entropy

MDPI

Article

Divergences Induced by the Cumulant and Partition Functions of Exponential Families and Their Deformations Induced by Comparative Convexity

Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; frank.nielsen.x@gmail.com

Abstract: Exponential families are statistical models which are the workhorses in statistics, information theory, and machine learning, among others. An exponential family can either be normalized subtractively by its cumulant or free energy function, or equivalently normalized divisively by its partition function. Both the cumulant and partition functions are strictly convex and smooth functions inducing corresponding pairs of Bregman and Jensen divergences. It is well known that skewed Bhattacharyya distances between the probability densities of an exponential family amount to skewed Jensen divergences induced by the cumulant function between their corresponding natural parameters, and that in limit cases the sided Kullback–Leibler divergences amount to reverse-sided Bregman divergences. In this work, we first show that the α-divergences between non-normalized densities of an exponential family amount to scaled α-skewed Jensen divergences induced by the partition function. We then show how comparative convexity with respect to a pair of quasi-arithmetical means allows both convex functions and their arguments to be deformed, thereby defining dually flat spaces with corresponding divergences when ordinary convexity is preserved.

Keywords: convex duality; exponential family; Bregman divergence; Jensen divergence; Bhattacharyya distance; Rényi divergence; α-divergences; comparative convexity; log convexity; exponential convexity; quasi-arithmetic means; information geometry

Citation: Nielsen, F. Divergences Induced by the Cumulant and Partition Functions of Exponential Families and Their Deformations Induced by Comparative Convexity. *Entropy* **2024**, *26*, 193. https://doi.org/10.3390/e26030193

Academic Editor: Carlo Cafaro

Received: 20 December 2023
Revised: 21 February 2024
Accepted: 21 February 2024
Published: 23 February 2024

1. Introduction

In information geometry [1], any strictly convex and smooth function induces a dually flat space (DFS) with a canonical divergence which can be expressed in charts either as dual Bregman divergences [2] or equivalently as dual Fenchel–Young divergences [3]. For example, the cumulant function of an exponential family [4] (also called the free energy) generates a DFS, that is, an exponential family manifold [5] with the canonical divergence yielding the reverse Kullback–Leibler divergence. Another typical example of a strictly convex and smooth function generating a DFS is the negative entropy of a mixture family, that is, a mixture family manifold with the canonical divergence yielding the (forward) Kullback–Leibler divergence [3]. In addition, any strictly convex and smooth function induces a family of scaled skewed Jensen divergences [6,7], which in limit cases includes the sided forward and reverse Bregman divergences.

In Section 2, we present two equivalent approaches to normalizing an exponential family: first by its cumulant function, and second by its partition function. Because both the cumulant and partition functions are strictly convex and smooth, they induce corresponding families of scaled skewed Jensen divergences and Bregman divergences, with corresponding dually flat spaces and related statistical divergences.

In Section 3, we recall the well-known result that the statistical α-skewed Bhattacharyya distances between the *probability densities* of an exponential family amount to a scaled α-skewed Jensen divergence between their natural parameters. In Section 4, we prove that the α-divergences [8] between the *unnormalized densities* of a exponential family amount to

scaled α-skewed Jensen divergence between their natural parameters (Proposition 5). More generally, we explain in Section 5 how to deform a convex function using comparative convexity [9]: When the ordinary convexity of the deformed convex function is preserved, we obtain new skewed Jensen divergences and Bregman divergences with corresponding dually flat spaces. Finally, Section 6 concludes this work with a discussion.

2. Dual Subtractive and Divisive Normalizations of Exponential Families

2.1. Natural Exponential Families

Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space [10], where $\mathcal{X}$ denotes the sample set (e.g., finite alphabet, $\mathbb{N}$, $\mathbb{R}^d$, space of positive-definite matrices $\mathrm{Sym}_{++}(d)$, etc.), $\mathcal{A}$ a σ-algebra on $\mathcal{X}$ (e.g., power set $2^{\mathcal{X}}$, Borel σ-algebra $\mathcal{B}(\mathcal{X})$, etc.), and μ a positive measure (e.g., counting measure or Lebesgue measure) on the measurable space $(\mathcal{X}, \mathcal{A})$.

A natural exponential family [4,11] (commonly abbreviated as NEF [12]) is a set of probability distributions $\mathcal{P} = \{P_\theta \ : \ \theta \in \Theta\}$ all dominated by μ such that their Radon–Nikodym densities $p_\theta(x) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x)$ can be expressed canonically as

$$p_\theta(x) \propto \tilde{p}_\theta(x) = \exp\left(\sum_{i=1}^{m} \theta_i x_i\right), \tag{1}$$

where θ is called the natural parameter and $x = (x_1, \ldots, x_m)$ denotes the linear sufficient statistic vector [11]. The order of the NEF [13] is m. When the parameter θ ranges in the full natural parameter space

$$\Theta = \left\{\theta \ : \ \int_{\mathcal{X}} \tilde{p}_\theta(x)\, \mathrm{d}\mu(x) < \infty\right\} \subset \mathbb{R}^m,$$

the family is called full. The NEF is said to be regular when Θ is topologically open.

The unnormalized positive density $\tilde{p}_\theta(x)$ is indicated with a tilde notation and the corresponding normalized probability density is obtained as $p_\theta(x) = \frac{1}{Z(\theta)}\tilde{p}_\theta(x)$, where $Z(\theta) = \int \tilde{p}_\theta(x)\, \mathrm{d}\mu(x)$ is the Laplace transform of μ (the density normalizer). For example, the family of exponential distributions $\mathcal{E} = \{\lambda e^{-\lambda x} \ : \ \lambda > 0\}$ is an NEF with densities defined on the support $\mathcal{X} = \mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \ : \ x \geq 0\}$, natural parameter $\theta = -\lambda$ in $\Theta = \mathbb{R}_{<0} = \{\theta \in \mathbb{R} \ : \ \theta < 0\}$, sufficient linear statistic x, and normalizer $Z(\theta) = -\frac{1}{\theta}$.

2.2. Exponential Families

More generally, exponential families include many well known distributions after reparameterization [4] of their ordinary parameter λ by $\theta(\lambda)$. The general canonical form of the densities of an exponential family is

$$p_\lambda(x) \propto \tilde{p}_\lambda(x) = \exp(\langle \theta(\lambda), t(x) \rangle)\, h(x), \tag{2}$$

where $t(x) = (t_1(x), \ldots, t_m(x))$ are the sufficient statistic vector (such that $1, t_1(x), \ldots, t_m(x)$ are linearly independent), $h(x)$ is an auxiliary term used to define the base measure with respect to μ, and $\langle \cdot, \cdot \rangle$ is an inner product (e.g., scalar product of $\mathbb{R}^m$, trace product of symmetric matrices, etc.). By defining a new measure ν such that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) = h(x)$, we may consider without loss of generality the densities $\bar{p}_\lambda(x) = \frac{\mathrm{d}P_\lambda}{\mathrm{d}\nu}(x)$ with $h(x) = 1$.

For example, the Bernoulli distributions, Gaussian or normal distributions, Gamma and Beta distributions, Poisson distributions, Rayleigh distributions, and Weibull distributions with prescribed shape parameter are just a few examples of exponential families with the inner product on $\mathbb{R}^m$ defined as the scalar product. The categorical distributions (i.e., discrete distributions on a finite alphabet sample space) form an exponential family as well [1]. Zero-centered Gaussian distributions and Wishart distributions are examples of exponential families parameterized by positive-definite matrices with inner products defined by the matrix trace product, which is $\langle A, B \rangle = \mathrm{tr}(AB)$.

Exponential families abound in statistics and machine learning. Any two probability measures Q and R with densities q and r with respect to a dominating measure, say, $\mu = \frac{Q+R}{2}$, define an exponential family

$$\mathcal{P}_{Q,R} = \left\{ p_\lambda(x) \propto q^\lambda(x) r^{1-\lambda}(x) \ : \ \lambda \in (0,1) \right\},$$

which is called the likelihood ratio exponential family [14], as the sufficient statistic is $t(x) = \log \frac{q(x)}{r(x)}$ (with auxiliary carrier term $h(x) = r(x)$), or the Bhattacharyya arc, as the cumulant function of $\mathcal{P}_{Q,R}$ is expressed as the negative of the skewed Bhattacharyya distances [7,15].

In machine learning, undirected graphical models [16] and energy-based models [17], including Markov random fields [18] and conditional random fields, are exponential families [19]. Exponential families are universal approximators of smooth densities [20].

From a theoretical standpoint, it is often enough to consider (without loss of generality) natural exponential families with densities expressed as in Equation (1). However, here we consider generic exponential families with the densities expressed in Equation (2) in order to report common examples encountered in practice, such as the multivariate Gaussian family [21].

When the natural parameter space Θ is not full but rather parameterized by $\lambda = c(\lambda')$ for $\lambda' \in \Lambda'$ with $\dim(\Lambda') < m$ and a smooth function $c(u)$, the exponential family is called a curved exponential family [1]. For example, the special family of normal distributions $\{ p_{\mu,\sigma^2=\mu^2} \ : \ \mu \in \mathbb{R} \}$ is a curved exponential family with $u = \mu$ and $c(u) = (u, u^2)$ [1].

2.3. Normalizations of Exponential Families

Recall that $\tilde{p}_\theta(x) = \exp(\langle \theta, t(x) \rangle) \, h(x)$ denotes the unnormalized density expressed using the natural parameter $\theta = \theta(\lambda)$. We can normalize $\tilde{p}_\theta(x)$ using either the partition function $Z(\theta)$ or equivalently using the cumulant function $F(\theta)$, as follows:

$$p_\theta(x) = \frac{\exp(\langle \theta, t(x) \rangle)}{Z(\theta)} \, h(x), \tag{3}$$

$$= \exp(\langle \theta, t(x) \rangle - F(\theta) + k(x)), \tag{4}$$

where $h(x) = \exp(k(x))$, $Z(\theta) = \int \tilde{p}_\theta(x) \mathrm{d}\mu(x)$, and $F(\theta) = \log Z(\theta) = \log \int \tilde{p}_\theta(x) \mathrm{d}\mu(x)$. Thus, the logarithm and exponential functions allow conversion to and from the dual normalizers Z and F:

$$Z(\theta) = \exp(F(\theta)) \Leftrightarrow F(\theta) = \log Z(\theta).$$

We may view Equation (3) as an exponential tilting [13] of density $h(x)\mathrm{d}\mu(x)$.

In the context of λ-deformed exponential families [22] which generalize exponential families, the function $Z(\theta)$ is called the divisive normalization factor (Equation (3)) and the function $F(\theta)$ is called the subtractive normalization factor (Equation (4)). Notice that $F(\theta)$ is called the cumulant function because when $X \sim p_\theta(x)$ is a random variable following a probability distribution of an exponential family, the function $F(\theta)$ appears in the cumulant generating function of X: $K_X(t) = \log E_X[e^{\langle t, X \rangle}] = F(\theta + t) - F(\theta)$. In statistical physics, the cumulant function is called the log-normalizer or log-partition function. Because $Z > 0$ and $F = \log Z$, we can deduce that $F \geq Z$, as $\log x \leq x$ for $x > 0$.

It is well known that the cumulant function $F(\theta)$ is a strictly convex function and that the partition function $Z(\theta)$ is strictly log-convex [11].

Proposition 1 ([11]). *The natural parameter space Θ of an exponential family is convex.*

Proposition 2 ([11]). *The cumulant function $F(\theta)$ is strictly convex and the partition function $Z(\theta)$ is positive and strictly log-convex.*

It can be shown that the cumulant and partition functions are smooth C^∞ analytic functions [4]. A remarkable property is that strictly log-convex functions are also strictly convex.

Proposition 3 ([23], Section 3.5). *A strictly log-convex function $Z : \Theta \subset \mathbb{R}^m \to \mathbb{R}$ is strictly convex.*

The converse of Proposition 3 is not necessarily true, however; certain convex functions are not log-convex, and as such the class of strictly log-convex functions is a proper subclass of strictly convex functions. For example, θ^2 is convex but log-concave, as $(\log \theta^2)'' = -\frac{2}{\theta^2} < 0$ (Figure 1).

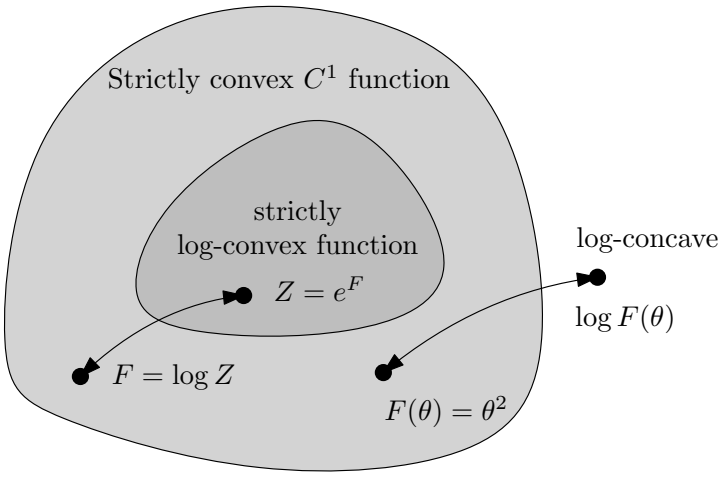

Figure 1. Strictly log-convex functions form a proper subset of strictly convex functions.

Remark 1. *Because $Z = \exp(F)$ is strictly convex (Proposition 3), F is exponentially convex.*

Definition 1. *The cumulant function F and partition function Z of a regular exponential family are both strictly convex and smooth functions inducing a pair of dually flat spaces with corresponding Bregman divergences* [2] *B_F (i.e., $B_{\log Z}$) and B_Z (i.e., $B_{\exp F}$):*

$$B_Z(\theta_1 : \theta_2) = Z(\theta_1) - Z(\theta_2) - \langle \theta_1 - \theta_2, \nabla Z(\theta_2) \rangle \geq 0, \quad (5)$$

$$B_{\log Z}(\theta_1 : \theta_2) = \log\left(\frac{Z(\theta_1)}{Z(\theta_2)}\right) - \left\langle \theta_1 - \theta_2, \frac{\nabla Z(\theta_2)}{Z(\theta_2)} \right\rangle \geq 0, \quad (6)$$

along with a pair of families of skewed Jensen divergences $J_{F,\alpha}$ and $J_{Z,\alpha}$:

$$J_{Z,\alpha}(\theta_1 : \theta_2) = \alpha Z(\theta_1) + (1-\alpha) Z(\theta_2) - Z(\alpha\theta_1 + (1-\alpha)\theta_2) \geq 0, \quad (7)$$

$$J_{\log Z,\alpha}(\theta_1 : \theta_2) = \log \frac{Z(\theta_1)^\alpha Z(\theta_2)^{1-\alpha}}{Z(\alpha\theta_1 + (1-\alpha)\theta_2)} \geq 0. \quad (8)$$

For a strictly convex function $F(\theta)$, we define the symmetric Jensen divergence as follows:

$$J_F(\theta_1, \theta_2) = J_{F,\frac{1}{2}}(\theta_1 : \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right).$$

Let $\mathcal{B}_\Theta$ denote the set of real-valued strictly convex and differentiable functions defined on an open set Θ, called Bregman generators. We may equivalently consider the set of strictly concave and differentiable functions $G(\theta)$ and let $F(\theta) = -G(\theta)$; see [24] (Equation (1)).

Remark 2. *The non-negativeness of the Bregman divergences for the cumulant and partition functions define the criteria for checking the strict convexity or log-convexity of a C^1 function:*

$$\begin{aligned} F(\theta) \text{ is strictly convex} &\Leftrightarrow \forall \theta_1 \neq \theta_2, B_F(\theta_1 : \theta_2) > 0, \\ &\Leftrightarrow \forall \theta_1 \neq \theta_2, F(\theta_1) > F(\theta_2) + \langle \theta_1 - \theta_2, \nabla F(\theta) \rangle, \end{aligned}$$

and

$$Z(\theta) \text{ is strictly log-convex} \Leftrightarrow \forall \theta_1 \neq \theta_2, B_{\log Z}(\theta_1 : \theta_2) > 0,$$

$$\Leftrightarrow \quad \forall \theta_1 \neq \theta_2, \log Z(\theta_1) > \log Z(\theta_2) + \left\langle \theta_1 - \theta_2, \frac{\nabla Z(\theta_2)}{Z(\theta_2)} \right\rangle.$$

The forward Bregman divergence $B_F(\theta_1 : \theta_2)$ and reverse Bregman divergence $B_F(\theta_2 : \theta_1)$ can be unified with the α-skewed Jensen divergences by rescaling $J_{F,\alpha}$ and allowing α to range in $\mathbb{R}$ [6,7]:

$$J^s_{F,\alpha}(\theta_1 : \theta_2) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1 : \theta_2), & \alpha \in \mathbb{R}\backslash\{0,1\}, \\ B_F(\theta_1 : \theta_2), & \alpha = 0, \\ 4\, J_F(\theta_1, \theta_2), & \alpha = \frac{1}{2}, \\ B^*_F(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1), & \alpha = 1. \end{cases}, \tag{9}$$

where ${B_F}^*$ denotes the reverse Bregman divergence obtained by swapping the parameter order (reference duality [6]): ${B_F}^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$.

Remark 3. *Alternatively, we may rescale J_F by a factor $\kappa(\alpha) = \frac{1}{\alpha(1-\alpha)4^{4\alpha(1-\alpha)}}$, i.e., $J^{\bar{s}}_{F,\alpha}(\theta_1 : \theta_2) = \kappa(\alpha)\, J_{F,\alpha}(\theta_1 : \theta_2)$ such that $\kappa(\frac{1}{2}) = 1$ and $J^{\bar{s}}_{F,\frac{1}{2}}(\theta_1 : \theta_2) = J_F(\theta_1, \theta_2)$.*

Next, in Section 3 we first recall the connections between these Jensen and Bregman divergences, which are divergences between parameters, and the statistical divergence counterparts between probability densities. Then, in Section 4 we introduce the novel connections between these parameter divergences and α-divergences between unnormalized densities.

3. Divergences Related to the Cumulant Function

Consider the scaled α-skewed Bhattacharyya distances [7,15] between two probability densities $p(x)$ and $q(x)$:

$$D^s_{B,\alpha}(p : q) = -\frac{1}{\alpha(1-\alpha)} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu, \quad \alpha \in \mathbb{R}\backslash\{0,1\}.$$

The scaled α-skewed Bhattacharyya distances can additionally be interpreted as Rényi divergences [25] scaled by $\frac{1}{\alpha}$: $D^s_{B,\alpha}(p : q) = \frac{1}{\alpha} D_{R,\alpha}(p : q)$, where the Rényi α-divergences are defined by

$$D_{R,\alpha}(p : q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha}\, \mathrm{d}\mu.$$

The Bhattacharyya distance $D_B(p, q) = -\log \int \sqrt{pq} \mathrm{d}\mu$ corresponds to one-fourth of $D^s_{B,\frac{1}{2}}(p : q)$: $D_B(p, q) = \frac{1}{4} D^s_{B,\frac{1}{2}}(p : q)$. Because $D^s_{B,\alpha}$ tends to the Kullback–Leibler divergence D_{KL} when $\alpha \to 1$ and to the reverse Kullback–Leibler divergence ${D_{\mathrm{KL}}}^*$ when $\alpha \to 0$, we have

$$D^s_{B,\alpha}(p : q) = \begin{cases} -\frac{1}{\alpha(1-\alpha)} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\}, \\ D_{\mathrm{KL}}(p : q), & \alpha = 1, \\ 4\, D_B(p, q) & \alpha = \frac{1}{2}, \\ {D_{\mathrm{KL}}}^*(p : q) = D_{\mathrm{KL}}(q : p) & \alpha = 0. \end{cases}$$

When both probability densities belong to the same exponential family $\mathcal{E} = \{p_\theta(x) : \theta \in \Theta\}$ with cumulant $F(\theta)$, we have the following proposition.

Proposition 4 ([7])**.** *The scaled α-skewed Bhattacharyya distances between two probability densities p_{θ_1} and p_{θ_2} of an exponential family amount to the scaled α-skewed Jensen divergence between their natural parameters:*

$$D^s_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J^s_{F,\alpha}(\theta_1, \theta_2). \tag{10}$$

Proof. The proof follows by first considering the α-skewed Bhattacharyya similarity coefficient $\rho_\alpha(p,q) = \int p^\alpha q^{1-\alpha} \mathrm{d}\mu$.

$$\begin{aligned}\rho_\alpha(p_{\theta_1} : p_{\theta_2}) &= \int \exp(\langle \theta_1, x\rangle - F(\theta_1))^\alpha \exp(\langle \theta_2, x\rangle - F(\theta_2))^{1-\alpha} \mathrm{d}\mu, \\ &= \int \exp(\langle \alpha\theta_1 + (1-\alpha)\theta_2), x\rangle) \exp(-(\alpha F(\theta_1) + (1-\alpha)F(\theta_2)))\mathrm{d}\mu.\end{aligned}$$

Multiplying the last equation by $\exp(F(\alpha\theta_1 + (1-\alpha)\theta_2)) \exp(-F(\alpha\theta_1 + (1-\alpha)\theta_2)) = \exp(0) = 1$ with $\bar{\theta} = \alpha\theta_1 + (1-\alpha)\theta_2$, we obtain

$$\rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \exp(-(\alpha F(\theta_1) + (1-\alpha)F(\theta_2)) \exp(F(\bar{\theta})) \int \exp(\langle \bar{\theta}, x\rangle - F(\bar{\theta}))\mathrm{d}\mu.$$

Because $\bar{\theta} \in \Theta$, we have $\int \exp(\langle \bar{\theta}, x\rangle - F(\bar{\theta}))\mathrm{d}\mu = 1$; therefore, we obtain

$$\rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \exp(-J_{F,\alpha}(\theta_1 : \theta_2)).$$

□

For practitioners in machine learning, it is well known that the Kullback–Leibler divergence between two probability densities p_{θ_1} and p_{θ_2} of an exponential family amounts to a Bregman divergence for the cumulant generator on a swapped parameter order (e.g., [26,27]):

$$D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = B_F(\theta_2 : \theta_1).$$

This is a particular instance of Equation (10) obtained for $\alpha = 1$:

$$D^{\mathrm{s}}_{B,1}(p_{\theta_1} : p_{\theta_2}) = J^{\mathrm{s}}_{F,1}(\theta_1, \theta_2).$$

This formula has been further generalized in [28] by considering truncations of exponential family densities. Let $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \mathcal{X}$ and $\mathcal{E}_1 = \{1_{\mathcal{X}_1}(x)\, p_\theta(x)\}$, $\mathcal{E}_2 = \{1_{\mathcal{X}_2}(x)\, q_{\theta'}(x)\}$ be two truncated families of $\mathcal{X}$ with corresponding cumulant functions

$$F_1(\theta) = \log \int_{\mathcal{X}_1} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu$$

and

$$F_2(\theta') = \log \int_{\mathcal{X}_2} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu \geq F_1(\theta').$$

Then, we have

$$\begin{aligned}D_{\mathrm{KL}}(p_{\theta_1} : q_{\theta'_2}) &= B_{F_2,F_1}(\theta'_2 : \theta_1), \\ &= F_2(\theta'_2) - F_1(\theta_1) - \langle \theta'_2 - \theta_1, \nabla F_1(\theta_1)\rangle.\end{aligned}$$

Truncated exponential families are normalized exponential families which may not be regular [29], i.e., the parameter space Θ may not be open.

4. Divergences Related to the Partition Function

Certain exponential families have intractable cumulant/partition functions (e.g., exponential families with sufficient statistics $t(x) = (x, x^2, \ldots, x^m)$ for high degrees m [20]) or cumulant/partition functions which require exponential time to compute [30] (e.g., graphical models [16], high-dimensional grid sample spaces, energy-based models [17] in deep learning, etc.). In such cases, the maximum likelihood estimator (MLE) cannot be used to infer the natural parameter of exponential densities. Many alternative methods have been proposed to handle such exponential families with untractable partition functions, e.g., score matching [31] or divergence-based inference [32,33]). Thus, it is important to consider dissimilarities between non-normalized statistical models.

The squared Hellinger distance [1] between two positive potentially unnormalized densities $\tilde{p}$ and $\tilde{q}$ is defined by

$$\begin{aligned} D_H^2(\tilde{p},\tilde{q}) &= \frac{1}{2}\int(\sqrt{\tilde{p}}-\sqrt{\tilde{q}})^2\,\mathrm{d}\mu, \\ &= \frac{\int\tilde{p}\mathrm{d}\mu+\int\tilde{q}\mathrm{d}\mu}{2}-\int\sqrt{\tilde{p}\tilde{q}}\mathrm{d}\mu. \end{aligned}$$

Notice that the Hellinger divergence can be interpreted as the integral of the difference between the arithmetical mean $A(\tilde{p},\tilde{q})=\frac{\tilde{p}+\tilde{q}}{2}$ minus the geometrical mean $G(\tilde{p},\tilde{q})=\sqrt{\tilde{p}\tilde{q}}$ of the densities: $D_H^2(\tilde{p},\tilde{q})=\int(A(\tilde{p},\tilde{q})-G(\tilde{p},\tilde{q}))\mathrm{d}\mu$. This further proves that $D_H(\tilde{p},\tilde{q})\geq 0$, as $A\geq G$. The Hellinger distance D_H satisfies the metric axioms of distances.

When considering unnormalized densities $\tilde{p}_{\theta_1}=\exp(\langle t(x),\theta_1\rangle)$ and $\tilde{p}_{\theta_2}=\exp(\langle t(x),\theta_2\rangle)$ of an exponential family $\mathcal{E}$ with a partition function $Z(\theta)=\int\tilde{p}_\theta\,\mathrm{d}\mu$, we obtain

$$D_H^2(\tilde{p}_{\theta_1},\tilde{p}_{\theta_2})=\frac{Z(\theta_1)+Z(\theta_2)}{2}-Z\left(\frac{\theta_1+\theta_2}{2}\right)=J_Z(\theta_1,\theta_2), \tag{11}$$

as $\sqrt{\tilde{p}_{\theta_1}\tilde{p}_{\theta_2}}=\tilde{p}_{\frac{\theta_1+\theta_2}{2}}$.

The Kullback–Leibler divergence [1] as extended to two positive densities $\tilde{p}$ and $\tilde{q}$ is defined by

$$D_{\mathrm{KL}}(\tilde{p}:\tilde{q})=\int\left(\tilde{p}\log\frac{\tilde{p}}{\tilde{q}}+\tilde{q}-\tilde{p}\right)\mathrm{d}\mu. \tag{12}$$

When considering unnormalized densities $\tilde{p}_{\theta_1}$ and $\tilde{p}_{\theta_2}$ of $\mathcal{E}$, we obtain

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2}) &= \int\left(\tilde{p}_{\theta_1}(x)\log\frac{\tilde{p}_{\theta_1}(x)}{\tilde{p}_{\theta_2}(x)}+\tilde{p}_{\theta_2}(x)-\tilde{p}_{\theta_1}(x)\right)\mathrm{d}\mu(x), & (13)\\ &= \int\left(e^{\langle t(x),\theta_1\rangle}\langle\theta_1-\theta_2,t(x)\rangle+e^{\langle t(x),\theta_2\rangle}-e^{\langle t(x),\theta_1\rangle}\right)\mathrm{d}\mu(x), & (14)\\ &= \left\langle\int t(x)e^{\langle t(x),\theta_1\rangle}\mathrm{d}\mu(x),\theta_1-\theta_2\right\rangle+Z(\theta_2)-Z(\theta_1), & (15)\\ &= \langle\theta_1-\theta_2,\nabla Z(\theta_1)\rangle+Z(\theta_2)-Z(\theta_1)=B_Z(\theta_2:\theta_1), & (16) \end{aligned}$$

as $\nabla Z(\theta)=\int t(x)\tilde{p}_\theta(x)\mathrm{d}\mu(x)$. Let $D^*_{\mathrm{KL}}(\tilde{p}:\tilde{q})=D_{\mathrm{KL}}(\tilde{q}:\tilde{p})$ denote the reverse KLD.

More generally, the family of α-divergences [1] between the unnormalized densities $\tilde{p}$ and $\tilde{q}$ is defined for $\alpha\in\mathbb{R}$ by

$$D_\alpha(\tilde{p}:\tilde{q})=\begin{cases}\frac{1}{\alpha(1-\alpha)}\int\left(\alpha\tilde{p}+(1-\alpha)\tilde{q}-\tilde{p}^\alpha\tilde{q}^{1-\alpha}\right)\mathrm{d}\mu, & \alpha\notin\{0,1\}\\ D^*_{\mathrm{KL}}(\tilde{p}:\tilde{q})=D_{\mathrm{KL}}(\tilde{q}:\tilde{p}) & \alpha=0,\\ 4\,D_H^2(\tilde{p},\tilde{q}) & \alpha=\frac{1}{2},\\ D_{\mathrm{KL}}(\tilde{p}:\tilde{q}) & \alpha=1.\end{cases}$$

We now have $D^*_\alpha(\tilde{p}:\tilde{q})=D_\alpha(\tilde{q}:\tilde{p})=D_{1-\alpha}(\tilde{p}:\tilde{q})$, and the α-divergences are homogeneous divergences of degree 1. For all $\lambda>0$, we have $D_\alpha(\lambda\tilde{q}:\lambda\tilde{p})=\lambda\,D_\alpha(\tilde{q}:\tilde{p})$. Moreover, because $\alpha\tilde{p}+(1-\alpha)\tilde{q}-\tilde{p}^\alpha\tilde{q}^{1-\alpha}$ can be expressed as the difference of the weighted arithmetic mean minus the weighted geometric mean $A(\tilde{p},\tilde{q};\alpha;1-\alpha)-G(\tilde{p},\tilde{q};\alpha;1-\alpha)$, it follows from the arithmetical–geometrical mean inequality that we have $D_\alpha(\tilde{p}:\tilde{q})\geq 0$.

When considering unnormalized densities $\tilde{p}_{\theta_1}$ and $\tilde{p}_{\theta_2}$ of $\mathcal{E}$, we obtain

$$D_\alpha(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2})=\begin{cases}\frac{1}{\alpha(1-\alpha)}J_{Z,\alpha}(\theta_1:\theta_2), & \alpha\notin\{0,1\}\\ B_Z(\theta_1:\theta_2) & \alpha=0,\\ 4\,J_Z(\theta_1,\theta_2) & \alpha=\frac{1}{2},\\ B^*_Z(\theta_1:\theta_2)=B_Z(\theta_2:\theta_1) & \alpha=1\end{cases}.$$

Proposition 5. *The α-divergences between the unnormalized densities of an exponential family amount to scaled α-Jensen divergences between their natural parameters for the partition function*

$$D_\alpha(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2})=J^{\mathrm{s}}_{Z,\alpha}(\theta_1:\theta_2).$$

When $\alpha \in \{0,1\}$, the oriented Kullback–Leibler divergences between unnormalized exponential family densities amount to reverse Bregman divergences on their corresponding natural parameters for the partition function

$$D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = B_Z(\theta_2 : \theta_1).$$

Proof. For $\alpha \notin \{0,1\}$, consider

$$D_\alpha(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) = \frac{1}{\alpha(1-\alpha)} \int \left(\alpha \tilde{p}_{\theta_1} + (1-\alpha)\tilde{p}_{\theta_2} - \tilde{p}_{\theta_1}^{\alpha} \tilde{p}_{\theta_2}^{1-\alpha}\right) \mathrm{d}\mu.$$

Here, we have $\int \alpha \tilde{p}_{\theta_1} \mathrm{d}\mu = \alpha Z(\theta_1)$, $\int (1-\alpha)\tilde{p}_{\theta_2} \mathrm{d}\mu = (1-\alpha)Z(\theta_2)$ and $\int \tilde{p}_{\theta_1}^{\alpha} \tilde{p}_{\theta_2}^{1-\alpha} \mathrm{d}\mu = \int \tilde{p}_{\alpha\theta_1 + (1-\alpha)\theta_2} \mathrm{d}\mu = Z(\alpha\theta_1 + (1-\alpha)\theta_2)$. It follows that

$$D_\alpha(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) = \frac{1}{\alpha(1-\alpha)} J_{Z,\alpha}(\theta_1 : \theta_2) = J^{\mathrm{s}}_{Z,\alpha}(\theta_1 : \theta_2).$$

□

Notice that the KLD extended to unnormalized densities can be written as a generalized relative entropy, i.e., it can be obtained as the difference of the extended cross-entropy minus the extended entropy (self cross-entropy):

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) &= H^{\times}(\tilde{p} : \tilde{q}) - H(\tilde{p}), \\ &= \int \left(\tilde{p} \log \frac{\tilde{p}}{\tilde{q}} + \tilde{q} - \tilde{p}\right) \mathrm{d}\mu \end{aligned}$$

with

$$H^{\times}(\tilde{p} : \tilde{q}) = \int \left(\tilde{p}(x) \log \frac{1}{\tilde{q}(x)} + \tilde{q}(x)\right) \mathrm{d}\mu(x) - 1$$

and

$$H(\tilde{p}) = H^{\times}(\tilde{p} : \tilde{p}) = \int \left(\tilde{p}(x) \log \frac{1}{\tilde{p}(x)} + \tilde{p}(x)\right) \mathrm{d}\mu(x) - 1.$$

Remark 4. *In general, we can consider two unnormalized positive densities $\tilde{p}(x)$ and $\tilde{q}(x)$. Let $p(x) = \frac{\tilde{p}(x)}{Z_p}$ and $q(x) = \frac{\tilde{q}(x)}{Z_q}$ denote their corresponding normalized densities (with normalizing factors $Z_p = \int \tilde{p} \, \mathrm{d}\mu$ and $Z_q = \int \tilde{q} \, \mathrm{d}\mu$); then, the KLD between $\tilde{p}$ and $\tilde{q}$ can be expressed using the KLD between their normalized densities and normalizing factors, as follows:*

$$D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = Z_p \left(D_{\mathrm{KL}}(p : q) + \log \frac{Z_p}{Z_q}\right) + Z_q - Z_p. \tag{17}$$

Similarly, we have

$$\begin{aligned} H^{\times}(\tilde{p} : \tilde{q}) &= Z_p \, H^{\times}(p : q) - Z_p \log Z_q + Z_q - 1, & (18) \\ H(\tilde{p}) &= Z_p \, H(p) - Z_p \log Z_p + Z_p - 1, & (19) \end{aligned}$$

and $D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = H^{\times}(\tilde{p} : \tilde{q}) - H(\tilde{p})$.

Notice that Equation (17) allows us to derive the following identity between B_Z and B_F:

$$\begin{aligned} B_Z(\theta_2 : \theta_1) &= Z(\theta_1) \, B_F(\theta_2 : \theta_1) + Z(\theta_1) \log \frac{Z(\theta_1)}{Z(\theta_2)} + Z(\theta_2) - Z(\theta_1), & (20) \\ &= \exp(F(\theta_1)) \, B_F(\theta_2 : \theta_1) + (\exp F(\theta_1))(F(\theta_1) - F(\theta_2)) + \exp(F(\theta_2)) - \exp(F(\theta_1)). & (21) \end{aligned}$$

Let $D_{\mathrm{skl}}(a : b) = a \log \frac{a}{b} + b - a$ be the scalar KLD for $a > 0$ and $b > 0$. Then, we can rewrite Equation (17) as

$$D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = Z_p \, D_{\mathrm{KL}}(p : q) + D_{\mathrm{skl}}(Z_p : Z_q),$$

and we have

$$B_Z(\theta_2 : \theta_1) = Z(\theta_1)\, B_F(\theta_2 : \theta_1) + D_{\mathrm{skl}}(Z(\theta_1) : Z(\theta_2)).$$

In addition, the KLD between the unnormalized densities $\tilde{p}$ and $\tilde{q}$ with support $\mathcal{X}$ can be written as a definite integral of a scalar Bregman divergence:

$$D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = \int_{\mathcal{X}} D_{\mathrm{skl}}(\tilde{p}(x) : \tilde{q}(x))\, \mathrm{d}\mu(x) = \int_{\mathcal{X}} B_{f_{\mathrm{skl}}}(\tilde{p}(x) : \tilde{q}(x))\, \mathrm{d}\mu(x),$$

where $f_{\mathrm{skl}}(x) = x \log x - x$. Because $B_{f_{\mathrm{skl}}}(a : b) \geq 0 \forall a > 0, b > 0$, we can deduce that $D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) \geq 0$ with equality iff $\tilde{p}(x) = \tilde{q}(x)$ μ almost everywhere.

Notice that $B_Z(\theta_2 : \theta_1) = Z(\theta_1)\, B_F(\theta_2 : \theta_1) + D_{\mathrm{skl}}(Z(\theta_1) : Z(\theta_2))$ can be interpreted as the sum of two divergences, that is, a conformal Bregman divergence with a scalar Bregman divergence.

Remark 5. *Consider the KLD between the normalized p_{θ_1} and unnormalized $\tilde{p}_{\theta_2}$ densities of the same exponential family. In this case, we have*

$$\begin{aligned} D_{\mathrm{KL}}(p_{\theta_1} : \tilde{p}_{\theta_2}) &= B_F(\theta_2 : \theta_1) - \log Z(\theta_2) + Z(\theta_2) - 1, && (22) \\ &= Z(\theta_2) - 1 - F(\theta_1) - \langle \theta_2 - \theta_1, \nabla F(\theta_2) \rangle, \\ &= B_{Z-1,F}(\theta_2 : \theta_1). && (23) \end{aligned}$$

The divergence $B_{Z-1,F}$ is a dual Bregman pseudo-divergence [28]*:*

$$B_{F_1,F_2}(\theta_1 : \theta_2) = F_1(\theta_1) - F_2(\theta_2) - \langle \theta_1 - \theta_2, \nabla F_2(\theta_2) \rangle,$$

for F_1 and F_2 that are two strictly convex and smooth functions such that $F_1 \geq F_2$. Indeed, we can check that generators $F_1(\theta) = Z(\theta) - 1$ and $F_2(\theta) = F(\theta)$ are both Bregman generators; then, we have $F_1(\theta) \geq F_2(\theta)$, as $e^x \geq x + 1$ for all x (with equality when $x = 0$), i.e., $Z(\theta) - 1 \geq F(\theta)$.

More generally, the α-divergences between p_{θ_1} and $\tilde{p}_{\theta_2}$ can be written as

$$D_\alpha(p_{\theta_1} : \tilde{p}_{\theta_2}) = \frac{1}{\alpha(1-\alpha)} \left(\alpha Z(\theta_1) + (1-\alpha) - \frac{Z(\alpha\theta_1 + (1-\alpha)\theta_2)}{Z(\theta_2)} \right), \quad (24)$$

with the (signed) α-skewed Bhattacharyya distances provided by

$$D_{B,\alpha}(p_{\theta_1} : \tilde{p}_{\theta_2}) = \log Z(\theta_2) - \log Z(\alpha\theta_1 + (1-\alpha)\theta_2).$$

Let us illustrate Proposition 5 with some examples.

Example 1. *Consider the family of exponential distributions $\mathcal{E} = \{p_\lambda(x) = 1_{x \geq 0}\, \lambda \exp(-\lambda x)\}$, where $\mathcal{E}$ is an exponential family with a natural parameter $\theta = \lambda$, parameter space $\Theta = \mathbb{R}_{>0}$, sufficient statistic $t(x) = -x$. The partition function is $Z(\theta) = \frac{1}{\theta}$, with $Z'(\theta) = -\frac{1}{\theta^2}$ and $Z''(\theta) = \frac{2}{\lambda^3} > 0$, while the cumulant function is $F(\theta) = \log Z(\theta) = -\log \theta$ with moment parameter $\eta = E_{p_\lambda}[t(x)] = F'(\theta) = -\frac{1}{\theta}$. The α-divergences between two unnormalized exponential distributions are*

$$D_\alpha(\tilde{p}_{\lambda_1} : \tilde{p}_{\lambda_2}) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{Z,\alpha}(\theta_1 : \theta_2) = \frac{(\lambda_1 - \lambda_2)^2)}{\alpha\lambda_1^2\lambda_2 + (1-\alpha)\lambda_1\lambda_2^2} & \alpha \notin \{0, 1\} \\ D_{\mathrm{KL}}(\tilde{p}_{\lambda_2} : \tilde{p}_{\lambda_1}) = B_Z(\theta_1 : \theta_2) = \frac{(\lambda_1 - \lambda_2)^2}{\lambda_1\lambda_2^2} & \alpha = 0, \\ 4\, J_Z(\theta_1, \theta_2) = \frac{(\lambda_1 - \lambda_2)^2}{2(\lambda_1\lambda_2^2 + \lambda_1^2\lambda_2)} & \alpha = \frac{1}{2}, \\ D_{\mathrm{KL}}(\tilde{p}_{\lambda_1} : \tilde{p}_{\lambda_2}) = B_Z(\theta_2 : \theta_1) = \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2\lambda_1^2} & \alpha = 1 \end{cases}. \quad (25)$$

Example 2. *Consider the family of univariate centered normal distributions with $\tilde{p}_{\sigma^2}(x) \propto \exp(-\frac{x^2}{2\sigma^2})$ and partition function $Z(\sigma^2) = \sqrt{2\pi\sigma^2}$ such that $p_{\sigma^2}(x) = \frac{1}{Z(\sigma^2)}\, \tilde{p}_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})$. Here, we have a natural parameter $\theta = \frac{1}{\sigma^2} \in \Theta = \mathbb{R}_{>0}$ and sufficient statistic $t(x) = -\frac{x^2}{2}$. The*

partition function expressed with the natural parameter is $Z(\theta) = \sqrt{\frac{2\pi}{\theta}}$, *with* $Z'(\theta) = -\sqrt{\frac{\pi}{2}}\theta^{-\frac{3}{2}}$ *and* $Z''(\theta) = \frac{3\sqrt{\pi}}{2^{\frac{3}{2}}}\theta^{-\frac{5}{2}} > 0$ *(strictly convex on* Θ*). The unnormalized KLD between* $\tilde{p}_{\sigma_1^2}$ *and* $\tilde{p}_{\sigma_2^2}$ *is*

$$D_{\mathrm{KL}}(\tilde{p}_{\sigma_1^2} : \tilde{p}_{\sigma_2^2}) = B_Z(\theta_2 : \theta_1) = \sqrt{\frac{\pi}{2}}\left(2\sigma_2 - 3\sigma_1 + \frac{\sigma_1^3}{\sigma_2^2}\right).$$

We can check that we have $D_{\mathrm{KL}}(\tilde{p}_{\sigma^2} : \tilde{p}_{\sigma^2}) = 0$.

For the Hellinger divergence, we have

$$D_H^2(\tilde{p}_{\sigma_1^2} : \tilde{p}_{\sigma_2^2}) = J_Z(\theta_1, \theta_2) = \sqrt{\frac{\pi}{2}}(\sigma_1 + \sigma_2) - 2\sqrt{\pi}\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}},$$

and we can check that $D_H(\tilde{p}_{\sigma^2} : \tilde{p}_{\sigma^2}) = 0$.

Consider the family of the d-variate case of centered normal distributions with unnormalized density

$$\tilde{p}_\Sigma(x) \propto \exp(-\frac{1}{2}x^\top\Sigma^{-1}x) = \exp\left(-\frac{1}{2}\mathrm{tr}(x^\top\Sigma^{-1}x)\right) = \exp\left(-\frac{1}{2}\mathrm{tr}(xx^\top\Sigma^{-1})\right)$$

obtained using the matrix trace cyclic property, where Σ *is the covariance matrix. Here, we have* $\theta = \Sigma^{-1}$ *(precision matrix) and* $\Theta = \mathrm{Sym}_{++}(d)$ *for* $t(x) = -\frac{1}{2}xx^\top$, *with the matrix inner product* $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. *The partition function* $Z(\Sigma) = (2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}$ *expressed with the natural parameter is* $Z(\theta) = (2\pi)^{\frac{d}{2}}\sqrt{\frac{1}{\det(\theta)}}$. *This is a convex function with*

$$\nabla Z(\theta) = -\frac{1}{2}(2\pi)^{\frac{d}{2}}\frac{\nabla_\theta\det(\theta)}{\det(\theta)^{\frac{3}{2}}} = -\frac{1}{2}(2\pi)^{\frac{d}{2}}\frac{\theta^{-1}}{\det(\theta)^{\frac{1}{2}}},$$

as $\nabla_\theta\det(\theta) = \det(\theta)\theta^{-\top}$ *using matrix calculus.*

Now, consider the family of univariate normal distributions

$$\mathcal{E} = \left\{p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right\}.$$

Let $\theta = \left(\theta_1 = \frac{1}{\sigma_2}, \theta_2 = \frac{\mu}{\sigma^2}\right)$ *and*

$$Z(\theta_1, \theta_2) = \sqrt{\frac{2\pi}{\theta_1}}\exp\left(\frac{1}{2}\frac{\theta_2^2}{\theta_1}\right).$$

The unnormalized densities are $\tilde{p}_\theta(x) = \exp\left(-\frac{\theta_1 x^2}{2} + x\theta_2\right)$, *and we have*

$$\nabla Z(\theta) = \begin{bmatrix} \sqrt{\frac{\pi}{2}}\frac{(\theta_1+\theta_2^2)\exp\left(\frac{\theta_2^2}{2\theta_1}\right)}{\theta_1^{\frac{5}{2}}} \\ \sqrt{2\pi}\frac{\theta_2\exp\left(\frac{\theta_2^2}{2\theta_1}\right)}{\theta_1^{\frac{3}{2}}} \end{bmatrix}.$$

It follows that $D_{\mathrm{KL}}[\tilde{p}_\theta : \tilde{p}_{\theta'}] = B_Z(\theta' : \theta)$.

5. Deforming Convex Functions and Their Induced Dually Flat Spaces

5.1. Comparative Convexity

The log-convexity can be interpreted as a special case of comparative convexity with respect to a pair (M, N) of comparable weighted means [9], as follows.

A function Z is (M, N)-convex if and only if for $\alpha \in [0, 1]$ we have

$$Z(M(x, y; \alpha, 1-\alpha)) \leq N(Z(x), Z(y); \alpha, 1-\alpha), \tag{26}$$

and is strictly (M, N)-convex iff we have strict inequality for $\alpha \in (0,1)$ and $x \neq y$. Furthermore, a function Z is (strictly) (M, N)-concave if $-Z$ is (strictly) (M, N)-convex.

Log-convexity corresponds to (A, G)-convexity, i.e., convexity with respect to the weighted arithmetical and geometrical means defined respectively by $A(x, y; \alpha, 1-\alpha) = \alpha x + (1-\alpha)y$ and $G(x, y; \alpha, 1-\alpha) = x^{\alpha} y^{1-\alpha}$. Ordinary convexity is (A, A)-convexity.

A weighted quasi-arithmetical mean [34] (also called a Kolmogorov–Nagumo mean [35]) is defined for a continuous and strictly increasing function h by

$$M_h(x, y; \alpha, 1-\alpha) = h^{-1}(\alpha h(x) + (1-\alpha)h(x)).$$

We let $M_h(x, y) = M_h\left(x, y; \frac{1}{2}, \frac{1}{2}\right)$. Quasi-arithmetical means include the arithmetical mean obtained for $h(u) = \mathrm{id}(u) = u$ and the geometrical mean for $h(u) = \log(u)$, and more generally power means

$$M_p(x, y; \alpha, 1-\alpha) = (\alpha x^p + (1-\alpha)y^p)^{\frac{1}{p}} = M_{h_p}(x, y; \alpha, 1-\alpha), \quad p \neq 0,$$

which are quasi-arithmetical means obtained for the family of generators $h_p(u) = \frac{u^p - 1}{p}$ with inverse $h_p^{-1}(u) = (1+up)^{\frac{1}{p}}$. In the limit $p \to 0$, we have $M_0(x, y) = G(x, y)$ for the generator $\lim_{p \to 0} h_p(u) = h_0(u) = \log u$.

Proposition 6 ([36,37]). *A function $Z(\theta)$ is strictly (M_ρ, M_τ)-convex with respect to two strictly increasing smooth functions ρ and τ if and only if the function $F = \tau \circ Z \circ \rho^{-1}$ is strictly convex.*

Notice that the set of strictly increasing smooth functions form a non-Abelian group, with the group operation as the function composition, the neutral element as the identity function, and the inverse element as the functional inverse function.

Because log-convexity is $(A = M_{\mathrm{id}}, G = M_{\log})$-convexity, a function Z is strictly log-convex iff $\log \circ Z \circ \mathrm{id}^{-1} = \log \circ Z$ is strictly convex. We have

$$Z = \tau^{-1} \circ F \circ \rho \Leftrightarrow F = \tau \circ Z \circ \rho^{-1}.$$

Starting from a given convex function $F(\theta)$, we can deform the function $F(\theta)$ to obtain a function $Z(\theta)$ using two strictly monotone functions τ and ρ: $Z(\theta) = \tau^{-1}(F(\rho(\theta)))$.

For a (M_ρ, M_τ)-convex function $Z(\theta)$ which is also strictly convex, we can define a pair of Bregman divergences B_Z and B_F with $F(\theta) = \tau(Z(\rho^{-1}(\theta)))$ and a corresponding pair of skewed Jensen divergences.

Thus, we have the following generic deformation scheme.

$$\underbrace{F = \tau \circ Z \circ \rho^{-1}}_{(M_{\rho^{-1}}, M_{\tau^{-1}})\text{-convex when } Z \text{ is convex}} \quad \underset{(\rho^{-1}, \tau^{-1})\text{-deformation}}{\overset{(\rho, \tau)\text{-deformation}}{\rightleftarrows}} \quad \underbrace{Z = \tau^{-1} \circ F \circ \rho}_{(M_\rho, M_\tau)\text{-convex when } F \text{ is convex}}$$

In particular, when the function Z is deformed by strictly increasing the power functions h_{p_1} and h_{p_2} for p_1 and p_2 in $\mathbb{R}$ as

$$Z_{p_1, p_2} = h_{p_2} \circ Z \circ h_{p_1}^{-1},$$

then Z_{p_1, p_2} is strictly convex when it is strictly (M_{p_1}, M_{p_2})-convex, and as such induces corresponding Bregman and Jensen divergences.

Example 3. *Consider the partition function $Z(\theta) = \frac{1}{\theta}$ of the exponential distribution family ($\theta > 0$ with $\Theta = \mathbb{R}_{>0}$). Let $Z_p(\theta) = (h_p \circ Z)(\theta) = \frac{\theta^{-p} - 1}{p}$; then, we have $Z_p''(\theta) = (1+p)\frac{1}{\theta^{2+p}} > 0$ when $p > -1$. Thus, we can deform Z smoothly by Z_p while preserving the convexity by ranging p from -1 to $+\infty$. In this way, we obtain a corresponding family of Bregman and Jensen divergences.*

The proposed convex deformation using quasi-arithmetical mean generators differs from the interpolation of convex functions using the technique of proximal averaging [38].

Note that in [37] the comparative convexity with respect to a pair of quasi-arithmetical means (M_ρ, M_τ) is used to define a (M_ρ, M_τ)-Bregman divergence, which turns out to be equivalent to a conformal Bregman divergence on the ρ-embedding of the parameters.

5.2. Dually Flat Spaces

We start with a refinement of the class of convex functions used to generate dually flat spaces.

Definition 2 (Legendre type function [39]). *(Θ, F) is of Legendre type if the function $F : \Theta \to \mathbb{R}$ is strictly convex and differentiable with $\Theta \neq \emptyset$ and*

$$\lim_{\lambda \to 0} \frac{d}{d\lambda} F(\lambda\theta + (1-\lambda)\bar{\theta}) = -\infty, \quad \forall \theta \in \Theta, \forall \bar{\theta} \in \partial\Theta. \tag{27}$$

Legendre-type functions $F(\Theta)$ admit a convex conjugate $F^*(\eta)$ via the Legendre transform $F^*(\eta) = \sup_{\theta \in \Theta} \langle \theta, \eta \rangle - F(\theta)$:

$$F^*(\eta) = \left\langle \nabla F^{-1}(\eta), \eta \right\rangle - F(\nabla F^{-1}(\eta)).$$

A smooth and strictly convex function $(\Theta, F(\theta))$ of Legendre type induces a dually flat space [1] $\mathcal{M}$, i.e., a smooth Hessian manifold [40] with a single global chart $(\Theta, \theta(\cdot))$ [1]. A canonical divergence $D(p : q)$ between two points p and q of $\mathcal{M}$ is viewed as a single-parameter contrast function [41] $\mathcal{D}(r_{pq})$ on the product manifold $\mathcal{M} \times \mathcal{M}$. The canonical divergence and its dual canonical divergence $\mathcal{D}^*(r_{qp}) = \mathcal{D}(r_{pq})$ can be expressed equivalently as either dual Bregman divergences or dual Fenchel–Young divergences (Figure 2):

$$\begin{aligned} \mathcal{D}(r_{pq}) &= B_F(\theta(p) : \theta(q)) = Y_{F,F^*}(\theta(p) : \eta(q)), \\ &= \mathcal{D}^*(r_{qp}) = B_{F^*}(\eta(q) : \eta(p)) = Y_{F^*,F}(\eta(q) : \theta(p)), \end{aligned}$$

where Y_{F,F^*} is the Fenchel–Young divergence:

$$Y_{F,F^*}(\theta(p) : \eta(q)) = F(\theta(p)) + F^*(\eta(q)) - \langle \theta(p), \eta(q) \rangle.$$

We have the dual global coordinate system $\eta = \nabla F(\theta)$ and the domain $H = \{\nabla F(\theta) : \theta \in \Theta\}$ which defines the dual Legendre-type potential function $(H, F^*(\eta))$. The Legendre-type function ensures that $F^{**} = F$ (a sufficient condition is to have F be convex and lower semi-continuous [42]).

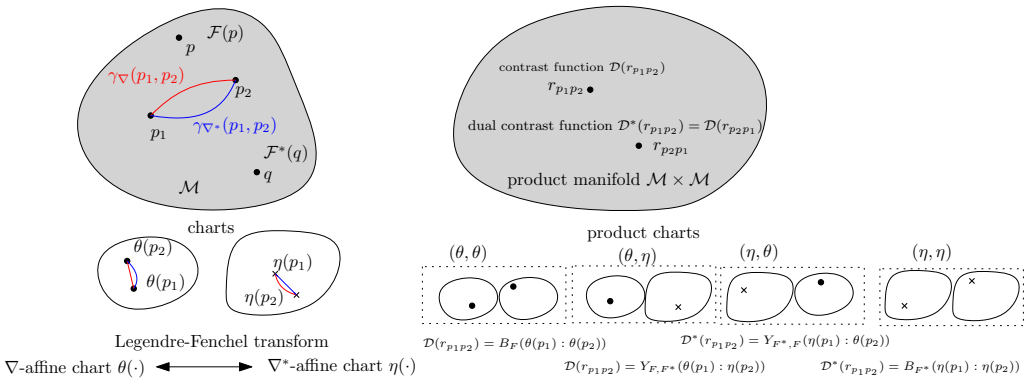

Figure 2. The canonical divergence $\mathcal{D}$ and dual canonical divergence $\mathcal{D}^*$ on a dually flat space $\mathcal{M}$ equipped with potential functions $\mathcal{F}$ and $\mathcal{F}^*$ can be viewed as single-parameter contrast functions on the product manifold $\mathcal{M} \times \mathcal{M}$: The divergence $\mathcal{D}$ can be expressed using either the $\theta \times \theta$-coordinate system as a Bregman divergence or the mixed $\theta \times \eta$-coordinate system as a Fenchel–Young divergence. Similarly, the dual divergence $\mathcal{D}$ can be expressed using either the $\eta \times \eta$-coordinate system as a dual Bregman divergence or the mixed $\eta \times \theta$-coordinate system as a dual Fenchel–Young divergence.

A manifold $\mathcal{M}$ is called dually flat, as the torsion-free affine connections ∇ and ∇^* induced by the potential functions $F(\theta)$ and $F^*(\eta)$ linked with the Legendre–Fenchel transformation are flat [1], that is, their Christoffel symbols vanishes in the dual coordinate system: $\Gamma(\theta) = 0$ and $\Gamma^*(\eta) = 0$.

The Legendre-type function $(\Theta, F(\theta))$ is not defined uniquely; the function $\bar{F}(\bar{\theta}) = F(A\theta + b) + C\theta + d$ with $\bar{\theta} = A\theta + b$ for A and C invertible matrices and b and d vectors defines the same dually flat space with the same canonical divergence $D(p, q)$:

$$D(p : q) = B_F(\theta(p) : \theta(q)) = B_{\bar{F}}(\bar{\theta}(p) : \bar{\theta}(q)).$$

Thus, a log-convex Legendre-type function $Z(\theta)$ induces two dually flat spaces by considering the DFSs induced by $Z(\theta)$ and $F(\theta) = \log Z(\theta)$. Let the gradient maps be $\eta = \nabla Z(\theta)$ and $\tilde{\eta} = \nabla F(\theta) = \frac{\eta}{Z(\theta)}$.

When $F(\theta)$ is chosen as the cumulant function of an exponential family, the Bregman divergence $B_F(\theta_1 : \theta_2)$ can be interpreted as a statistical divergence between corresponding probability densities, meaning that the Bregman divergence amounts to the reverse Kullback–Leibler divergence: $B_F(\theta_1 : \theta_2) = D^*_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2})$, where D^*_{KL} is the reverse KLD.

Notice that deforming a convex function $F(\theta)$ into $F(\rho(\theta))$ such that $F \circ \rho$ remains strictly convex has been considered by Yoshizawa and Tanabe [43] to build a two-parameter deformation $\rho_{\alpha,\beta}$ of the dually flat space induced by the cumulant function $F(\theta)$ of the multivariate normal family. Additionally, see the method of Hougaard [44] for obtaining other exponential families from a given exponential family.

Thus, in general, there are many more dually flat spaces with corresponding divergences and statistical divergences than the usually considered exponential family manifold [5] induced by the cumulant function. It is interesting to consider their use in information sciences.

6. Conclusions and Discussion

For machine learning practioners, it is well known that the Kullback–Leibler divergence (KLD) between two probability densities p_{θ_1} and p_{θ_2} of an exponential family with cumulant function F (free energy in thermodynamics) amounts to a reverse Bregman divergence [26] induced by F, or equivalently to a reverse Fenchel–Young divergence [27]

$$D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = B_F(\theta_2 : \theta_1) = Y_{F,F^*}(\theta_2 : \eta_1),$$

where $\eta = \nabla F(\theta)$ is the dual moment or expectation parameter.

In this paper, we have shown that the KLD as extended to positive unnormalized densities $\tilde{p}_{\theta_1}$ and $\tilde{p}_{\theta_2}$ of an exponential family with a convex partition function $Z(\theta)$ (Laplace transform) amounts to a reverse Bregman divergence induced by Z, or equivalently to a reverse Fenchel–Young divergence

$$D_{\mathrm{KL}}(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) = B_Z(\theta_2 : \theta_1) = Y_{Z,Z^*}(\theta_2 : \tilde{\eta}_1),$$

where $\tilde{\eta} = \nabla Z(\theta)$.

More generally, we have shown that the scaled α-skewed Jensen divergences induced by the cumulant and partition functions between natural parameters coincide with the scaled α-skewed Bhattacharyya distances between probability densities and the α-divergences between unnormalized densities, respectively:

$$\begin{aligned} D^s_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) &= J^s_{F,\alpha}(\theta_1 : \theta_2), \\ D_{\alpha}(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) &= J^s_{Z,\alpha}(\theta_1 : \theta_2). \end{aligned}$$

We have noted that the partition functions Z of exponential families are both convex and log-convex, and that the corresponding cumulant functions are both convex and exponentially convex.

Figure 3 summarizes the relationships between statistical divergences and between the normalized and unnormalized densities of an exponential family, as well as the corresponding divergences between their natural parameters. Notice that Brekelmans and Nielsen [45] considered deformed uni-order likelihood ratio exponential families (LREFs)

for annealing paths and obtained an identity for the α-divergences between unnormalized densities and Bregman divergences induced by multiplicatively scaled partition functions.

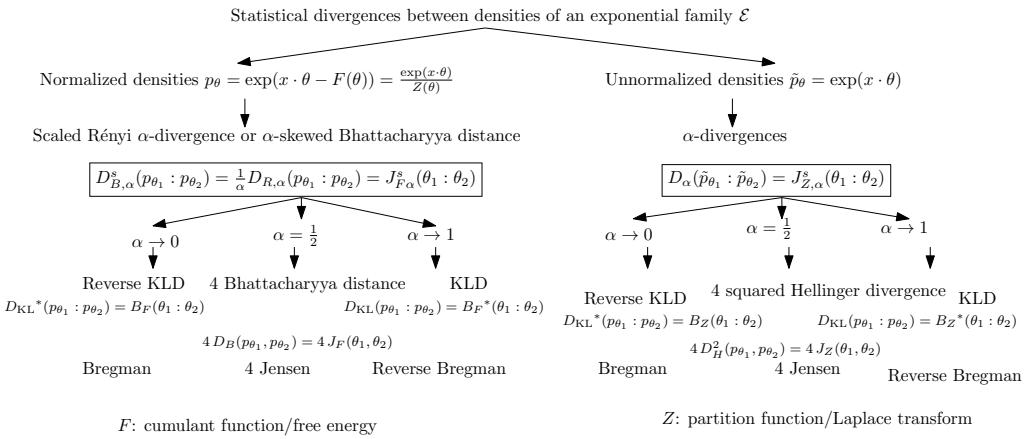

Figure 3. Statistical divergences between normalized p_θ and unnormalized $\tilde{p}_\theta$ densities of an exponential family $\mathcal{E}$ with corresponding divergences between their natural parameters. Without loss of generality, we consider a natural exponential family (i.e., $t(x) = x$ and $k(x) = 0$) with cumulant function F and partition function Z, with J_F and B_F respectively denoting the Jensen and Bregman divergences induced by the generator F. The statistical divergences $D_{R,\alpha}$ and $D_{B,\alpha}$ denote the Rényi α-divergences and skewed α-Bhattacharyya distances, respectively. The superscript "s" indicates rescaling by the multiplicative factor $\frac{1}{\alpha(1-\alpha)}$, while the superscript "*" denotes the reverse divergence obtained by swapping the parameter order.

Because the log-convex partition function is also convex, we have generalized the principle of building pairs of convex generators using the comparative convexity with respect to a pair of quasi-arithmetical means, and have further discussed the induced dually flat spaces and divergences. In particular, by considering the convexity-preserving deformations obtained by power mean generators, we have shown how to obtain a family of convex generators and dually flat spaces. Notice that some parametric families of Bregman divergences, such as the α-divergences [46], β-divergences [47], and V-geometry [48] of symmetric positive-definite matrices, yield families of dually flat spaces.

Banerjee et al. [49] proved a duality between regular exponential families and a subclass of Bregman divergences, which they accordingly termed regular Bregman divergences. In particular, this duality allows the Maximum Likelihood Estimator (MLE) of an exponential family with a cumulant function F to be viewed as a right-sided Bregman centroid with respect to the Legendre–Fenchel dual F^*. In [50], the scope of this duality was further extended for arbitrary Bregman divergences by introducing a class of generalized exponential families.

Concave deformations have been recently studied in [51], where the authors introduced the $\log_\phi$-concavity induced by a positive continuous function ϕ generating a deformed logarithm $\log_\phi$ as the $(A, \log_\phi)$-comparative concavity (Definition 1.2 in [51]), as well as the weaker notion of F-concavity which corresponds to the (A, F)-concavity (Definition 2.1 in [51], requiring strictly increasing functions F). Our deformation framework $Z = \tau^{-1} \circ F \circ \rho$ is more general, as it is double-sided. We jointly deform the function F by $F_\tau = \tau^{-1} \circ F$ and its argument θ by $\theta_\rho = \rho(\theta)$.

Exponentially concave functions have been considered as generators of L-divergences in [24]; α-exponentially concave functions G such that $\exp(\alpha G)$ are concave for $\alpha > 0$ generalize the L-divergences to L_α-divergences, which can be expressed equivalently using a generalization of the Fenchel–Young divergence based on the c-transforms [24]. When $\alpha < 0$, exponentially convex functions are considered instead of exponentially concave functions. The information geometry induced by L_α-divergences are dually projectively flat with constant curvature, and reciprocally possess a dually projectively flat

structure with constant curvature, inducing (locally) a canonical $L_{-\alpha}$-divergence. Wong and Zhang [52] investigated a one-parameter deformation of convex duality, called λ-duality, by considering functions f such that $\frac{1}{\lambda}(e^{\lambda f} - 1)$ are convex for $\lambda \neq 0$. They defined the λ-conjugate transform as a particular case of the c-transform [24] and studie the information geometry of the induced λ-logarithmic divergences. The λ-duality yields a generalization of exponential and mixture families to λ-exponential and λ-mixture families related to the Rényi divergence.

Finally, certain statistical divergences, called projective divergences, are invariant under rescaling, and as such can define dissimilarities between non-normalized densities. For example, the γ-divergences [32] D_γ are such that $D_\gamma(p : q) = D_\gamma(\tilde{p} : \tilde{q})$ (with γ-divergences tending to the KLD when $\gamma \to 0$) or the Cauchy–Schwarz divergence [53].

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author heartily thanks the three reviewers for their helpful comments which led to this improved paper.

Conflicts of Interest: Author Frank Nielsen is employed by the company Sony Computer Science Laboratories Inc. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare no conflicts of interest.

References

1. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016.
2. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]
3. Nielsen, F.; Hadjeres, G. Monte Carlo information-geometric structures. In *Geometric Structures of Information*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 69–103.
4. Brown, L.D. Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. In *Lecture Notes-Monograph Series*; Cornell University: Ithaca, NY, USA, 1986; Volume 9.
5. Scarfone, A.M.; Wada, T. Legendre structure of κ-thermostatistics revisited in the framework of information geometry. *J. Phys. Math. Theor.* **2014**, *47*, 275002. [CrossRef]
6. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195. [CrossRef] [PubMed]
7. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
8. Cichocki, A.; Amari, S.I. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568. [CrossRef]
9. Niculescu, C.; Persson, L.E. *Convex Functions and Their Applications*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 23, first edition published in 2006.
10. Billingsley, P. *Probability and Measure*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
11. Barndorff-Nielsen, O. *Information and Exponential Families*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
12. Morris, C.N. Natural exponential families with quadratic variance functions. *Ann. Stat.* **1982**, *10*, 65–80. [CrossRef]
13. Efron, B. *Exponential Families in Theory and Practice*; Cambridge University Press: Cambridge, UK, 2022.
14. Grünwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
15. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
16. Wainwright, M.J.; Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends® Mach. Learn.* **2008**, *1*, 1–305.
17. LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F. A tutorial on energy-based learning. In *Predicting Structured Data*; University of Toronto: Toronto, ON, USA, 2006; Volume 1.
18. Kindermann, R.; Snell, J.L. *Markov Random Fields and Their Applications*; American Mathematical Society: Providence, RI, USA, 1980; Volume 1.
19. Dai, B.; Liu, Z.; Dai, H.; He, N.; Gretton, A.; Song, L.; Schuurmans, D. Exponential family estimation via adversarial dynamics embedding. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2019; Volume 32.
20. Cobb, L.; Koppstein, P.; Chen, N.H. Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. Am. Stat. Assoc.* **1983**, *78*, 124–130. [CrossRef]

21. Garcia, V.; Nielsen, F. Simplification and hierarchical representations of mixtures of exponential families. *Signal Process.* **2010**, *90*, 3197–3212. [CrossRef]
22. Zhang, J.; Wong, T.K.L. λ-Deformed probability families with subtractive and divisive normalizations. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2021; Volume 45, pp. 187–215.
23. Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
24. Wong, T.K.L. Logarithmic divergences from optimal transport and Rényi geometry. *Inf. Geom.* **2018**, *1*, 39–78. [CrossRef]
25. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
26. Azoury, K.S.; Warmuth, M.K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.* **2001**, *43*, 211–246. [CrossRef]
27. Amari, S.I. *Differential-Geometrical Methods in Statistics*, 1st ed.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 28.
28. Nielsen, F. Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences. *Entropy* **2022**, *24*, 421. [CrossRef] [PubMed]
29. Del Castillo, J. The singly truncated normal distribution: A non-steep exponential family. *Ann. Inst. Stat. Math.* **1994**, *46*, 57–66. [CrossRef]
30. Wainwright, M.J.; Jaakkola, T.S.; Willsky, A.S. A new class of upper bounds on the log partition function. *IEEE Trans. Inf. Theory* **2005**, *51*, 2313–2335. [CrossRef]
31. Hyvärinen, A.; Dayan, P. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **2005**, *6*, 695–709.
32. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
33. Eguchi, S.; Komori, O. *Minimum Divergence Methods in Statistical Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2022.
34. Kolmogorov, A. *Sur la Notion de la Moyenne*; Cold Spring Harbor Laboratory: Cold Spring Harbor, NY, USA, 1930.
35. Komori, O.; Eguchi, S. A unified formulation of k-Means, fuzzy c-Means and Gaussian mixture model by the Kolmogorov–Nagumo average. *Entropy* **2021**, *23*, 518. [CrossRef]
36. Aczél, J. A generalization of the notion of convex functions. *Det K. Nor. Vidensk. Selsk. Forh. Trondheim* **1947**, *19*, 87–90.
37. Nielsen, F.; Nock, R. Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Process. Lett.* **2017**, *24*, 1123–1127. [CrossRef]
38. Bauschke, H.H.; Goebel, R.; Lucet, Y.; Wang, X. The proximal average: Basic theory. *SIAM J. Optim.* **2008**, *19*, 766–785. [CrossRef]
39. Rockafellar, R.T. Conjugates and Legendre transforms of convex functions. *Can. J. Math.* **1967**, *19*, 200–205. [CrossRef]
40. Shima, H. *The Geometry of Hessian Structures*; World Scientific: Singapore, 2007.
41. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391. [CrossRef]
42. Rockafellar, R. *Convex Analysis*; Princeton Landmarks in Mathematics and Physics; Princeton University Press: Princeton, NJ, USA, 1997.
43. Yoshizawa, S.; Tanabe, K. Dual differential geometry associated with the Kullbaek-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT J. Math.* **1999**, *35*, 113–137. [CrossRef]
44. Hougaard, P. *Convex Functions in Exponential Families*; Department of Mathematical Sciences, University of Copenhagen: Copenhagen, Denmark, 1983.
45. Brekelmans, R.; Nielsen, F. Variational representations of annealing paths: Bregman information under monotonic embeddings. *Inf. Geom.* **2024**. [CrossRef]
46. Amari, S.I. α-Divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931. [CrossRef]
47. Hennequin, R.; David, B.; Badeau, R. Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Process. Lett.* **2010**, *18*, 83–86. [CrossRef]
48. Ohara, A.; Eguchi, S. Group invariance of information geometry on q-Gaussian distributions induced by Beta-divergence. *Entropy* **2013**, *15*, 4732–4747. [CrossRef]
49. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J.; Lafferty, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
50. Frongillo, R.; Reid, M.D. Convex Found. *Gen. Maxent Model.* **2014**, *1636*, 11–16.
51. Ishige, K.; Salani, P.; Takatsu, A. Hierarchy of deformations in concavity. *Inf. Geom.* **2022**, *7*, 251–269. [CrossRef]
52. Zhang, J.; Wong, T.K.L. λ-Deformation: A canonical framework for statistical manifolds of constant curvature. *Entropy* **2022**, *24*, 193. [CrossRef] [PubMed]
53. Jenssen, R.; Principe, J.C.; Erdogmus, D.; Eltoft, T. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *J. Frankl. Inst.* **2006**, *343*, 614–629. [CrossRef]