties of exponential families [33], in particular convex functions of Legendre type, Banerjee et al. [25] also showed that there is a one-to-one correspondence between what they call regular Bregman divergences and regular exponential families. We expect analogous results hold for λ-logarithmic divergences and λ-exponential families.

C. Statistical consequences

The representation (VI.8) leads to some interesting consequences. In the following results we consider a λ-exponential family whose density satisfies (VI.8).

Theorem VI.3 (MLE as right barycenter)**.** *Let $\hat{\vartheta}$ be a maximum likelihood estimate with respect to the λ-exponential family $\{p(\cdot;\vartheta)\}_{\vartheta\in\Omega}$ and data points $y_1,\dots,y_n$ under i.i.d. sampling. Then $\hat{\eta} = \nabla^{c_\lambda}\varphi_\lambda(\hat{\vartheta})$ is a right barycenter of the data points with respect to $\mathbf{L}_{\lambda,\psi_\lambda}$:*

$$\hat{\eta} \in \operatorname*{arg\,min}_{\eta\in\Omega'} \sum_{i=1}^n \frac{1}{n}\mathbf{L}_{\lambda,\psi_\lambda}[y_i:\eta].$$

Proof. Consider observations $y_1,\dots,y_n \in \mathcal{Y}$. By (VI.8), the log-likelihood is given by

$$\sum_{i=1}^n \log p(y_i;\vartheta) = \sum_{i=1}^n \left(-\mathbf{L}_{\lambda,\psi_\lambda}[y_i:\eta] + \psi_\lambda(y_i)\right).$$

Since the second term is independent of ϑ, maximizing the likelihood is equivalent to minimizing the sum of the divergences. It follows that the MLE $\hat{\eta}$ is the right barycenter. □

By Theorem VI.3, minimization of the λ-logarithmic divergence can be interpreted probabilistically as maximum likelihood estimation of an underlying λ-exponential family. This gives a probabilistic basis for using λ-logarithmic divergences as loss functions. Using the duality (VI.2) between exponential family and Bregman divergence, statistical methodologies such as clustering and principal component analysis were investigate in the literature [25], [58]. Some preliminary results about these statistical methodologies using λ-logarithmic divergence are reported in the recent work [48].

The next result, which is closely related to Theorem VI.3, provides a new probabilistic interpretation of the dual variable η. In classical deformation theory (Section II-B), $\eta = \widetilde{\mathbb{E}}_\vartheta[Y]$ is the escort expectation which involves the escort distribution. Here, we show that the dual parameter has a probabilistic interpretation under the original distribution without performing the escort transformation.

Theorem VI.4 (Dual variable as right barycenter)**.** *The dual variable η is the unique right barycenter of the distribution $p(\cdot;\vartheta)$:*

$$\eta = \operatorname*{arg\,min}_{\eta'\in\Omega'} \mathbb{E}_\vartheta\left[\mathbf{L}_{\lambda,\psi_\lambda}[Y:\eta']\right]. \tag{VI.9}$$

Proof. Let $\eta' \in \Omega'$ (corresponding to $\vartheta' \in \Omega$) be given. By (VI.8), we have

$$\begin{aligned} 0 &\le \mathbf{H}(p(\cdot;\vartheta)||p(\cdot;\vartheta')) \\ &= \int p(y;\vartheta)\log\frac{p(y;\vartheta)}{p(y;\vartheta')}\mathrm{d}\nu(y) \\ &= \mathbb{E}_\vartheta\left[-\mathbf{L}_{\lambda,\psi_\lambda}[y:\eta] + \mathbf{L}_{\lambda,\psi_\lambda}[y:\eta']\right]. \end{aligned}$$

Rearranging gives $\mathbb{E}_\vartheta[\mathbf{L}_{\lambda,\psi_\lambda}[y:\eta]] \le \mathbb{E}_\vartheta[\mathbf{L}_{\lambda,\psi_\lambda}[y:\eta']]$ and the theorem is proved. □

VII. Conclusion

In this paper we showed that the λ-duality, which is a one-parameter deformation of convex duality, leads to fresh perspectives on the q-exponential family (where $q = 1 - \lambda$) and related concepts. In particular, the mathematical properties of the λ-exponential family under the λ-duality nicely parallel those of the exponential family. Furthermore, the λ-mixture family, which also conforms to the same duality, may be understood as another face of the λ-exponential family. In Table I we summarize the analogy between the classical objects and the framework of this paper.

Convex Duality ($\lambda = 0$)	**λ-Duality** ($\lambda \neq 0$)
convex function (II.1)	c_λ-convex function, where $\frac{1}{\lambda}(e^{\lambda f} - 1)$ is convex (Thm II.2)
Bregman divergence (II.3)	λ-logarithmic divergence (Def II.5)
exponential family (I.1)	λ-exponential family (Def III.1)
mixture family (I.2)	λ-mixture family (Def IV.6)
Shannon entropy (I.6)	Rényi entropy (I.4)
KL-divergence (II.6)	Rényi divergence (III.12)
Dually flat geometry [1]	Dually projectively flat geometry (with constant curvature λ) [26]
2-Wasserstein transport [24], [40]	Dirichlet transport ($\lambda = -1$) [28], [27] (also see Eg III.18)

TABLE I: Analogy between our framework and the classical one based on convex duality.

We expect that our framework will be helpful in further studies in the theory and application of generalized exponential families. A natural direction is to study implications of the λ-duality in the context of statistical physics. An explicit example is the *porous medium equation* which was analyzed in [59] using the q-Gaussian distribution and information geometry (see also [60]). As described in Section VI-C, the λ-logarithmic divergences leads to many potential applications in statistics and machine learning, some of which are being investigated by the authors. The geometric concepts studied in this paper, such as the escort transformation, are closely related to the Aitchison geometry of the probability simplex in compositional data analysis [61], and it is of interest to explore deeper links between compositional data analysis and information geometry. Finally, we believe that the λ-duality and logarithmic divergences can be applied to optimization in both convex and non-convex settings.

Appendix
Proof of Theorem II.2(i)

We follow the approach of [26, Section 3] and adapt the notations there to our setting. Consider the function $\frac{1}{\lambda}e^{\lambda f}$ which is convex on Ω by assumption (here the constant term in F_λ is dropped for convenience). The tangent hyperplane based at $u' \in \Omega$ is given by

$$\begin{aligned} u \mapsto & \frac{1}{\lambda}e^{\lambda f(u')} + e^{\lambda f(u')}\nabla f(u') \cdot (u - u') \\ & = \frac{1}{\lambda}e^{\lambda f(u')}[(1 - \lambda \nabla f(u') \cdot u') + \lambda \nabla f(u') \cdot u]. \end{aligned}$$

By convex duality, $\frac{1}{\lambda}e^{\lambda f}$ is the maximum of the tangent hyperplanes where u' varies over Ω:

$$\begin{aligned} \frac{1}{\lambda}e^{\lambda f(u)} &= \max_{u' \in \Omega} \frac{e^{\lambda f(u')}}{\lambda}[(1 - \lambda \nabla f(u') \cdot u') + \nabla f(u') \cdot u] \\ &= \max_{u' \in \Omega} \frac{e^{\lambda f(u')}}{\lambda}(1 - \lambda \nabla f(u') \cdot u') \left[1 + \frac{\lambda \nabla f(u') \cdot u}{1 - \lambda \nabla f(u') \cdot u'}\right]. \end{aligned}$$

Note that in this step we use the assumption that $1 - \lambda \nabla f(u') \cdot u'$ is strictly positive. (A sufficient condition is that 0 belongs to the closure of the domain. This construction also motivates the definition of the λ-deformed Legendre transformation $v' = \nabla^{c_\lambda} f(u')$.) Now take logarithm and rearrange. Here we consider the case $\lambda > 0$; the other case $\lambda < 0$ can be handled similarly. For $u \in \Omega$, we have

$$f(u) = \max_{v' \in \Omega} \left\{ \frac{1}{\lambda} \log(1 + \lambda u \cdot v') - \widetilde{g}(v') \right\},$$

where $v' = \frac{\nabla f(u')}{1 - \lambda \nabla f(u') \cdot u'}$ and $\widetilde{g}(v')$ absorbs the other terms which depend only on v'.

Since the mapping $u' \mapsto v'$ is a diffeomorphism from Ω to the range Ω' (see Proposition 1 and Theorem 11 of [26]), we may rewrite the above as

$$\begin{aligned} f(u) &= \max_{v' \in \Omega'} \left\{ \frac{1}{\lambda} \log(1 + \lambda u \cdot v') - g(v') \right\} \\ &= \max_{v' \in \Omega'} \left\{ -c_\lambda(u, v') - g(v') \right\}, \end{aligned}$$

where $g(v') = \widetilde{g}(u')$. This shows that f is c_λ-convex on Ω.

Acknowledgment

The first author acknowledges support by NSERC Discovery Grant RGPIN-2019-04419 and a Connaught New Researcher Award. The second author acknowledges support by AFOSR Grant FA9550-19-1-0213. We also thank the referees and the associate editor for their careful reading and helpful comments.

References

[1] S.-I. Amari, *Information Geometry and Its Applications*. Springer, 2016.
[2] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.
[3] J. Naudts, *Generalised Thermostatistics*. Springer, 2011.
[4] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1-2, pp. 479–487, 1988.
[5] T. Leinster, *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press, 2021.
[6] A. Rényi, "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.
[7] J. Naudts, "Deformed exponentials and logarithms in generalized thermostatistics," *Physica A: Statistical Mechanics and its Applications*, vol. 316, no. 1-4, pp. 323–334, 2002.
[8] S.-I. Amari and A. Ohara, "Geometry of q-exponential family of probability distributions," *Entropy*, vol. 13, no. 6, pp. 1170–1185, 2011.
[9] J. Naudts, "The q-exponential family in statistical physics," *Central European Journal of Physics*, vol. 7, no. 3, pp. 405–413, 2009.
[10] S.-I. Amari, A. Ohara, and H. Matsuzoe, "Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 18, pp. 4308–4319, 2012.
[11] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. Springer, 2009.

[12] P. Douglas, S. Bergamini, and F. Renzoni, "Tunable Tsallis distributions in dissipative optical lattices," *Physical Review Letters*, vol. 96, no. 11, p. 110601, 2006.

[13] N. Ding, "Statistical machine learning in the t-exponential family of distributions," Ph.D. dissertation, Purdue University, 2013.

[14] K. Lee, S. Kim, S. Lim, S. Choi, and S. Oh, "Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning," *arXiv preprint arXiv:1902.00137*, 2019.

[15] A. F. Martins, M. Treviso, A. Farinhas, P. M. Aguiar, M. A. Figueiredo, M. Blondel, and V. Niculae, "Sparse continuous distributions and Fenchel-Young losses," *arXiv preprint arXiv:2108.01988*, 2021.

[16] V. Masrani, R. Brekelmans, T. Bui, F. Nielsen, A. Galstyan, G. Ver Steeg, and F. Wood, "q-paths: Generalizing the geometric annealing path using power means," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 1938–1947.

[17] A. Moreno, S. Nagesh, Z. Wu, W. Dempsey, and J. M. Rehg, "Kernel deformed exponential families for sparse continuous attention," *arXiv preprint arXiv:2111.01222*, 2021.

[18] R. Nock, Z. Cranko, A. K. Menon, L. Qu, and R. C. Williamson, "f-gans in an information geometric nutshell," *arXiv preprint arXiv:1707.04385*, 2017.

[19] J. Naudts, "Estimators, escort probabilities, and ϕ-exponential families in statistical physics," *Journal of Inequalities in Pure & Applied Mathematics*, vol. 5, no. 4, p. 102, 2004.

[20] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Computation*, vol. 16, no. 1, pp. 159–195, 2004.

[21] ——, "Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds," *Entropy*, vol. 15, no. 12, pp. 5384–5418, 2013.

[22] ——, "On monotone embedding in information geometry," *Entropy*, vol. 17, no. 7, pp. 4485–4499, 2015.

[23] S. Eguchi, "Information geometry and statistical pattern recognition," *Sugaku Expositions*, vol. 19, no. 2, pp. 197–216, 2006.

[24] C. Villani, *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[25] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[26] T.-K. L. Wong, "Logarithmic divergences from optimal transport and Rényi geometry," *Information Geometry*, vol. 1, no. 1, pp. 39–78, 2018.

[27] S. Pal and T.-K. L. Wong, "Exponentially concave functions and a new information geometry," *The Annals of Probability*, vol. 46, no. 2, pp. 1070–1113, 2018.

[28] ——, "The geometry of relative arbitrage," *Mathematics and Financial Economics*, vol. 10, no. 3, pp. 263–293, 2016.

[29] ——, "Multiplicative Schröodinger problem and the Dirichlet transport," *Probability Theory and Related Fields*, vol. 178, no. 1, pp. 613–654, 2020.

[30] T.-K. L. Wong, "Information geometry in portfolio theory," in *Geometric Structures of Information*. Springer, 2019, pp. 105–136.

[31] T.-K. L. Wong and J. Yang, "Pseudo-Riemannian geometry encodes information geometry in optimal transport," *Information Geometry*, vol. Advance Online Publication, 2021.

[32] S.-I. Amari, "Integration of stochastic models by minimizing α-divergence," *Neural computation*, vol. 19, no. 10, pp. 2780–2796, 2007.

[33] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, 2014.

[34] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[35] S. Abe, "Geometry of escort distributions," *Physical Review E*, vol. 68, no. 3, p. 031101, 2003.

[36] J.-F. Bercher, "Source coding with escort distributions and Rényi entropy bounds," *Physics Letters A*, vol. 373, no. 36, pp. 3235–3238, 2009.

[37] V. Pawlowsky-Glahn and A. Buccianti, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, 2011.

[38] A. M. Scarfone, H. Matsuzoe, and T. Wada, "A study of Rényi entropy based on the information geometry formalism," *Journal of Physics A: Mathematical and Theoretical*, vol. 53, no. 14, p. 145003, 2020.

[39] F. Santambrogio, *Optimal Transport for Applied Mathematicians*. Birkäuser, 2015.

[40] C. Villani, *Optimal Transport: Old and New*. Springer, 2008.

[41] L. Ambrosio and N. Gigli, "A user's guide to optimal transport," in *Modelling and Optimisation of Flows on Networks*. Springer, 2013, pp. 1–155.

[42] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[43] G. Pistone, "Nonparametric information geometry," in *International Conference on Geometric Science of Information*. Springer, 2013, pp. 5–36.

[44] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society, 2000.

[45] M. A. Kumar and K. V. Mishra, "Cramér–Rao lower bounds arising from generalized csiszár divergences," *Information Geometry*, vol. 3, no. 1, pp. 33–59, 2020.

[46] T. Van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[47] P. A. Morales and F. E. Rosas, "Generalization of the maximum entropy principle for curved statistical manifolds," *Physical Review Research*, vol. 3, no. 3, p. 033216, 2021.

[48] Z. Tao and T.-K. L. Wong, "Projections with logarithmic divergences," in *International Conference on Geometric Science of Information*. Springer, 2021, pp. 477–486.

[49] J. M. Dickey, "Three multidimensional-integral identities with Bayesian applications," *The Annals of Mathematical Statistics*, pp. 1615–1628, 1968.

[50] A. Rodríguez and C. Tsallis, "Connection between Dirichlet distributions and a scale-invariant probabilistic model based on Leibniz-like pyramids," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 12, p. P12027, 2014.

[51] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[52] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[53] G. Mateu-Figueras, G. S. Monti, and J. Egozcue, "Distributions on the simplex revisited," in *Advances in Compositional Data Analysis*. Springer, 2021, pp. 61–82.

[54] R. J. McCann, "A convexity principle for interacting gases," *Advances in Mathematics*, vol. 128, no. 1, pp. 153–179, 1997.

[55] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[56] T.-K. L. Wong and J. Yang, "Logarithmic divergences: geometry and interpretation of curvature," in *International Conference on Geometric Science of Information*. Springer, 2019, pp. 413–422.

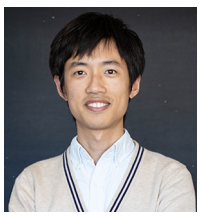
[57] J. Naudts and J. Zhang, "Rho–tau embedding and gauge freedom in information geometry," *Information Geometry*, vol. 1, no. 1, pp. 79–115, 2018.

[58] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family," in *Advances in Neural Information Processing Systems*, 2002, pp. 617–624.

[59] A. Ohara and T. Wada, "Information -Gaussian densities and behaviors of solutions to related diffusion equations," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 3, p. 035002, 2009.

[60] A. Takatsu, "Wasserstein geometry of porous medium equation," in *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, vol. 29, no. 2. Elsevier, 2012, pp. 217–232.

[61] I. Erb and N. Ay, "The information-geometric perspective of compositional data analysis," in *Advances in Compositional Data Analysis*. Springer, 2021, pp. 21–43.



Ting-Kam Leonard Wong is an Assistant Professor in Statistics at the Department of Statistical Sciences, University of Toronto. He obtained his BSc and MPhil degrees at the Chinese University of Hong Kong, and received his PhD in Mathematics from the University of Washington. Before joining U of T in 2018, he spent two years at the University of Southern California as a non-tenure track Assistant Professor in Financial Mathematics. His current research interests include mathematical finance, probability, optimal transport and information geometry, as well as applications in statistics and machine learning.

Jun Zhang is with the University of Michigan as a Full Professor of Psychology and of Statistics. He has held visiting positions at the University of Melbourne (Australia), CNRS Marseille (France), University of Waterloo (Canada), RIKEN Brain Science Institute (Japan), Center for Mathematical Sciences and Applications (CMSA) at Harvard University, and Shanghai Advanced Institute of Finance (China). He has served as the President, Vice President, and Executive Board Member of the Society for Mathematical Psychology, and as a member of the Council and a Member-at-Large on the Governing Board of the Federation of Associations in Brain and Behavioral Sciences (FABBS). He is a Fellow of the Association for Psychological Sciences (APS) and a Fellow of Psychonomic Society. He is the founding co-editor of the journal Information Geometry, has served as an Associate Editor of the Journal of Mathematical Psychology, and is a member of editorial boards of various journals. He directs the M3 Lab ("Mind, Machine and Mathematics") conducting research in neuronal and brain signal analysis, computation vision, cognitive modeling, machine learning, brain-like computation and artificial intelligence. His current research effort is devoted to information geometry and geometrization of science of information.