λ-Deformed probability families with subtractive and divisive normalizations

Jun Zhang[a,*] and Ting-Kam Leonard Wong[b]

[a]*Department of Psychology and Department of Statistics, University of Michigan, Ann Arbor, MI, United States*

[b]*Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada*

[*]*Corresponding author: e-mail: junz@umich.edu*

Abstract

This chapter investigates deformations to the exponential family and the mixture family of probability density functions, under both subtractive and divisive normalizations. We study the two normalizations under the so-called "λ-deformation" which captures the Tsallis deformation and Rényi deformation as two sides of the same coin; the resulting deformed exponential families only differ by a reparameterization. Our λ-deformation is essentially a parametric deformation of the exponential function by $\exp_\lambda(t) = (1+\lambda t)^{1/\lambda}$ coupled with a λ-deformed Legendre duality characterized by $\kappa_\lambda(t) = \frac{1}{\lambda}\log(1+\lambda t) = \log(\exp_\lambda(t))$. This λ-deformed Legendre duality and the regular Legendre duality induce the same conjugate variable once the notion of λ-gradient is introduced. The deformed λ-exponential and λ-mixture families turn out to be nearly identical apart from a transformation of the underlying random variables. We remark that the induced information geometry has a pair of constant-curvature dual connections and a conformal Hessian metric, where the λ parameter encodes the curvature value.

Keywords: Legendre duality, c-Duality, Deformed exponential family, Tsallis entropy, Rényi entropy, Conformal Hessian, Constant curvature

1 Introduction

Exponential families of probability distributions play important roles in, among other disciplines, probability and statistics, information theory, statistical physics, etc. with a well-understood geometry (Amari, 2016). Given a dominating measure μ, an exponential family is a parameterized probability density $p^{(e)}(\cdot|\theta)$ of the form

Handbook of Statistics. https://doi.org/10.1016/bs.host.2021.06.003

$$p^{(e)}(\zeta|\theta) = e^{\theta \cdot F(\zeta) - \phi(\theta)}, \tag{1}$$

where $\theta = (\theta^1, \ldots, \theta^d) \in \mathbb{R}^d$ are parameters, $F(\zeta) = (F_1(\zeta), \ldots, F_d(\zeta))$ are functions of a random variable denoted ζ with values in some sample space $\mathcal{X}$, with the shorthand $\theta \cdot F = \sum_{i=1}^d \theta^i F_i(\zeta)$. The dominating measure $\mu(d\zeta) \equiv d\mu$ may possibly incorporate another function, i.e., $d\nu = e^{F_0(\zeta)} d\mu$ through a change of measure $d\mu \to d\nu$, where $e^{F_0(\zeta)}$ plays the role of the Radon–Nikodym derivative. Also, $\phi(\theta)$ is the cumulant generating function defined by the probability normalization $\int p^{(e)} d\mu = 1$, that is,

$$\phi(\theta) = \log \int e^{\theta \cdot F(\zeta)} \, d\mu.$$

In statistical physics, $Z(\theta) = e^{\phi(\theta)}$ is called the partition function. The natural parameter space is the set of θ such that $\phi(\theta) < \infty$. It is easily shown that the cumulant generating function $\phi(\theta) = \log Z(\theta)$ is convex in θ, with its Hessian matrix being positive semidefinite:

$$\frac{\partial^2 \phi}{\partial \theta^i \partial \theta^j} = \int p^{(e)}(\zeta|\theta) \left(F_i(\zeta) - \int p^{(e)} F_i \, d\mu \right) \left(F_j(\zeta) - \int p^{(e)} F_j \, d\mu \right) d\mu. \tag{2}$$

The convex conjugate of $\phi(\theta)$, defined in terms of the ordinary Legendre duality and denoted by ϕ^*, turns out to be the (negative) Shannon entropy function $\phi^*(\eta) := -\mathbf{S}[p^{(e)}]$, where $p^{(e)}$ is parameterized by the conjugate variable $\eta = (\eta_1, \ldots, \eta_d)$, called the expectation parameters and defined by $\eta = \nabla \phi$, the gradient of ϕ, which evaluates to

$$\eta_i = \int p^{(e)}(\zeta|\theta) F_i(\zeta) \, d\mu. \tag{3}$$

The Kullback–Leibler divergence (also called the relative entropy) $\mathbf{D}[p||p'] = \int p \log(p/p') \, d\mu$ between two members $p = p^{(e)}(\cdot|\theta)$ and $p' = p^{(e)}(\cdot|\theta')$ of an exponential family equals nothing but the *Bregman divergence* of ϕ. Namely, we have

$$\mathbf{D}[p||p'] = \mathbf{B}_\phi(\theta', \theta) = \phi(\theta') - \phi(\theta) - \nabla\phi(\theta) \cdot (\theta' - \theta), \tag{4}$$

with

$$\mathbf{B}_\phi(\theta, \theta') = \mathbf{B}_{\phi^*}(\eta', \eta).$$

Consequently, the second-order approximation of $\mathbf{B}_\phi(\theta, \theta')$ around $\theta \approx \theta'$, which gives the Riemannian metric on the statistical manifold of a parametric family known as the *Fisher–Rao metric*, turns out to be a Hessian metric (2) for an exponential family. For details, see Amari (2016, Chapter 1).

The *mixture family* is another probability family which is very useful in both theory and applications. Let $P_0(\zeta), P_1(\zeta), \ldots, P_d(\zeta)$ be a set of affinely independent probability densities, all with respect to a common dominating measure μ. Given mixture parameters $\eta_i > 0$ for $i = 0, \ldots, d$ with $\sum_{i=0}^{d} \eta_i = 1$, the mixture family $p^{(m)}$ is defined by

$$p^{(m)}(\zeta|\eta) = \sum_{i=0}^{d} \eta_i P_i(\zeta) = P_0(\zeta) + \sum_{i=1}^{d} \eta_i (P_i(\zeta) - P_0(\zeta)), \tag{5}$$

with $\eta = (\eta_1, \ldots, \eta_d) \in \{\eta_i > 0 : \sum_{i=1}^{d} \eta_i < 1\}$ taken as the (independent) parameters. The negative Shannon entropy

$$\psi(\eta) = -\mathbf{S}[p^{(m)}(\cdot\,|\eta)] = \int p^{(m)}(\zeta|\eta) \log p^{(m)}(\zeta|\eta)\, d\mu$$

of a mixture family $p^{(m)}(\cdot|\eta)$ can be shown to be convex in the mixture parameters η. The Kullback–Leibler divergence $\mathbf{D}[p||p']$ between two members $p = p^{(m)}(\cdot|\eta)$ and $p' = p^{(m)}(\cdot|\eta')$ of a mixture family also turns out to be a Bregman divergence, where $\psi(\eta)$ is the convex potential function:

$$\mathbf{D}[p||p'] = \mathbf{B}_{\psi}(\eta, \eta'). \tag{6}$$

The convex conjugate of $\psi(\eta)$ is $-\int P_0(\zeta)\ \log p^{(m)}\ d\mu$, with conjugate parameters $\theta = \nabla\psi$ given by $\theta^i = \int (P_i(\zeta) - P_0(\zeta))\ \log p^{(m)}\ d\mu$.

Shannon entropy and Kullback–Leibler divergence are widely used in science and engineering. It is well known (see e.g., Amari and Nagaoka, 2000) that the principle of maximum entropy (or of minimum Kullback–Leibler divergence) with linear constraints on the expectation leads to an exponential family of probability density functions (thus generalizing the Boltzmann–Gibbs distribution), and that the corresponding manifold of parametric density functions is a dually flat Hessian manifold. In such a manifold, the Riemannian metric g is the Hessian of a convex potential function (2), and the natural and mixture coordinates are biorthogonal, meaning $g\left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \eta_j}\right) = \delta_{ij}$. As an important special case, we note that positive probability measures over a finite set (i.e., the probability simplex) is at the same time an exponential family and a mixture family, and in this case the expectation parameter of the exponential representation coincides with the mixture parameter of the mixture representation. This probability simplex will be studied in Section 5.

1.1 Deformation models

Despite the elegance and profound applications of these classic pillars in statistics and information theory, considerations in e.g., nonextensive statistical physics and nonlinear diffusion led to generalizations of the Shannon entropy and the Boltzmann–Gibbs distribution.

There have been systematic efforts to generalize the Shannon entropy $\mathbf{S}[p] = -\int p \log p\ d\mu$ (of *one* probability density function p) and Kullback–Leibler divergence or relative entropy $\mathbf{D}[p||p'] = \int p \log (p/p')\ d\mu$ (between *two* probability density functions p and p') to more general analytic forms; here and throughout the chapter the notation $[\cdot]$ indicates a functional which involves integrating out the random variable ζ. To use a functional form other than $\exp$ (exponential function) or $\log$ (logarithm function) is referred to as *deformation* in the statistical and information-theoretic contexts, and the resulting probability families are called "deformed" families. Typically, this is done by regarding the logarithm function $\log$, or equivalently the exponential function $\exp$, as special cases of some parametric classes of functions that include $\exp$ and $\log$ as special members.

In the context of statistical physics, Tsallis introduced in Tsallis (1988) the generalized entropy

$$\mathbf{H}_q^{\text{Tsallis}}[p] = \frac{1}{q-1}\left(1 - \int (p(\zeta))^q \, d\mu\right) = \int p \log_q\left(\frac{1}{p}\right), \, d\mu, \tag{7}$$

for a given index $q \in \mathbb{R}$, $q \neq 1$, through the q-logarithm and q-exponential functions (Tsallis, 1994):

$$\log_q(t) = \frac{1}{1-q}\left(t^{1-q} - 1\right), \quad \exp_q(t) = [1 + (1-q)t]_+^{1/(1-q)}, \tag{8}$$

where $x_+ = \max(x, 0)$. The q-logarithm reduces to the standard logarithm as $q \to 1$. Taking the limit $q \to 1$ in (7) recovers the Shannon entropy $\mathbf{S}[p]$.

Later, Kaniadakis (2001) introduced the κ-model, where $\kappa \in \mathbb{R}$, $\kappa \neq 0$:

$$\log_\kappa(t) = \frac{1}{2\kappa}(t^\kappa - t^{-\kappa}), \quad \exp_\kappa(t) = \left(\kappa t + \sqrt{1 + \kappa^2 t^2}\right)^{\frac{1}{\kappa}}.$$

Similarly, taking $\kappa \to 0$ yields the standard exponential/logarithm.

Using an axiomatic approach, Rényi (1961) introduced the Rényi entropy for a given index parameter $q \in \mathbb{R}_{>0}$, $q \neq 1$. For our purpose, we use $\lambda = 1 - q$ and write

$$\mathbf{H}_\lambda^{\text{Rényi}}[p] := \frac{1}{\lambda}\log\left(\int p^{1-\lambda}(\zeta)\, d\mu\right). \tag{9}$$

It is a monotonic transformation of the Tsallis entropy:

$$\mathbf{H}_q^{\text{Tsallis}}[p] = \frac{1}{\lambda}\left(e^{\lambda \mathbf{H}_\lambda^{\text{Rényi}}[p]} - 1\right). \tag{10}$$

In the context of information geometry, Amari and Nagaoka (2000) adopted the α-*embedding* function $l^{(\alpha)} : \mathbb{R}_{>0} \to \mathbb{R}$, defined as

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1, \\ \dfrac{2}{1-\alpha}\, t^{(1-\alpha)/2} & \alpha \neq 1. \end{cases}$$

The α-embedding leads to the so-called α-divergence. Note that the q-embedding and α-embedding functions are slightly different but related:

$$\log_q(t) = l^{(\alpha)}(t) - \frac{2}{1-\alpha}, \quad \alpha = 2q - 1.$$

These parametric deformation schemes were investigated by a general approach using nonparametric embedding functions, as explained below.

1.2 Deformed probability families: General approach

Going beyond these "parametric" approaches of extending the exponential/logarithmic function is the nonparametric approach, that is, to use an infinite-dimensional function class that includes exp or log as a special member. Several approaches can be found in the literature, including the ϕ-deformed exponential approach by Naudts (2004, 2008, 2011), the conjugate (ρ, τ)-embedding approach by the first author Zhang (2004, 2013, 2015), and the U-model by Murata et al. (2004) and Eguchi (2006). The ϕ-model and U-model are both one-function models, while the (ρ, τ)-model uses two free functions. The aforementioned parametric deformation such as α-, q-, and κ-models can be viewed as special cases of the above function class models by adopting the specific α-, q-, and κ-embedding functions. These models have been applied to nonparametric probability families in Newton (2012), Montrucchio and Pistone (2017), and de Andrade et al. (2021).

It eventually became clear in the 2018 paper Naudts and Zhang (2018) by Naudts and the first author that (i) the ϕ-model and U-model turned out to be equivalent; (ii) they are special cases of the (ρ, τ)-model upon a particular fixing of the "gauge freedom"; (iii) the corresponding (ρ, τ)-geometry of the manifold of ϕ-exponential family can have different appearances (gauge freedom), such as a Hessian geometry (under one type of gauge selection) and a conformal Hessian geometry (under another type of gauge selection). The work of Naudts and Zhang (2018) unifies intermediary results in Ohara et al. (2012), Amari et al. (2012), and Matsuzoe (2014), and provides a general deformation framework that preserves the rigid interlock of (i) the function form of entropy, cross-entropy, and relative entropy (divergence); (ii) the functional form of the probability family with the corresponding normalization and potential, and the duality between the natural and expectation parameterizations; (iii) the expressions of the Riemannian metric (Fisher–Rao metric in general and Hessian metric in particular) and of the conjugate connections.

The deformation approaches described above are based on the application of Legendre (convex) duality to various embedding functions. As explained in Section 3.1, under the classical Legendre duality, the exponential and mixture families are naturally associated to the Shannon entropy $\mathbf{S}[\cdot]$ and the Kullback–Leibler divergence $\mathbf{D}[\cdot||\cdot]$. The hallmark of these models is that a generalized Pythagorean theorem holds for the induced (Bregman) divergences.

More recently, there is another line of work by the second author and his collaborators (Pal and Wong, 2016, 2018, 2020; Wong, 2018, 2019; Wong and Yang, 2021), based on the geometry of optimal transport, that deforms the Legendre duality. Under the framework of the current paper, the deformation is given by

$$\kappa_\lambda(t) = \frac{1}{\lambda} \log(1 + \lambda t), \tag{11}$$

where $\lambda \neq 0$ is a parameter.[a] This generalized duality will be explained in Section 3. When $\lambda \to 0$, then $\kappa_\lambda(t) \to t$, so $\exp(\kappa_\lambda(t)) \to \exp(t)$. This approach leads to a novel divergence function, called λ-logarithmic divergence below, that induces a geometry with constant curvature and satisfies a generalized Pythagorean theorem. As explained below, this deformation framework amounts to a divisive normalization, which is different from the above-mentioned (ρ, τ) approach that uses a subtractive normalization.

In our recent paper Wong and Zhang (2021), the subtractive and divisive approaches to deformed exponential family are examined through the concept of λ-duality using the deformation function (11). In this chapter we provide a review of the subtractive and divisive normalization approaches, including an in-depth exposition of the λ-duality in Wong and Zhang (2021). Our framework leads to a unified way of looking at the Tsallis entropy (related to the subtractive denormalization) and Rényi entropy (related to the divisive normalization), as well as generating new insights in the distinction between the exponential and mixture families through the lens of deformation theory.

1.3 Chapter outline

The rest of this chapter is organized as follows:

Section 2 first discusses the issue of subtractive and divisive normalizations while deforming the exponential family. The special case of λ-deformation, namely deforming the exponential and logarithm by a suite of four closely linked parametric functions (which include κ_λ defined by (11)), is investigated at length. We introduce the λ-exponential family which unifies the q-exponential family and the $\mathcal{F}^{(\pm\alpha)}$-families introduced in Wong (2018), and show that the two deformations differ only with a change of parameterization. In other words, the λ-exponential family can take the forms of both subtractive normalization and divisive normalization, which differ only by reparameterization, to yield Tsallis' q-exponential family or the $\mathcal{F}^{(\pm\alpha)}$ -families. They are associated with, respectively, Tsallis and Rényi entropies, which are themselves related by a monotonic transform. Dual to λ-exponential family is the λ-mixture family, which are again equivalent after a suitable reparameterization. So our λ-deformation approach yields a unified view of the generalized exponential and mixture families, with both appearances of subtractive or divisive normalization.

[a]Note that this κ_λ (which is a family of functions with λ as the parameter) is different from and unrelated to the κ (which is a constant) in the κ-model of Kaniadakis.

Section 3 then investigates in detail the λ-duality (as a generalization of the Legendre duality), which underlies both the λ-exponential and λ-mixture families. The λ-duality corresponds to a generalized convexity notion that we call λ-convexity; it reduces to ordinary convexity for $\lambda = 0$. The generalized gradient of a λ-convex function involves an adjustment factor. This λ-duality is a special case of the so-called c-duality in optimal transport theory (Villani, 2003, 2008). Under the λ-duality we obtain the λ-logarithmic divergence, for which a generalized Pythagorean theorem holds. For a λ-exponential family under the λ-duality, the Rényi entropy (9) and Rényi divergence are the natural objects generalizing Shannon entropy and Kullback–Leibler divergence.

In Section 4 we show that appropriately defined potential functions of the λ-exponential and λ-mixture families are indeed λ-convex functions. In fact, they are related to the Rényi entropy, and the induced λ-logarithmic divergences are Rényi divergences. We state the Rényi entropy maximizing property of λ-exponential family, where the constraint is cast in terms of the escort expectation. Also, the λ-exponential family can be constructed from the λ-logarithmic divergence functions. This generalizes a known connection between exponential family and Bregman divergence.

Section 5 provides the calculation of the deformed probability simplex as an example of the λ-exponential and λ-mixture families. Finally, Section 6 provides a short summary and conclusion.

2 λ-Deformation of exponential and mixture families

2.1 λ-Deformation

Consider an exponential family

$$p^{(e)}(\zeta|\theta) = \exp\{\theta \cdot F(\zeta) - \phi(\theta)\} = \frac{1}{Z(\theta)} \exp\{\theta \cdot F(\zeta)\}.$$

The cumulant generating function $\phi(\theta)$ (also called the potential function or subtractive normalization function) and the partition function $Z(\theta)$ (also called the divisive normalization function) are used to ensure the probability normalization $\int p^{(e)}(\zeta|\theta) d\mu = 1$. Clearly, they are related to each other by the simple relation $\phi = \log Z$. For the exponential family the equivalence of using ϕ and Z is due to the elementary but crucial property $\exp(x+y) = \exp(x)\exp(y)$ of the exponential function. However, when we deform the exponential embedding, the first issue we encounter is the need to treat separately the subtractive and divisive normalizations which are no longer equivalent in general.

Let us introduce some notations. Consider the following deformed logarithm and exponential functions:[b]

[b]This reduces to $\exp_q$ defined in (12) where $q = 1 - \lambda$. The subscript indicates the parameterization used.

$$\log_\lambda(t) = \frac{1}{\lambda}\left(t^\lambda - 1\right), \quad \exp_\lambda(t) = (1 + \lambda t)^{1/\lambda}.$$

More precisely, we define $\exp_\lambda : \mathbb{R} \to [0, \infty)$ (where $\lambda \in \mathbb{R}, \lambda \neq 0$) by

$$\exp_\lambda(t) = [1 + \lambda\, t]_+^{1/\lambda}. \tag{12}$$

Also $\frac{d}{dt}\exp_\lambda(t) = [\exp_\lambda(t)]^{1-\lambda}$, so $\exp_\lambda(\cdot)$ is convex if and only if $\lambda < 1$. For this reason we restrict λ to this range as in Amari and Ohara (2011) and Naudts (2011). In applications, we assume implicitly that $1 + \lambda t > 0$.

Note that our notation differs slightly from Tsallis' deformed-logarithm and deformed-exponential functions given by Eq. (8); the utility of our approach will become clear below. Next, we construct another pair of inverse functions κ_λ, γ_λ by

$$\kappa_\lambda = \log \circ \exp_\lambda, \quad \gamma_\lambda = \log_\lambda \circ \exp,$$

where $\circ$ denotes function composition. Explicitly, they are

$$\kappa_\lambda(t) = \frac{1}{\lambda}\log(1 + \lambda t), \quad \gamma_\lambda(t) = \frac{1}{\lambda}\left(e^{\lambda t} - 1\right).$$

This suite of four functions, namely $\exp_\lambda$, $\log_\lambda$, κ_λ, and γ_λ, will be called λ-deformation and used in the λ-duality. Regular exponential and logarithmic functions are recovered when $\lambda \to 0$, whence both κ_λ and γ_λ reduce to the identity.

Interestingly, the deformation functions γ_λ used in this chapter are closely related to the famous Box–Cox transformation (Box and Cox, 1964) in statistics, used to transform monotonically the data for stabilizing variance and for improving the validity of correlation measures. Though the same symbol λ is used as a free parameter, our λ-deformation theory goes far beyond the mere formula by developing the generalized Legendre duality behind it.

2.2 Deformation: Subtractive approach

Let ρ be a smooth and strictly increasing function defined on $[0, \infty)$. The *ρ-deformed exponential family* (Naudts, 2004; Zhang, 2004; Naudts and Zhang, 2018), under subtractive normalization, takes the form

$$\rho(p(\zeta|\theta)) = \theta \cdot F(\zeta) - \phi(\theta),$$

where $\theta \cdot F(\zeta) = \sum_{i=1}^d \theta^i F_i(\zeta)$, and $\phi(\theta)$ is specified by the normalization

$$1 = \int p(\zeta|\theta)\, d\mu = \int \rho^{-1}(\theta \cdot F(\zeta) - \phi(\theta))\, d\mu.$$

Here it is assumed implicitly that $\theta \cdot F - \phi$ is in the domain of ρ^{-1}. Taking derivative on both sides and rearranging,[f] we obtain

[f]We assume that the family is sufficiently regular such that all needed analytical operations, including differentiation under the integral sign, can be performed.

$$\frac{\partial \phi}{\partial \theta^i}(\theta) = \frac{\int (\rho^{-1})'(\rho(p))F_i(\zeta)\, d\mu}{\int (\rho^{-1})'(\rho(p))\, d\mu}.$$

Given ρ, the transformation: $p \mapsto (\rho^{-1})'(\rho(p))$ is called the *escort map*, and

$$\widetilde{p}(\zeta) = \frac{(\rho^{-1})'(\rho(p(\zeta))}{\int (\rho^{-1})'(\rho(p(\zeta))\, d\mu}$$

is called the *escort transformation* of the probability density function $p(\zeta)$.

Specializing to the λ-deformation, where we take the embedding function $\rho(t) = \log_\lambda(t)$, so that $\rho^{-1}(t) = \exp_\lambda(t)$, the corresponding deformed-exponential family under subtractive normalization is given by

$$p(\zeta|\theta) = \exp_\lambda(\theta \cdot F(\zeta) - \phi_\lambda(\theta)). \tag{13}$$

This *subtractive* normalization (i.e., a subtractive term within $\exp_\lambda$ in (13)) together with $\int p(\zeta|\theta)d\mu = 1$ defines the function $\phi_\lambda(\theta)$, called the (subtractive) λ-potential function, which satisfies

$$\frac{\partial \phi_\lambda}{\partial \theta^i} = \int \widetilde{p}(\zeta|\theta)F_i(\zeta)\, d\mu \tag{14}$$

with the escort transformation given by

$$\widetilde{p}(\zeta|\theta) = \frac{(p(\zeta|\theta))^{1-\lambda}}{\int (p(\zeta|\theta))^{1-\lambda}\, d\mu}. \tag{15}$$

Clearly, when $\lambda \to 0$ we recover the conventional exponential family (1). Traditionally, one uses the parameter $q = 1 - \lambda$ and calls (13) the *q-exponential family* (also called the *Tsallis distribution*). The q-exponential family was originally obtained as the solution to maximization of Tsallis entropy subject to constraints on the escort expectation. See Section 4.3 for more discussions.

2.3 Deformation: Divisive approach

To deform the exponential family through divisive normalization, we use a smooth monotone function $c(\cdot)$, and define a parametric probability family which takes the form

$$\log p(\zeta|\vartheta) = c(\vartheta \cdot F(\zeta)) - \varphi(\vartheta)$$

Note that we use the symbol ϑ to distinguish it from the parameter θ in the subtractive case. Here

$$\varphi(\vartheta) = \log \int e^{c(\vartheta \cdot F(\zeta))} \, d\mu \tag{16}$$

is the divisive normalization function, and it is assumed that

$$\int e^{c(\vartheta \cdot F(\zeta))} d\mu < \infty$$

in the domain of ϑ (the natural parameter set).

In the special case of λ-deformation, the function $c(\cdot)$ takes the form

$$\kappa_\lambda(t) = \frac{1}{\lambda} \log (1 + \lambda t). \tag{17}$$

From this deformation function, the resulting family is

$$p(\zeta|\vartheta) = (1 + \lambda\vartheta \cdot F(\zeta))_+^{1/\lambda} \, e^{-\varphi_\lambda(\vartheta)}, \tag{18}$$

where

$$\varphi_\lambda(\vartheta) = \log \int (1 + \lambda \, \vartheta \cdot F(\zeta))_+^{1/\lambda} d\mu$$

is finite on the parameter set. This family unifies the $\mathcal{F}^{(\pm\alpha)}$-*families* introduced in Wong (2018).

It is easy to show that $\Pr_\vartheta(1 + \lambda \, \vartheta \cdot F(\zeta) > 0) = 1$, because $1 + \lambda\vartheta \cdot F(\zeta) \leq 0$ leads to $\exp_\lambda(\vartheta \cdot F(\zeta))$ being either 0 or $+\infty$. Note that in contrast to the exponential family, it is possible that the support of the density depends on the parameter ϑ, as in the case of q-exponential family, see Wong and Zhang (2021). To avoid technicalities, we assume that the support of $p(\zeta|\vartheta)$ is independent of ϑ.

2.4 Relation between the two normalizations

Under the λ-deformation there is an intrinsic link between the subtractive and divisive normalizations, namely the link between Eqs. (13) and (18). We begin with the observation

$$e^{\kappa_\lambda(t)} = (1 + \lambda t)_+^{1/\lambda} = \exp_\lambda(t). \tag{19}$$

In fact, we can find conditions for which the divisively normalized density $p(\zeta|\vartheta)$ and the subtractively normalized $p(\zeta|\theta)$ are equal:

$$p(\zeta|\vartheta) = p(\zeta|\theta),$$

or, explicitly,

$$(1 + \lambda\vartheta \cdot F(\zeta))_+^{1/\lambda} \, e^{-\varphi_\lambda(\vartheta)} = (1 + \lambda(\theta \cdot F(\zeta) - \phi_\lambda(\theta)))_+^{1/\lambda}.$$

Taking λ-th power and equating both sides, we obtain the conditions

$$\theta = \vartheta e^{-\lambda\varphi_\lambda(\vartheta)} \Leftrightarrow \vartheta = \frac{\theta}{1-\lambda\phi_\lambda(\theta)}, \tag{20}$$

$$\phi_\lambda(\theta) = \frac{1}{-\lambda}(e^{-\lambda\varphi_\lambda(\vartheta)} - 1) \Leftrightarrow \varphi_\lambda(\vartheta) = -\frac{1}{\lambda}\log(1-\lambda\phi_\lambda(\theta)). \tag{21}$$

We summarize the above observations with the following

Proposition 1. (Reparameterization equivalence). *The deformed exponential family $p^{(\lambda)}(\zeta|\theta)$ under subtractive normalization can be reparameterized as the deformed exponential family $p^{(\lambda)}(\zeta|\vartheta)$ under divisive normalization:*

$$p^{(\lambda)}(\zeta|\theta) = \exp_\lambda(\theta \cdot F(\zeta) - \phi_\lambda(\theta)) = \exp_\lambda(\vartheta \cdot F(\zeta))e^{-\varphi_\lambda(\vartheta)} = p^{(\lambda)}(\zeta|\vartheta). \tag{22}$$

Remark 1. This means that the q-exponential family of Tsallis and the $F^{(\pm\alpha)}$-family of Wong (2018) are in fact reparameterizations of one another.

2.5 λ-Exponential and λ-mixture families

To formalize the above discussion and our λ-deformation framework, we make the following definition.

Definition 1. (λ-exponential family). Let $\lambda \neq 0$. The λ-exponential family, with respect to a given reference (dominating) measure μ and a fixed vector of random functions $F(\zeta) = (F_1(\zeta), \dots, F_d(\zeta))$, is the parameterized density given by the expressions

$$p^{(\lambda)}(\zeta|\cdot) = \exp_\lambda(\vartheta \cdot F(\zeta))e^{-\varphi_\lambda(\vartheta)} = \exp_\lambda(\theta \cdot F(\zeta) - \phi_\lambda(\theta)), \tag{23}$$

where the last two expressions are related by Proposition 1.

Note again that we use ϑ for the divisive normalization setting and distinguish it from θ for the subtractive normalization setting. The function $\phi_\lambda(\theta)$ is called *subtractive* λ-potential and used for *subtractive* normalization, while $\varphi_\lambda(\vartheta)$ is called *divisive* λ-potential and used for *divisive* normalization. They satisfy

$$\phi_\lambda(\theta) = \gamma_{-\lambda}(\varphi_\lambda(\vartheta)) \Leftrightarrow \varphi_\lambda(\vartheta) = \kappa_{-\lambda}(\phi_\lambda(\theta)),$$

where

$$\kappa_{-\lambda}(t) = -\frac{1}{\lambda}\log(1-\lambda t), \qquad \gamma_{-\lambda}(t) = \frac{1}{\lambda}\left(1-e^{-\lambda t}\right).$$

For later convenience, we also note

$$e^{-\lambda\varphi_\lambda(\vartheta)} = 1 - \lambda\phi_\lambda(\theta). \tag{24}$$

We next define a mixture-type family dual to the λ-exponential family, in an analogous way that an exponential family is dual to the mixture family. The form of the family will be justified in Section 4 where we show that it is compatible with the λ-duality.

Definition 2. (λ-mixture family). Let $\lambda \neq 0, 1$ be given. The λ-mixture family with respect to a fixed set of densities $P_0(\zeta), P_1(\zeta), \ldots, P_d(\zeta)$ is defined by

$$p^{(\lambda)}(\zeta|\eta) = \frac{1}{Z_\lambda(\eta)}\left(\sum_{i=0}^{d}\eta_i\widetilde{P}_i(\zeta)\right)^{1/(1-\lambda)}, \tag{25}$$

where $\eta = (\eta_1, \cdots, \eta_d)$ is the mixture parameter satisfying $0 \leq \eta_i \leq 1$ and $\eta_0 = 1 - \sum_{i=1}^{d}\eta_i > 0$. Here $\widetilde{P}_i$ denotes the escort transformation of P_i given by (15):

$$\widetilde{P}_i(\zeta) = \frac{(P_i(\zeta))^{1-\lambda}}{\int (P_i(\zeta))^{1-\lambda}\, d\mu},$$

and $Z_\lambda(\eta)$ represents the integral

$$Z_\lambda(\eta) = \int \left(\sum_{i=0}^{d}\eta_i\widetilde{P}_i(\zeta)\right)^{1/(1-\lambda)} d\mu$$

which is assumed to converge for all η and can be differentiated under the integral sign.

Denote

$$\widetilde{\eta} = \eta Z_\lambda(\eta)^{-1/(1-\lambda)}.$$

Then we have the following succinct expression for a λ-mixture family:

$$p^{(\lambda)}(\zeta|\widetilde{\eta}) = \left(\sum_{i=0}^{d}\widetilde{\eta}_i\widetilde{P}_i(\zeta)\right)^{1/(1-\lambda)}.$$

Let a λ-mixture family be given. Because

$$\begin{aligned}
&\frac{1}{Z_\lambda(\eta)}\left(\sum_{i=0}^{d}\eta_i\widetilde{P}_i(\zeta)\right)^{1/(1-\lambda)} \\
&= e^{-\log Z_\lambda(\eta)}\left(\left(1 - \sum_{i=1}^{d}\eta_i\right)\widetilde{P}_0(\zeta) + \sum_{i=1}^{d}\eta_i\widetilde{P}_i(\zeta)\right)^{1/(1-\lambda)} \\
&= \left(1 + \sum_{i=1}^{d}\eta_i\frac{\widetilde{P}_i(\zeta) - \widetilde{P}_0(\zeta)}{\widetilde{P}_0(\zeta)}\right)^{1/(1-\lambda)} e^{-\log Z_\lambda(\eta)}(\widetilde{P}_0(\zeta))^{1/(1-\lambda)},
\end{aligned}$$

we can set

$$F_i(\zeta) = \frac{1}{1-\lambda} \frac{\widetilde{P}_i(\zeta) - \widetilde{P}_0(\zeta)}{\widetilde{P}_0(\zeta)},$$

with $d\nu = (\widetilde{P}_0(\zeta))^{1/(1-\lambda)} d\mu$. The density of $p^{(\lambda)}(\zeta|\eta)$ with respect to ν then has the form

$$p^{(\lambda)}(\zeta|\eta) = (1 + (1-\lambda)\,\eta \cdot F(\zeta))^{1/(1-\lambda)} e^{-\psi_{1-\lambda}(\eta)},$$

where $\psi_{1-\lambda}(\eta) = \log Z_\lambda(\eta)$. Thus we have shown the following:

Proposition 2. (Relation between λ-exponential and λ-mixture families). *Suppose $\lambda \neq 0, 1$. A λ-mixture family with pure densities*

$$P(\zeta) = \{P_0(\zeta), P_1(\zeta), \cdots, P_d(\zeta)\}$$

becomes a λ-exponential family with the vector of random functions

$$F(\zeta) = \{F_1(\zeta), \cdots, F_d(\zeta)\}$$

after a transformation of its random variables $P(\zeta) \to F(\zeta)$ and the dominating measure.

Remark 2. That the λ-mixture family may be regarded as the "dual" of the λ-exponential family is true only for $\lambda \neq 0, 1$, as it is well known that in general a mixture family cannot be expressed as an exponential family. As shown in Section 5, in the case of the probability simplex (where reference densities are index functions $P_i(\zeta) = \delta_i(\zeta)$, i.e., ζ is on discrete support), then it is at the same time a λ-exponential and λ-mixture family, for any λ. Probability simplex is special in this regard.

3 Deforming Legendre duality: λ-Duality

3.1 From Bregman divergence to λ-logarithmic divergence

Given a function f on $\mathbb{R}^d$, its convex conjugate is defined by

$$f^*(y) = \sup_x (x \cdot y - f(x)), \quad y \in \mathbb{R}^d. \tag{26}$$

If f is convex and lower semicontinuous, then $(f^*)^* = f$. When f is convex and differentiable, the Legendre transformation

$$y = \nabla f(x), \tag{27}$$

which can be motivated by the first-order condition in Eq. (26), defines a "dual variable" y, and we have $x = \nabla f^*(y)$ (provided the second derivative or Hess f is positive definite). The function f also defines a Bregman divergence given by

$$\mathbf{B}_f(x, x') = f(x) - f(x') - \nabla f(x') \cdot (x - x') \geq 0. \tag{28}$$

The Bregman divergence satisfies the *reference-representation biduality* (Zhang, 2005, 2013) in the sense that

$$\mathbf{B}_f(x,x') = \mathbf{B}_{f^*}(y',y).$$

Note that when f is convex and differentiable, the nonnegativity of the Bregman divergence encodes the fact that for any x, x' we have

$$f(x) - f(x') \geq \nabla f(x') \cdot (x - x').$$

Now, instead of assuming f to be convex, we assume that

$$g_\lambda(x) := \gamma_\lambda(f(x)) = \frac{1}{\lambda}\left(e^{\lambda f(x)} - 1\right)$$

as a function of x is convex and differentiable on some convex domain Ω.

Definition 3. (λ-exponential convexity and concavity). Let $\Omega \subset \mathbb{R}^d$ be convex set. A smooth function $f : \Omega \to \mathbb{R}$ is said to be λ-exponentially convex (or concave, respectively), or in short λ-convex (or λ-concave), if $g_\lambda(x) \equiv \gamma_\lambda(f(x))$ is convex (or concave) in x and that the Hessian of g_λ is strictly positive (or negative) definite.

Note that the additive term $-1/\lambda$ in $g_\lambda(x) = \frac{1}{\lambda}(e^{\lambda f(x)} - 1)$ is not necessary; it is included so that we have $\lim_{\lambda \to 0} g_\lambda(x) = f(x)$ and we recover ordinary convexity.

The following can be easily proven:

Proposition 3. *For λ a fixed positive number, $\lambda > 0$, we have*
(i) f is λ-convex if and only if $-f$ is $(-\lambda)$-concave;
(ii) f is λ-concave if and only if $-f$ is $(-\lambda)$-convex.

Remark 3. Previously, (Pal and Wong, 2016, 2018; Wong, 2018) used the notions of α-exponentially convexity and α-exponentially concavity, where α is positive. When speaking of λ-convexity or λ-concavity, we do *not* restrict to using only positive λ values.

For a λ-convex function f, we now apply the convexity argument to $g_\lambda(x) = \gamma_\lambda(f(x))$ and write

$$g_\lambda(x) - g_\lambda(x') \geq \nabla g_\lambda(x') \cdot (x - x').$$

In terms of f, we have, after some manipulations,

$$\gamma_\lambda(f(x) - f(x')) \geq \nabla f(x') \cdot (x - x').$$

Because γ_λ is a monotone increasing function, we have

$$f(x) - f(x') \geq (\gamma_\lambda)^{-1}(\nabla f(x') \cdot (x - x')) = \kappa_\lambda(\nabla f(x') \cdot (x - x'))$$

as long as $1 + \lambda\nabla f(x') \cdot (x - x') > 0$ (so that the logarithm in κ_λ is well-defined). So we can define a quantity that is nonnegative and vanishes when $x = x'$; it is a divergence function using the language of information geometry (Amari, 2016).

Definition 4. (λ-logarithmic divergence). Let f be a λ-convex function. Let $x, x' \in \Omega$ be such that $1 + \lambda\nabla f(x') \cdot (x - x') > 0$. Then we define the λ-logarithmic divergence between x, x' by

$$\begin{aligned} \mathbf{L}_{\lambda,f}(x,x') &= f(x) - f(x') - \kappa_\lambda(\nabla f(x') \cdot (x - x')) \\ &= f(x) - f(x') - \frac{1}{\lambda}\log(1 + \lambda\nabla f(x') \cdot (x - x')). \end{aligned} \tag{29}$$

Below we will give conditions under which $1 + \lambda\nabla f(x') \cdot (x - x') > 0$.

3.2 λ-Deformed Legendre duality

Legendre duality has been generalized to c-duality which is used extensively in the theory of optimal transport (Villani, 2003, 2008). In our context, the bilinear pairing $x \cdot y$ in (26) is replaced by $c(x \cdot y)$, where c is monotone but non-linear in general. In particular, we take $c(t) = \kappa_\lambda(t)$. By Theorem 1, the requirement of convexity of $f(x)$ is replaced by the requirement of convexity of $g(x) = \gamma_\lambda(f(x))$ on a convex set $\Omega \ni x$, and the λ-conjugate variable is given by the λ-deformed Legendre transformation. Following Wong and Zhang (2021), we made the following definition.

Definition 5. (λ-conjugate function). Fix $\lambda \in \mathbb{R}, \lambda \neq 0$. Given a function $f : \Omega \to \mathbb{R}$ on some $\Omega \subset \mathbb{R}^d$, we define $f^{(\lambda)}$ on $\Omega' \subset \mathbb{R}^d$ by

$$f^{(\lambda)}(u) = \sup_{x\in\Omega}(\kappa_\lambda(x \cdot u) - f(x)), \quad u \in \mathbb{R}^d. \tag{30}$$

We call $f^{(\lambda)}$ the λ-conjugate of f.

The following result connects the notion of λ-convexity and the deformed duality under the function κ_λ.

Theorem 1. (λ-duality). *Let $\lambda \neq 0$ and let f be a smooth and λ-convex function on some open convex set $\Omega \subset \mathbb{R}^d$. Further assume $1 - \lambda\nabla f(x) \cdot x > 0$ on Ω.*[c]

(i) Consider $f^{(\lambda)}$ as a function on some other open convex set Ω'. Then the operation of λ-conjugation is involutive on Ω: $(f^{(\lambda)})^{(\lambda)} = f$;

(ii) For $x \in \Omega$, define the λ-gradient ∇^λ by

$$\nabla^\lambda f(x) := \frac{1}{1 - \lambda\nabla f(x) \cdot x}\nabla f(x). \tag{31}$$

[c]For the geometric meaning of the condition $1 - \lambda\nabla f(x) \cdot x > 0$ see Wong (2018, Section 3.3); by Theorem 6 it always holds for the potential function of the λ-exponential family.

Then $u = \nabla^\lambda f(x)$ is a C^1 diffeomorphism from Ω onto its range Ω' which is an open set;

(iii) For $x \in \Omega$ and $u = \nabla^\lambda f(x)$ we have $1 + \lambda x \cdot u > 0$ and the following identity holds

$$f(x) + f^{(\lambda)}(u) \equiv \frac{1}{\lambda} \log(1 + \lambda x \cdot u) = \kappa_\lambda(x \cdot u). \tag{32}$$

In reminiscent of Legendre duality, we call the above set of relations the λ-deformed Legendre duality, or λ-duality in short.

Proof. The proof of this theorem is given in scattered parts in Wong (2018, Section 3.3). Here we just rephrased the results in terms of the λ-duality. □

3.3 Relationship between λ-conjugation and Legendre conjugation

Clearly, when $\lambda \to 0$, then the λ-duality reduces to the regular Legendre duality, so that formally $f^{(\lambda)} \to f^*$. However, for a fixed $\lambda \neq 0$, how are the two dualities related? More specifically, because a smooth convex function f is automatically λ-convex for $\lambda > 0$, how is its Legendre dual f^* related to $f^{(\lambda)}$, and how is the conjugate variable ∇f related to $\nabla^\lambda f$?

Henceforth we fix a λ-convex function f. Assume $\lambda > 0$ and that $1 + \lambda x \cdot u > 0$ for all $x \in \Omega$, $u \in \Omega'$. Then we have the following identities:

$$\begin{aligned}(1 + \lambda x \cdot u)e^{-\lambda f(x)} &= e^{-\lambda f(x)} + \lambda\, e^{-\lambda f(x)} x \cdot u \\ &= (1 - \lambda\, g(\tilde{x})) + \lambda\, \tilde{x} \cdot u \\ &= 1 + \lambda\, (\tilde{x} \cdot u - g(\tilde{x})),\end{aligned}$$

as long as we set (going from the first to the second line)

$$\tilde{x} = x\, e^{-\lambda\, f(x)}, \tag{33}$$

$$g(\tilde{x}) = \frac{1}{-\lambda}\left(e^{-\lambda f(x)} - 1\right) = \gamma_{-\lambda}(f(x)). \tag{34}$$

Therefore, for $u \in \Omega'$ we have

$$\begin{aligned}f^{(\lambda)}(u) &= \sup_{x \in \Omega} \frac{1}{\lambda}(\log(1 + \lambda(x \cdot u))) - f(x) \\ &= \sup_{\tilde{x} \in \Omega'} \frac{1}{\lambda}(\log(1 + \lambda\, (\tilde{x} \cdot u - g(\tilde{x})))) \\ &= \frac{1}{\lambda} \log\left(1 + \lambda \sup_{\tilde{x} \in \Omega} (\tilde{x} \cdot u - g(\tilde{x}))\right) \\ &= \frac{1}{\lambda} \log(1 + \lambda g^*(u)) \\ &= \kappa_\lambda(g^*(u)),\end{aligned}$$

or

$$g^*(u) = \gamma_\lambda(f^{(\lambda)}(u)),$$

where g^* is the (regular) Legendre conjugate of the function $g(x) \equiv \gamma_\lambda(f(x))$. So, the functions γ_λ and $\gamma_{-\lambda}$ serve as link functions from the $(f, f^{(\lambda)})$-pair of the λ-deformed Legendre conjugation to the (g, g^*)-pair of the (regular) Legendre conjugation:

$$g(\widetilde{x}) = \gamma_{-\lambda}(f(x)), \quad g^*(u) = \gamma_\lambda(f^{(\lambda)}(u)).$$

What about the conjugate variables of the λ-conjugation and of the Legendre conjugation? The next proposition says that $u = \nabla_x^\lambda f(x)$ which is dual to x with respect to f under λ-conjugation, and $u = \nabla_{\tilde{x}} g(\widetilde{x})$ dual to $\widetilde{x}$ with respect to g under Legendre-conjugation, are in fact equal! This justifies our using the same symbol u.

Proposition 4. *We have*

$$\nabla_x^\lambda f(x) = \nabla_{\tilde{x}} g(\widetilde{x}). \tag{35}$$

Proof. We will use matrix notations where the gradient is regarded as a column vector. Applying the multivariate chain rule to Eq. (34), we have[d]

$$(\nabla_{\tilde{x}} g(\widetilde{x}))^\top = e^{-\lambda f(x)} (\nabla_x f(x))^\top \frac{\partial x}{\partial \widetilde{x}}(\widetilde{x}),$$

where $\frac{\partial x}{\partial \widetilde{x}}(\widetilde{x})$ is the Jacobian of the transformation $\widetilde{x} \mapsto x$ and $(\cdot)^\top$ denotes the transpose.
From (33), we have

$$\frac{\partial \widetilde{x}}{\partial x}(x) = e^{-\lambda f(x)} \left(\mathrm{I} - \lambda\, x\, (\nabla_x f(x))^\top \right).$$

Since $1 - \lambda \nabla_x f(x) \cdot x > 0$ by assumption, by the Shermon–Morrison formula (see e.g., Wong, 2018) we can invert the Jacobian to get

$$\frac{\partial x}{\partial \widetilde{x}}(\widetilde{x}) = e^{\lambda f(x)} \left(\mathrm{I} + \frac{\lambda\, x\, (\nabla_x f(x))^\top}{1 - \lambda \nabla_x f(x) \cdot x} \right).$$

Substituting this into the above and using Eq. (31), we have $\nabla_x^\lambda f(x) = \nabla_{\tilde{x}} g(\widetilde{x})$ and the proposition is proved. □

---

[d]For two vectors x and y, we denote their outer product by $x\, y^\top$, which is a rank-1 square matrix with the (i, j)-entry $x^i y^j$.

Let $u = \nabla^\lambda f(x)$ denote the λ-conjugate parameter corresponding to x with respect to $f(x)$, and let $\tilde{u} = \nabla_{\tilde{x}} g(\widetilde{x})$ be the Legendre-conjugate parameter corresponding to $\widetilde{x}$ with respect to $g(\widetilde{x})$. Then Proposition 4 says $u(x) = \tilde{u}(\widetilde{x})$, where $\widetilde{x}$ and x are linked by (20). Therefore,

Theorem 2. (Connecting λ-duality to Legendre duality). *For $u = \nabla^\lambda f(x) = \nabla_{\tilde{x}} g(\widetilde{x})$, we have the following identity as expression of λ-duality:*

$$\kappa_\lambda(x \cdot u) = f(x) + f^{(\lambda)}(u), \tag{36}$$

and the following identity as expression of the (regular) Legendre duality:

$$\widetilde{x} \cdot u = g(\widetilde{x}) + g^*(u). \tag{37}$$

where $\widetilde{x} = x e^{-\lambda f(x)}$ as in (33).

Remark 4. The equivalence of Eqs. (37) and (36) can be seen by re-writing the former as

$$e^{-\lambda f(x)} x \cdot u = \gamma_{-\lambda}(f(x)) + \gamma_\lambda(f^{(\lambda)}(u)) = \frac{1}{\lambda}\left(e^{\lambda f^{(\lambda)}(u)} - e^{-\lambda f(x)}\right).$$

Multiplying $e^{\lambda f(x)}$ on both sides, we obtain

$$x \cdot u = \frac{1}{\lambda}\left(e^{\lambda(f^{(\lambda)}(u) + f(x))} - 1\right) = \gamma_\lambda(f^{(\lambda)}(u) + f(x)).$$

Noting $(\gamma_\lambda)^{-1} = \kappa_\lambda$ verifies (36).

Next we connect the λ-duality with the λ-logarithmic divergence. We begin with a lemma.

Lemma 1. *We have, for $u = \nabla_x^\lambda f(x)$, $x = \nabla_u^\lambda f^{(\lambda)}(u)$, $\hat{u} = \nabla_x f(x)$, $\hat{x} = \nabla_u f^\lambda(u)$ and arbitrary y, v (such that the logarithms are well-defined), the identities*

$$\kappa_\lambda(u \cdot y) - \kappa_\lambda(u \cdot x) = \kappa_\lambda(\hat{u} \cdot (y - x)), \tag{38}$$

$$\kappa_\lambda(v \cdot x) - \kappa_\lambda(u \cdot x) = \kappa_\lambda((v - u) \cdot \hat{x}). \tag{39}$$

Proof. From (31), we have

$$u \cdot x = \frac{\nabla f(x) \cdot x}{1 - \lambda\, \nabla f(x) \cdot x},$$

so that

$$1 + \lambda\, u \cdot x = \frac{1}{1 - \lambda\, \nabla f(x) \cdot x}$$

and

$$1 + \lambda\, u \cdot y = \frac{1 + \lambda\, \nabla f(x) \cdot (y - x)}{1 - \lambda\, \nabla f(x) \cdot x} = (1 + \lambda u \cdot x)(1 + \lambda\, \nabla f(x) \cdot (y - x)).$$

Taking logarithm and rearranging, we obtain Eq. (38).

On the other hand, because

$$x = \nabla_u^\lambda f^{(\lambda)}(u) = \frac{\nabla f^{(\lambda)}(u)}{1 - \lambda \nabla f^{(\lambda)}(u) \cdot u},$$

we also have

$$1 + \lambda x \cdot u = \frac{1}{1 - \lambda \nabla f^{(\lambda)}(u) \cdot u}. \tag{40}$$

The proof of Eq. (38) is similar. □

In this lemma above, y and v are arbitrary; it is interesting that some form of linearity holds even though κ_λ is itself nonlinear. As a consequence, we can obtain:

Theorem 3. *The λ-logarithmic divergence satisfies reference-representation biduality (see Zhang, 2005, 2013), namely*

$$\mathbf{L}_{\lambda, f^{(\lambda)}}(v, u) = \mathbf{L}_{\lambda, f}(x, y),$$

where $u = \nabla^\lambda f(x) \longleftrightarrow x = \nabla^\lambda f^{(\lambda)}(u)$ *and* $v = \nabla^\lambda f(y) \longleftrightarrow y = \nabla^\lambda f^{(\lambda)}(v)$.

Proof. We have

$$\begin{aligned}
\mathbf{L}_{\lambda, f^{(\lambda)}}(v, u) &= f^{(\lambda)}(v) - f^{(\lambda)}(u) - \kappa_\lambda\Big((\nabla_u f^{(\lambda)}(u)) \cdot (v - u)\Big) \\
&= f^{(\lambda)}(v) - f^{(\lambda)}(u) - (\kappa_\lambda(x \cdot v) - \kappa_\lambda(x \cdot u)) \\
&= (\kappa_\lambda(y \cdot v) - f(y)) - \kappa_\lambda(x \cdot v) + f(x) \\
&= f(x) - f(y) - (\kappa_\lambda(x \cdot v) - \kappa_\lambda(y \cdot v)) \\
&= f(x) - f(y) - \kappa_\lambda\big((x - y) \cdot \nabla_y f(y)\big) \\
&= \mathbf{L}_{\lambda, f}(x, y).
\end{aligned}$$

□

3.4 Information geometry of λ-logarithmic divergence

Recall that we have defined $g_\lambda(x) = \gamma_\lambda(f(x)) = \frac{1}{\lambda}(e^{\lambda f(x)} - 1)$, and used (ordinary) convexity of g_λ to define λ-convexity of f. The Hessian of g is given by

$$\begin{aligned} \mathbf{G}(x) &\equiv \text{Hess}\, g_\lambda(x) \\ &= e^{\lambda f(x)} \left(\text{Hess} f(x) + \lambda (\nabla f(x))(\nabla f(x))^\top \right) \\ &= e^{\lambda f(x)}\, \mathbf{g}^{(\lambda)}(x), \end{aligned} \tag{41}$$

where

$$\mathbf{g}^{(\lambda)}(x) = \text{Hess} f(x) + \lambda\, (\nabla f(x))(\nabla f(x))^\top. \tag{42}$$

Because $e^{\lambda f(x)} > 0$, we can see that $\mathbf{g}^{(\lambda)}$ is positive definite if and only if $\mathbf{G}$ is positive definite. Furthermore, $\mathbf{g}^{(\lambda)}$ is a rank-one correction to a Hessian metric Hess$f$; so $\mathbf{g}^{(\lambda)}$ is positive-definite for $\lambda > 0$ whenever Hess$f$ is positive-definite. Stated in another way, when Hess$f$ is positive-definite, then both $\mathbf{g}^{(\lambda)}$ and $\mathbf{G}$ are positive-definite.

The value of λ regulates the "degree" of convexity. From Eq. (42) we see that if a function f is λ_1-convex, then it is λ_2-convex for all $\lambda_2 \geq \lambda_1$. Since 0-convexity is just the regular convexity, we have that

(i) if f is λ-convex for a negative $\lambda < 0$, then f is convex;
(ii) if f is convex, then f is λ-convex for all $\lambda > 0$.

Information geometry (see e.g., Amari, 2016, Chapter 6) tells us that taking the mixed second derivative of the divergence function yields a Riemannian metric. Indeed, we have

Proposition 5. *The Riemannian metric associated with the λ-logarithmic divergence is* $\mathbf{g}^{(\lambda)}$

$$-(\nabla_x)(\nabla_{x'})^\top \mathbf{L}_{\lambda,f}(x, x')\big|_{x=x'} = \mathbf{g}^{(\lambda)}(x).$$

Proof. Direct differentiation. □

So $\mathbf{g}^{(\lambda)}(x)$ defined by Eq. (42) serves as a metric tensor on the manifold with coordinates $x \in \Omega$:

$$ds^2 = (dx)^\top\, \mathbf{g}^{(\lambda)}(x)\ \ dx = e^{-\lambda f(x)} (dx)^\top\, \mathbf{G}(x)\ \ dx.$$

We see from Eq. (42) that $\mathbf{g}^{(\lambda)}$ is a conformal Hessian metric, namely, conformal (by a factor $e^{\lambda f}$) to the Hessian $\mathbf{G}$, which is itself a Riemannian metric for dually flat spaces. It was shown in Wong (2018) and Wong and Yang (2019) that this conformal Hessian geometry comes with a dual pair of projectively flat connections, and is a space of constant (sectional) curvature with these connections (with λ encoding the curvature value of a constant-curvature space) in the information-geometric sense. See Wong and Yang (2021) for additional results related to the geometry of optimal transport.

Remarkably, the λ-logarithmic divergence satisfies a generalized Pythagorean theorem.

Theorem 4. (Generalized Pythagorean theorem). *Let x_P, x_Q, and x_R be three points such that the (unparameterized) primal geodesic from x_Q to x_R and the (unparameterized) dual geodesic from x_Q to x_P intersect at x_Q orthogonally (with respect to a Riemannian metric $\mathbf{g}^{(\lambda)}$). Then*

$$\mathbf{L}_{\lambda,f}(x_Q, x_P) + \mathbf{L}_{\lambda,f}(x_R, x_Q) = \mathbf{L}_{\lambda,f}(x_R, x_P) \tag{43}$$

In the above statement, by a (unparameterized) primal (resp. dual) geodesic we mean the straight line in the x (resp. v) coordinates with respect to the λ-duality induced by f.

Proof. See (Wong, 2018, theorem 16). □

4 λ-Deformed entropy and divergence

4.1 Relation between potential functions and Rényi entropy

We now show that above λ-duality theory is naturally compatible with the λ-exponential and λ-mixture families. In what follows we assume $\lambda < 1$.

Proposition 6. (λ-convexity of potential functions).

(i) The divisive potential function $\varphi_\lambda(x)$ of the λ-exponential family, defined by Eq. (16)*, is λ-convex. Moreover,* $1 - \lambda\varphi_\lambda(x) \cdot x > 0$.

(ii) The potential function $\psi_\lambda(\eta)$ for the λ-mixture family, defined by

$$\psi_\lambda(\eta) = \log \int \left(\sum_{i=0}^{d} \eta_i \widetilde{P}_i \right)^{1/(1-\lambda)} d\mu = \log Z_\lambda(\eta), \tag{44}$$

is λ-convex.

Proof. See Wong and Zhang (2021). □

The following theorem, first proved in Wong (2018) (formulated in terms of $\mathcal{F}^{(\pm\alpha)}$-families), 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 Kullback–Leibler divergence (Bregman divergence of the potential function) is replaced by the Rényi entropy.

Theorem 5. (Link to Rényi entropy).

(i) For the λ-exponential family, the λ-conjugate of the divisive potential function $\varphi_\lambda(\vartheta)$, where η is the λ-conjugate variable $\eta = \nabla^\lambda \varphi_\lambda$, is denoted as ψ_λ and given by

$$\psi_\lambda(\eta) = -\boldsymbol{H}_\lambda^{\text{Rényi}}[p(\cdot\,|\vartheta)]. \tag{45}$$

(ii) For the λ-mixture family, we have

$$\psi_\lambda(\eta) = -\boldsymbol{H}_\lambda^{\text{Rényi}}[p(\cdot\,|\eta)]. \tag{46}$$

Remark 5. By (14), the dual variable $\eta = \nabla\phi_\lambda(\theta) = \nabla^\lambda\varphi_\lambda(\vartheta)$ of the λ-exponential family is the *escort expectation*:

$$\eta = \frac{\int (p(\zeta|\theta))^{1-\lambda}\, F(\zeta)\, d\mu}{\int (p(\zeta|\theta))^{1-\lambda}\, d\mu} = \int \widetilde{p}(\zeta|\theta)\, F(\zeta)\, d\mu. \tag{47}$$

4.2 Relation between λ-logarithmic divergence and Rényi divergence

Recall that the Rényi divergence (with Rényi index $q = 1 - \lambda$) is defined by

$$\boldsymbol{H}_\lambda^{\text{Rényi}}[p||p'] = \frac{-1}{\lambda} \log \int (p(\zeta))^{1-\lambda}(p'(\zeta))^\lambda d\mu. \tag{48}$$

Note that here we use the subscript λ rather than q.

Theorem 6. *The λ-logarithmic divergence of the divisive λ-potential is the Rényi divergence:*

(i) For a λ-exponential family, using φ_λ as the potential, we have

$$\mathbf{L}_{\lambda,\varphi_\lambda}(\vartheta,\vartheta') = \boldsymbol{H}_\lambda^{\text{Rényi}}[p(\cdot\,|\vartheta')||p(\cdot\,|\vartheta)]. \tag{49}$$

(ii) For a λ-mixture family, using ψ_λ as the potential, we have

$$\mathbf{L}_{\lambda,\psi_\lambda}(\eta,\eta') = \boldsymbol{H}_\lambda^{\text{Rényi}}[p(\cdot\,|\eta)||p(\cdot\,|\eta')]. \tag{50}$$

Proof. We give the proof of (ii). Let η, η' be given. By a direct computation, we have

$$\begin{aligned} 1+\lambda\nabla\psi(\eta')\cdot(\eta-\eta') &= \frac{\int \left(\sum_i \eta_i'\widetilde{P}_i\right)^{\lambda/(\lambda-1)}\left(\sum_i \eta_i\widetilde{P}_i\right) d\mu}{\int \left(\sum_i \eta_i'\widetilde{P}_i\right)^{1/(1-\lambda)} d\mu} \\ &= \frac{(Z(\eta))^{1-\lambda}}{(Z(\eta'))^{1-\lambda}} \int (p(\zeta|\eta))^{1-\lambda}(p(\zeta|\eta'))^\lambda d\mu. \end{aligned}$$

It follows that

$$\psi(\eta) - \psi(\eta') - \frac{1}{\lambda}\log\left(1 + \lambda\nabla\psi(\eta')\cdot(\eta - \eta')\right)$$
$$= \frac{1}{\lambda}\log\int p(\zeta|\eta)^{1-\lambda}p(\zeta|\eta')^{\lambda}d\mu$$
$$= \boldsymbol{H}_{\lambda}^{\text{Rényi}}[p(\cdot\,|\eta)||p(\cdot\,|\eta')] \geq 0.$$

□

4.3 Entropy maximizing property of λ-exponential family

In Banerjee et al. (2005), it is proved that for a regular exponential family, the log likelihood of an exponential family can be associated to a Bregman divergence:

$$\log p(\zeta|\theta) = -\mathbf{B}_{\psi}(y,\eta) + \psi(y). \tag{51}$$

Note here $\psi = \phi^*$ and the parameter appearing on the right-hand side is not θ but in the form of the dual parameter $\eta = \nabla\phi(\theta) = \mathbb{E}_{\theta}[F(\zeta)]$, and $y = F(\zeta)$ is a transformation of sample value ζ.

Here, we show that the λ-exponential family satisfies an analogous property.

Theorem 7. *Under suitable conditions, we have*

$$\log p(\zeta|\vartheta) = -\mathbf{L}_{\lambda,\psi_{\lambda}}(y,\eta) + \psi_{\lambda}(y), \quad y = F(\zeta). \tag{52}$$

Proof. We note

$$\log p(\zeta|\vartheta) = \frac{1}{\lambda}\log\left(1 + \lambda\vartheta\cdot y\right) - \varphi_{\lambda}(\vartheta). \tag{53}$$

The right-hand side is

$$\frac{1}{\lambda}\log\left(1 + \lambda\vartheta\cdot y\right) - \frac{1}{\lambda}\log\left(1 + \lambda\vartheta\cdot\eta\right) + \psi_{\lambda}(\eta)$$
$$= \frac{1}{\lambda}\log\left(1 + \lambda\nabla\psi_{\lambda}(\eta)\cdot(y - \eta)\right) + \psi_{\lambda}(\eta)$$
$$= -\mathbf{L}_{\lambda,\psi_{\lambda}}(y,\eta) + \psi_{\lambda}(y),$$

where the last step implicitly assumes that y belongs to the domain of ψ_{λ} so that the divergence $\mathbf{L}_{\lambda,\psi_{\lambda}}(y,\eta)$ is well-defined. □

Let $\lambda < 1$ and let a reference measure μ be given. Let $F = (F_1, \ldots, F_d)$ be a vector of statistics. Consider the Rényi entropy maximization problem

$$\max_{p\sim\mu} \boldsymbol{H}_{\lambda}^{\text{Rényi}}[p] \text{ subject to } \widetilde{\mathbb{E}}_{p}[F(\zeta)] = c, \tag{54}$$

where $\widetilde{\mathbb{E}}_p$ is the expectation with respect to the escort distribution (with exponent $1-\lambda$) and $\sim$ means equivalence (mutually absolutely continuous). It is well known that the solution to Eq. (54) can be written in the form of a q-exponential family; see the q-Max-Ent theorem in Amari and Ohara (2011). An alternative proof using the λ-exponential family and the associated λ-logarithmic divergence was given in Wong and Zhang (2021)

Proposition 7. *Consider the λ-exponential family (23) and suppose for some parameter ϑ_0 we have $\widetilde{\mathbb{E}}_{\vartheta_0}[F(\zeta)] = c$. Then the distribution $p(\zeta|\vartheta_0)$ is the unique solution to Eq. (54).*

Note that the Rényi entropy is monotonically linked to the Tsallis entropy: a straightforward computation gives

$$\boldsymbol{H}_q^{\text{Tsallis}}[p] = \gamma_\lambda\left(\boldsymbol{H}_\lambda^{\text{Rényi}}[p]\right) \Leftrightarrow \boldsymbol{H}_\lambda^{\text{Rényi}}[p] = \kappa_\lambda\left(\boldsymbol{H}_q^{\text{Tsallis}}[p]\right).$$

So maximizing the Tsallis entropy is equivalent to maximizing the Rényi entropy, and the solution to the problem (54) remains the same if we use instead the Tsallis entropy. Some arguments that favor the Rényi entropy as a physical concept are given in (Naudts, 2011, Section 9.3); for example, the Rényi entropy satisfies the additive property but the Tsallis entropy does not.[e]

5 Example: λ-Deformation of the probability simplex

To illustrate our formulation in a concrete setting, we now consider the probability simplex Δ^d regarded as the set of positive discrete probability measures on the finite set $\mathcal{X} = \{0, 1, \ldots, d\}$ which is the state space of the categorical random variable ζ:

$$p(\zeta|\vartheta) = \begin{cases} p(i|\vartheta) = u_i, & \text{if } \zeta = i, \quad i \in \{0, 1, \ldots, d\}; \\ 0, & \text{otherwise.} \end{cases} \tag{55}$$

Further examples can be found in Wong and Zhang (2021).

5.1 λ-Exponential representation

For any $\lambda \neq 0$, Δ^d can be expressed as a λ-exponential family, with parameter $\vartheta = (\vartheta^1, \ldots, \vartheta^d)$. To see this, let $F_i(\zeta) = \delta_i(\zeta)$ be the indicator of $i = 1, \ldots, d$ where the probability mass has support on. Then we may write

[e]See Van Erven and Harremos (2014) for a useful overview of the properties of the Rényi entropy and divergence. In particular, we mention that the Rényi divergence is *additive*: Given two product measures $p_1 \otimes p_2$ and $p_1' \otimes p_2'$, we have

$$\boldsymbol{H}_q^{\text{Rényi}}[p_1 \otimes p_1' || p_2 \otimes p_2'] = \boldsymbol{H}_q^{\text{Rényi}}[p_1 || p_2] + \boldsymbol{H}_q^{\text{Rényi}}[p_1' || p_2'].$$

$$p(\zeta|\vartheta) = \left(1 + \lambda \sum_{j=1}^{d} \vartheta^j \delta_j(\zeta)\right)^{1/\lambda} e^{-\varphi_\lambda(\vartheta)}, \tag{56}$$

so

$$u_i = p(i|\vartheta) = (1 + \lambda\vartheta^i)^{1/\lambda} e^{-\varphi_\lambda(\vartheta)},$$

and $u_0 = 1 - \sum_{i=1}^{d} u_i$. The potential function φ_λ, given by

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

is λ-convex for $\lambda < 1$. We then obtain

$$\vartheta^i = \frac{1}{\lambda}\left[\left(\frac{u_i}{u_0}\right)^{\lambda} - 1\right] \tag{57}$$

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 $\theta^i = \log(u_i/u_0)$. By Proposition 4, the dual variable $\eta = \nabla^{\lambda}\varphi_\lambda(\vartheta)$ gives the escort distribution:

$$\eta_i = \frac{(p(i|\vartheta))^{1-\lambda}}{\sum_{j=0}^{d}(p(j|\vartheta))^{1-\lambda}} = \frac{(u_i)^{1-\lambda}}{\sum_{j=0}^{d}(u_j)^{1-\lambda}}, \quad i = 1, \ldots, d. \tag{58}$$

And we can define $\eta_0 = 1 - \sum_{i=1}^{d} \eta_i$.

5.2 λ-Mixture representation

Dually, for $i \in \mathcal{X}$, let $P_i(\zeta) = \delta_i(\zeta)$ be the density of the point mass at i and consider the corresponding λ-mixture family $p(\cdot|\eta)$ given by (25) (not to be confused with $p(\cdot|\vartheta)$):

$$p(\zeta|\eta) = \frac{1}{Z_\lambda(\eta)} \int \left(\sum_{j=0}^{d} \eta_j \delta_j(\zeta)\right)^{1/(1-\lambda)} d\mu.$$

That is, $p(\zeta|\eta) = 0$ for $\zeta \neq i \in \{0, 1, \cdots, d\}$ and

$$p(i|\eta) = \frac{(\eta_i)^{1/(1-\lambda)}}{\sum_{j=0}^{d}(\eta_j)^{1/(1-\lambda)}}, \quad i = 0, 1, \cdots, d. \tag{59}$$

Note that if η is given by Eq. (58), then we have $p(\cdot|\vartheta) = p(\cdot|\eta)$. Thus the probability simplex is both a λ-exponential family and a λ-mixture family.

Consider the mixture variable η. Since the set $\{\eta \in (0,1)^{1+d} : \eta_0 + \cdots + \eta_d = 1\}$ is not open, we consider instead the open domain

$$\Omega' = \{\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_d) \in (0,1)^d : \hat{\eta}_1 + \cdots + \hat{\eta}_d < 1\},$$

where the notation $\hat{\eta}$ signifies that the 0-th coordinate is dropped. Consider the negative Rényi entropy ψ, given by Eq. (46), as a function of $\hat{\eta}$. Note that

$$\frac{\partial \psi}{\partial \hat{\eta}_i} = \frac{\partial \psi}{\partial \eta_i} - \frac{\partial \psi}{\partial \eta_0}, \quad i = 1, \ldots, d. \tag{60}$$

Since $\frac{1}{\lambda}(e^{\lambda\psi(\hat{\eta})} - 1)$ is convex in $\hat{\eta}$ (which is a linear transformation of η), so we may consider the λ-duality. The following result completes the circle of idea (see Wong and Zhang, 2021 for a more general result).

Proposition 8. *With respect to the probability simplex Δ^d, the primal variable ϑ in (57) when Δ^d is considered as a λ-exponential family with the form (56) equals the conjugate variable given by $\hat{\vartheta} = \nabla^{\lambda}_{\hat{\eta}}\psi(\hat{\eta})$ when Δ^d is considered as a λ-mixture family with mixture variable η:*

$$\vartheta^i = \hat{\vartheta}^i, \quad i = 1, \ldots, d. \tag{61}$$

Proof. Given $\hat{\eta} \in \Omega'$, consider η where $\eta_i = \hat{\eta}_i$ for $1 \le i \le d$ and $\eta_0 = 1 - \sum_i \hat{\eta}_i$. Using (9) and (60), we have

$$\psi(\hat{\eta}) = \frac{1-\lambda}{\lambda} \log\left(\sum_{j=0}^{d} \eta_j^{\frac{1}{1-\lambda}}\right),$$

and

$$\frac{\partial \psi}{\partial \hat{\eta}_i} = \frac{1}{\lambda} \frac{1}{\sum_{j=0}^{d} \eta_j^{1/(1-\lambda)}} \left(\eta_j^{\frac{\lambda}{1-\lambda}} - \eta_0^{\frac{\lambda}{1-\lambda}}\right).$$

So

$$1 - \lambda \sum_{i=1}^{d} \frac{\partial \psi}{\partial \hat{\eta}_i} \hat{\eta}_i = \frac{\eta_0^{1/(1-\lambda)-1}}{\sum_{j=0}^{d} \eta_j^{1/(1-\lambda)}}.$$

Computing the λ-gradient (31), we have, for any $i = 1, \ldots, d$,

$$\hat{\vartheta}^i = \frac{1}{\lambda}\left[\left(\frac{u_i}{u_0}\right)^{\lambda} - 1\right].$$

Hence we obtain (61). □

6 Summary and conclusion

This chapter summarizes a novel approach for studying subtractive and divisive normalizations in deformed exponential family models. A typical example of the former is the q-exponential family with associated Tsallis entropy whereas an example of the latter is the $\mathcal{F}^{(\pm\alpha)}$-family and the associated Rényi entropy. Our first conclusion is that these two versions of deformation to an exponential family are two faces of the same coin: under a reparameterization, they are one and the same. Nevertheless, using different dualities lead to genuinely different mathematical structures.

We introduce the λ-exponential family which reparameterizes (13) by

$$p^{(\lambda)}(\zeta|\cdot) = \exp_\lambda(\vartheta \cdot F(\zeta))e^{-\varphi_\lambda(\vartheta)} = \exp_\lambda(\theta \cdot F(\zeta) - \phi_\lambda(\theta)).$$

The λ-exponential family is also linked to λ-mixture family, when $\lambda \neq 0, 1$, via a reparameterization of the random functions $F(\zeta)$ above. A basic example is the finite simplex which is both a λ-exponential family and a λ-mixture family. Several other examples are considered in Wong and Zhang (2021).

The coincidence of these two parameterizations of the deformed family is associated with the λ-duality, which is the main focus of our exposition. The λ-duality is a "deformation" of the usual Legendre duality reviewed in Section 3.1. In a nutshell, instead of convex functions, we work with λ-convex functions f such that $\frac{1}{\lambda}(e^{\lambda f} - 1)$ is convex, for a fixed $\lambda \neq 0$. Also, instead of the convex conjugate, we use the λ-conjugate given by

$$f^{(\lambda)}(u) = \sup_x \left(\frac{1}{\lambda} \log(1 + \lambda x \cdot u) - f(x)\right).$$

Turning the above into an inequality leads to the nonnegative λ-logarithmic divergence associated to the λ-convexity (as generalization of the Bregman divergence associated to the regular convexity):

$$\mathbf{L}_{\lambda,f}(x, y) = f(x) - f(y) - \frac{1}{\lambda} \log(1 + \lambda \nabla f(y) \cdot (x - y)).$$

Coming back to the probability families, we first verified that the subtractive potential $\phi_\lambda(\theta)$ is convex in θ and the divisive potential $\varphi_\lambda(\vartheta)$ is λ-convex in ϑ. Subtractive normalization using $\phi_\lambda(\theta)$ is associated with the regular Legendre duality, whereas multiplicative normalization using $\varphi_\lambda(\vartheta)$ is associated with the λ-duality. This gives an interpretation of distinctiveness of Rényi entropy (used in the latter) from the Tsallis entropy (used in the former) based on their intimate connection to λ-duality (for $\lambda \neq 0$) or to the Legendre duality. As λ is the parameter that controls the curvature in the Riemannian geometry of these probability families (see Wong, 2018), our framework provides a simple parametric deformation from dually flat geometry (of the exponential model) to the dually projectively flat geometry (of the λ-exponential model). We expect that this framework will generate new insights in the applications of the q-exponential family and related concepts in statistical physics and information science.

References

Amari, S.-I., 2016. Information Geometry and Its Applications. Springer.

Amari, S.-I., Nagaoka, H., 2000. Methods of Information Geometry. American Mathematical Society.

Amari, S.-I., Ohara, A., 2011. Geometry of q-exponential family of probability distributions. Entropy 13 (6), 1170–1185.

Amari, S.-I., Ohara, A., Matsuzoe, H., 2012. Geometry of deformed exponential families: invariant, dually-flat and conformal geometries. Phys. A Stat. Mech. Appl. 391 (18), 4308–4319.

Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J., 2005. Clustering with Bregman divergences. J. Mach. Learn. Res. 6 (Oct), 1705–1749.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. B (Methodol.) 26 (2), 211–243.

de Andrade, L.H.F., Vieira, F.L.J., Cavalcante, C.C., 2021. On normalization functions and ϕ-families of probability distributions. In: Progress in Information Geometry: Theory and Applications, Springer Nature, p. 19.

Eguchi, S., 2006. Information geometry and statistical pattern recognition. Sugaku Expo. 19 (2), 197–216.

Kaniadakis, G., 2001. Non-linear kinetics underlying generalized statistics. Phys. A Stat. Mech. Appl. 296 (3-4), 405–425.

Matsuzoe, H., 2014. Hessian structures on deformed exponential families and their conformal structures. Differ. Geom. Appl. 35, 323–333.

Montrucchio, L., Pistone, G., 2017. Deformed exponential bundle: the linear growth case. In: International Conference on Geometric Science of Information, Springer, pp. 239–246.

Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S., 2004. Information geometry of U-boost and Bregman divergence. Neural Comput. 16 (7), 1437–1481.

Naudts, J., 2004. Estimators, escort probabilities, and ϕ-exponential families in statistical physics. J. Inequalities Pure Appl. Math. 5 (4), 102.

Naudts, J., 2008. Generalised exponential families and associated entropy functions. Entropy 10 (3), 131–149.

Naudts, J., 2011. Generalised Thermostatistics. Springer.

Naudts, J., Zhang, J., 2018. Rho-tau embedding and gauge freedom in information geometry. Inf. Geom. 1 (1), 79–115.

Newton, N.J., 2012. An infinite-dimensional statistical manifold modelled on Hilbert space. J. Funct. Anal. 263 (6), 1661–1681.

Ohara, A., Matsuzoe, H., Amari, S.-I., 2012. Conformal geometry of escort probability and its applications. Mod. Phys. Lett. B 26 (10), 1250063.

Pal, S., Wong, T.-K.L., 2016. The geometry of relative arbitrage. Math. Financ. Econ. 10 (3), 263–293.

Pal, S., Wong, T.-K.L., 2018. Exponentially concave functions and a new information geometry. Ann. Probab. 46 (2), 1070–1113.

Pal, S., Wong, T.-K.L., 2020. Multiplicative Schröodinger problem and the Dirichlet transport. Probab. Theory Relat. Fields 178 (1), 613–654.

Rényi, A., 1961. 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.

Tsallis, C., 1988. Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. 52 (1-2), 479–487.

Tsallis, C., 1994. What are the numbers that experiments provide. Quimica Nova 17 (6), 468–471.

Van Erven, T., Harremos, P., 2014. Rényi divergence and Kullback-Leibler divergence. IEEE Trans. Inf. Theory 60 (7), 3797–3820.

Villani, C., 2003. Topics in Optimal Transportation. American Mathematical Society.

Villani, C., 2008. Optimal Transport: Old and New. Springer.

Wong, T.-K.L., 2018. Logarithmic divergences from optimal transport and Rényi geometry. Inf. Geom. 1 (1), 39–78.

Wong, T.-K.L., 2019. Information geometry in portfolio theory. In: Geometric Structures of Information, Springer, pp. 105–136.

Wong, T.-K.L., Yang, J., 2019. Logarithmic divergence: geometry and interpretation of curvature. In: International Conference on Geometric Science of Information. (pp. 413–422) Springer, Cham.

Wong, T.-K.L., Yang, J., 2021. Pseudo-Riemannian geometry encodes information geometry in optimal transport. Information Geometry. Inf. Geom. (in press).

Wong, T.-K.L., Zhang, J., 2021. Tsallis and Rényi deformations linked via a new λ-duality. arXiv preprint arXiv:2107.11925.

Zhang, J., 2004. Divergence function, duality, and convex analysis. Neural Comput. 16 (1), 159–195.

Zhang, J., 2005. Referential duality and representational duality on statistical manifolds. In: Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan, vol. 1216, pp. 58–67.

Zhang, J., 2013. Nonparametric information geometry: from divergence function to referential-representational biduality on statistical manifolds. Entropy 15 (12), 5384–5418.

Zhang, J., 2015. On monotone embedding in information geometry. Entropy 17 (7), 4485–4499.