entropy

MDPI

Article

Statistical Divergences between Densities of Truncated Exponential Families with Nested Supports: Duo Bregman and Duo Jensen Divergences

Frank Nielsen

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

Abstract: By calculating the Kullback–Leibler divergence between two probability measures belonging to different exponential families dominated by the same measure, we obtain a formula that generalizes the ordinary Fenchel–Young divergence. Inspired by this formula, we define the duo Fenchel–Young divergence and report a majorization condition on its pair of strictly convex generators, which guarantees that this divergence is always non-negative. The duo Fenchel–Young divergence is also equivalent to a duo Bregman divergence. We show how to use these duo divergences by calculating the Kullback–Leibler divergence between densities of truncated exponential families with nested supports, and report a formula for the Kullback–Leibler divergence between truncated normal distributions. Finally, we prove that the skewed Bhattacharyya distances between truncated exponential families amount to equivalent skewed duo Jensen divergences.

Keywords: exponential family; statistical divergence; truncated exponential family; truncated normal distributions

Citation: Nielsen, F. Statistical Divergences between Densities of Truncated Exponential Families with Nested Supports: Duo Bregman and Duo Jensen Divergences. *Entropy* **2022**, *24*, 421. https://doi.org/10.3390/e24030421

Academic Editors: Karagrigoriou Alexandros and Makrides Andreas

Received: 2 March 2022
Accepted: 16 March 2022
Published: 17 March 2022

1. Introduction

1.1. Exponential Families

Let $(\mathcal{X}, \Sigma)$ be a measurable space, and consider a regular minimal exponential family [1] $\mathcal{E}$ of probability measures P_θ all dominated by a base measure μ ($P_\theta \ll \mu$):

$$\mathcal{E} = \{P_\theta \ : \ \theta \in \Theta\}. \tag{1}$$

The Radon–Nikodym derivatives or densities of the probability measures P_θ with respect to μ can be written canonically as

$$p_\theta(x) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x) = \exp\Big(\theta^\top t(x) - F(\theta) + k(x)\Big), \tag{2}$$

where θ denotes the natural parameter, $t(x)$ the sufficient statistic [1–4], and $F(\theta)$ the log-normalizer [1] (or cumulant function). The optional auxiliary term $k(x)$ allows us to change the base measure μ into the measure ν such that $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) = e^{k(x)}$. The order D of the family is the dimension of the natural parameter space Θ:

$$\Theta = \left\{\theta \in \mathbb{R}^D \ : \ \int_{\mathcal{X}} \exp\Big(\theta^\top t(x) + k(x)\Big)\, \mathrm{d}\mu(x) < \infty\right\}, \tag{3}$$

where $\mathbb{R}$ denotes the set of reals. The sufficient statistic $t(x) = (t_1(x), \ldots, t_D(x))$ is a vector of D functions. The sufficient statistic $t(x)$ is said to be minimal when the $D+1$ functions $1, t_1(x), \ldots, t_D(x)$ are linearly independent [1]. The sufficient statistics $t(x)$ are such that the probability $\Pr[X|\theta] = \Pr[X|t(X)]$. That is, all information necessary for the statistical inference of parameter θ is contained in $t(X)$. Exponential families are characterized as

families of parametric distributions with finite-dimensional sufficient statistics [1]. Exponential families $\{p_\lambda\}$ include among others the exponential, normal, gamma/beta, inverse gamma, inverse Gaussian, and Wishart distributions once a reparameterization $\theta = \theta(\lambda)$ of the parametric distributions $\{p_\lambda\}$ is performed to reveal their natural parameters [1].

When the sufficient statistic $t(x)$ is x, these exponential families [1] are called natural exponential families or tilted exponential families [5] in the literature. Indeed, the distributions P_θ of the exponential family $\mathcal{E}$ can be interpreted as distributions obtained by tilting the base measure μ [6]. In this paper, we consider either discrete exponential families like the family of Poisson distributions (univariate distributions of order $D = 1$ with respect to the counting measure) or continuous exponential families like the family of normal distributions (univariate distributions of order $D = 2$ with respect to the Lebesgue measure). The Radon–Nikodym derivative of a discrete exponential family is a probability mass function (pmf), and the Radon–Nikodym derivative of a continuous exponential family is a probability density function (pdf). The support of a pmf $p(x)$ is $\mathrm{supp}(p) = \{x \in \mathbb{Z} \,:\, p(x) > 0\}$ (where $\mathbb{Z}$ denotes the set of integers) and the support of a d-variate pdf $p(x)$ is $\mathrm{supp}(p) = \{x \in \mathbb{R}^d \,:\, p(x) > 0\}$. The Poisson distributions have support $\mathbb{N} \cup \{0\}$ where $\mathbb{N}$ denotes the set of natural numbers $\{1, 2, \ldots, \}$. Densities of an exponential family all have coinciding support [1].

1.2. Truncated Exponential Families with Nested Supports

In this paper, we shall consider truncated exponential families [7] with nested supports. A truncated exponential family is a set of parametric probability distributions obtained by truncation of the support of an exponential family. Truncated exponential families are exponential families but their statistical inference is more subtle [8,9]. Let $\mathcal{E}_{\mathrm{Trunc}} = \{q_\theta\}$ be a truncated exponential family of $\mathcal{E} = \{p_\theta\}$ with nested supports $\mathrm{supp}(q_\theta) \subset \mathrm{supp}(p_\theta)$. The canonical decompositions of densities p_θ and q_θ have the following expressions:

$$p_\theta(x) = \exp\Big(\theta^\top t(x) + k(x) - F(\theta)\Big), \tag{4}$$

$$q_\theta(x) = \frac{p_\theta(x)}{Z^{\mathcal{X}_{\mathrm{Trunc}}}(\theta)} = \exp\Big(\theta^\top t(x) + k(x) - F_{\mathrm{Trunc}}(\theta)\Big), \tag{5}$$

where the log-normalizer of the truncated exponential family is:

$$F_{\mathrm{Trunc}}(\theta) = F(\theta) + \log Z^{\mathcal{X}_{\mathrm{Trunc}}}(\theta), \tag{6}$$

where $Z^{\mathcal{X}_{\mathrm{Trunc}}}(\theta)$ is a normalizing term that takes into account the truncated support $\mathcal{X}_{\mathrm{Trunc}}$. These equations show that densities of truncated exponential families only differ by their log-normalizer functions. Let $\mathcal{X}_{\mathrm{Trunc}}$ denote the support of the distributions of $\mathcal{E}_{\mathrm{Trunc}} = \mathrm{supp}(q_\theta)$ and $\mathcal{X} = \mathrm{supp}(p_\theta)$ the support of $\mathcal{E}$. Family $\mathcal{E}_{\mathrm{Trunc}}$ is a truncated exponential family of $\mathcal{E}$ that can be notationally written as $\mathcal{E}_{\mathcal{X}_{\mathrm{Trunc}}}$. Family $\mathcal{E}$ can also be interpreted as the (un)truncated exponential family $\mathcal{E}_{\mathcal{X}}$ with densities $p_\theta^{\mathcal{X}} = p_\theta$. A truncated exponential family $\mathcal{E}_{\mathcal{X}_{\mathrm{Trunc}}}$ of $\mathcal{E}$ is said to have nested support when $\mathcal{X}_{\mathrm{Trunc}} \subset \mathcal{X}$. For example, the family of half-normal distributions defined on the support $\mathcal{X}_{\mathrm{Trunc}} = [0, \infty)$ is a nested truncated exponential family of the family of normal distributions defined on the support $\mathcal{X} = (-\infty, \infty)$.

1.3. Kullback–Leibler Divergence Between Exponential Family Distributions

For two σ-finite probability measures P and Q on $(\mathcal{X}, \Sigma)$ such that P is dominated by Q ($P \ll Q$), the Kullback–Leibler divergence between P and Q is defined by

$$D_{\mathrm{KL}}[P : Q] = \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P = E_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right], \tag{7}$$

where $E_P[X]$ denotes the expectation of a random variable $X \sim P$ [10]. When $P \not\ll Q$, we set $D_{\mathrm{KL}}[P : Q] = +\infty$. Gibbs' inequality [11] $D_{\mathrm{KL}}[P : Q] \geq 0$ shows that the Kullback–Leibler

divergence (KLD for short) is always non-negative. The proof of Gibbs' inequality relies on Jensen's inequality and holds for the wide class of f-divergences [12] induced by convex generators $f(u)$:

$$I_f[P:Q] = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}P \geq f\left(\int_{\mathcal{X}} \frac{\mathrm{d}Q}{\mathrm{d}P}\, \mathrm{d}P\right) \geq f(1). \tag{8}$$

The KLD is an f-divergence obtained for the convex generator $f(u) = -\log u$.

1.4. Kullback–Leibler Divergence Between Exponential Family Densities

It is well-known that the KLD between two distributions P_{θ_1} and P_{θ_2} of $\mathcal{E}$ amounts to computing an equivalent Fenchel–Young divergence [13]:

$$D_{\mathrm{KL}}[P_{\theta_1} : P_{\theta_2}] = \int_{\mathcal{X}} p_{\theta_1}(x) \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\, \mathrm{d}\mu(x) = Y_{F,F^*}(\theta_2, \eta_1), \tag{9}$$

where $\eta = \nabla F(\theta) = E_{P_\theta}[t(x)]$ is the moment parameter [1] and

$$\nabla F(\theta) = \left[\frac{\partial}{\partial \theta_1} F(\theta), \ldots, \frac{\partial}{\partial \theta_D} F(\theta)\right]^\top, \tag{10}$$

is the gradient of F with respect to $\theta = [\theta_1, \ldots, \theta_D]^\top$. The Fenchel–Young divergence is defined for a pair of strictly convex conjugate functions [14] $F(\theta)$ and $F^*(\eta)$ related by the Legendre–Fenchel transform by

$$Y_{F,F^*}(\theta_1, \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2. \tag{11}$$

Amari (1985) first introduced this formula as the canonical divergence of dually flat spaces in information geometry [15] (Equation 3.21), and proved that the Fenchel–Young divergence is obtained as the KLD between densities belonging to the same exponential family [15] (Theorem 3.7). Azoury and Warmuth expressed the KLD $D_{\mathrm{KL}}[P_{\theta_1} : P_{\theta_2}]$ using dual Bregman divergences in [13] (2001):

$$D_{\mathrm{KL}}[P_{\theta_1} : P_{\theta_2}] = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2), \tag{12}$$

where a Bregman divergence [16] $B_F(\theta_1 : \theta_2)$ is defined for a strictly convex and differentiable generator $F(\theta)$ by:

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2). \tag{13}$$

Acharyya termed the divergence Y_{F,F^*} the Fenchel–Young divergence in his PhD thesis [17] (2013), and Blondel et al. called such divergences Fenchel–Young losses (2020) in the context of machine learning [18] (Equation (9) in Definition 2). This term was also used by the author the Legendre–Fenchel divergence in [19]. The Fenchel–Young divergence stems from the Fenchel–Young inequality [14,20]:

$$F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2, \tag{14}$$

with equality if and only if $\eta_2 = \nabla F(\theta_1)$.

Figure 1 visualizes the 1D Fenchel–Young divergence and gives a geometric proof that $Y_{F,F^*}(\theta_1, \eta_2) \geq 0$ with equality if and only if $\eta_2 = F'(\theta_1)$. Indeed, by considering the behavior of the Legendre–Fenchel transformation under translations:

- if $F_t(\theta) = F(\theta + t)$ then $F_t^*(\eta) = F^*(\eta) - \eta^\top t$ for all $t \in \mathbb{R}$, and
- if $F_\lambda(\theta) = F(\theta) + \lambda$ then $F_\lambda^*(\eta) = F^*(\eta) - \lambda$ for all $\lambda \in \mathbb{R}$,

we may assume without loss of generality that $F(0) = 0$. The function $F'(\theta)$ is strictly increasing and continuous since $F(\theta)$ is a strictly convex and differentiable convex function. Thus we have $F(\theta) = \int_0^\theta F'(\theta)\,\mathrm{d}\theta$ and $F^*(\eta) = \int_0^\eta F^{*\prime}(\eta)\,\mathrm{d}\eta = \int_0^\eta F'^{-1}(\eta)\,\mathrm{d}\eta$.

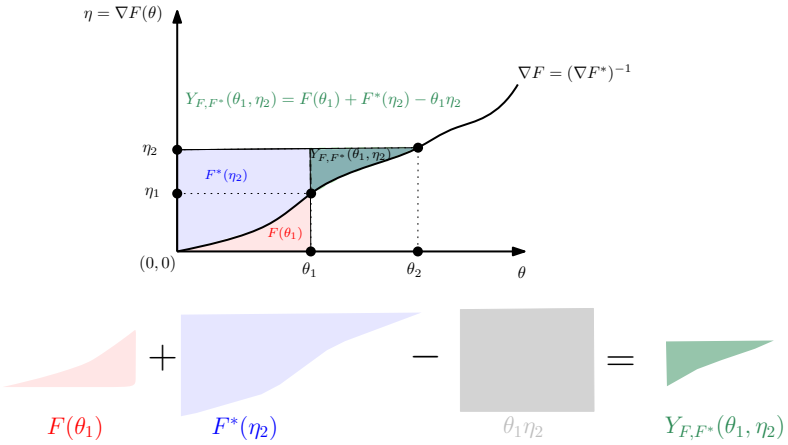


Figure 1. Visualizing the Fenchel–Young divergence.

The Bregman divergence $B_F(\theta_1 : \theta_2)$ amounts to a dual Bregman divergence [13] between the dual parameters with swapped order: $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$ where $\eta_i = \nabla F(\theta_i)$ for $i \in \{1, 2\}$. Thus the KLD between two distributions P_{θ_1} and P_{θ_2} of $\mathcal{E}$ can be expressed equivalently as follows:

$$D_{\mathrm{KL}}[P_{\theta_1} : P_{\theta_2}] = Y_{F,F^*}(\theta_2 : \eta_1) = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2) = Y_{F^*,F}(\eta_1 : \eta_2). \tag{15}$$

The symmetrized Kullback–Leibler divergence $D_J[P_{\theta_1} : P_{\theta_2}]$ between two distributions P_{θ_1} and P_{θ_2} of $\mathcal{E}$ is called Jeffreys' divergence [21] and amounts to a symmetrized Bregman divergence [22]:

$$\begin{aligned} D_J[P_{\theta_1} : P_{\theta_2}] &= D_{\mathrm{KL}}[P_{\theta_1} : P_{\theta_2}] + D_{\mathrm{KL}}[P_{\theta_2} : P_{\theta_1}], & (16)\\ &= B_F(\theta_2 : \theta_1) + B_F(\theta_1 : \theta_2), & (17)\\ &= (\theta_2 - \theta_1)^\top (\eta_2 - \eta_1) := S_F(\theta_1, \theta_2). & (18) \end{aligned}$$

Note that the Bregman divergence $B_F(\theta_1 : \theta_2)$ can also be interpreted as a surface area:

$$B_F(\theta_1 : \theta_2) = \int_{\theta_2}^{\theta_1} (F'(\theta) - F'(\theta_2))\mathrm{d}\theta. \tag{19}$$

Figure 2 illustrates the sided and symmetrized Bregman divergences.

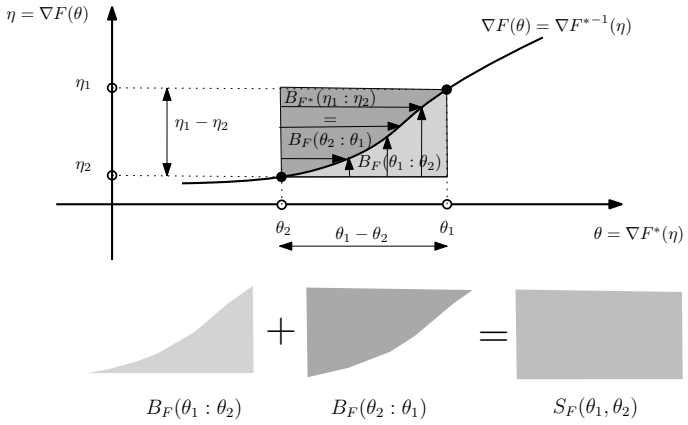


Figure 2. Visualizing the sided and symmetrized Bregman divergences.

1.5. Contributions and Paper Outline

We recall in Section 2 the formula obtained for the Kullback–Leibler divergence between two exponential family densities equivalent to each other [23] (Equation (29)). Inspired by this formula, we give a definition of the duo Fenchel–Young divergence induced by a pair of strictly convex functions F_1 and F_2 (Definition 1) in Section 3, and prove that the divergence is always non-negative provided that F_1 upper bounds F_2. We then define the duo Bregman divergence (Definition 2) corresponding to the duo Fenchel–Young divergence. In Section 4, we show that the Kullback–Leibler divergence between a truncated density and a density of a same parametric exponential family amounts to a duo Fenchel–Young divergence or equivalently to a duo Bregman divergence on swapped parameters (Theorem 1). That is, we consider a truncated exponential family [7] $\mathcal{E}_1$ of an exponential family $\mathcal{E}_1$ such that the common support of the distributions of $\mathcal{E}_1$ is contained in the common support of the distributions of $\mathcal{E}_2$ and both canonical decompositions of the families coincide (see Equation (2)). In particular, when $\mathcal{E}_2$ is also a truncated exponential family of $\mathcal{E}$, then we express the KLD between two truncated distributions as a duo Bregman divergence. As examples, we report the formula for the Kullback–Leibler divergence between two densities of truncated exponential families (Corollary 1), and illustrate the formula for the Kullback–Leibler divergence between truncated exponential distributions (Example 6) and for the Kullback–Leibler divergence between truncated normal distributions (Example 7).

In Section 5, we further consider the skewed Bhattacharyya distance between densities of truncated exponential families and prove that it amounts to a duo Jensen divergence (Theorem 2). Finally, we conclude in Section 6.

2. Kullback–Leibler Divergence Between Different Exponential Families

Consider now two exponential families [1] $\mathcal{P}$ and $\mathcal{Q}$ defined by their Radon–Nikodym derivatives with respect to two positive measures $\mu_{\mathcal{P}}$ and $\mu_{\mathcal{Q}}$ on $(\mathcal{X}, \Sigma)$:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \tag{20}$$
$$\mathcal{Q} = \{Q_{\theta'} : \theta' \in \Theta'\}. \tag{21}$$

The corresponding natural parameter spaces are

$$\Theta = \left\{\theta \in \mathbb{R}^D : \int_{\mathcal{X}} \exp(\theta^\top t_{\mathcal{P}}(x) + k_{\mathcal{P}}(x))\, \mathrm{d}\mu_{\mathcal{P}}(x) < \infty\right\}, \tag{22}$$
$$\Theta' = \left\{\theta' \in \mathbb{R}^{D'} : \int_{\mathcal{X}} \exp({\theta'}^\top t_{\mathcal{Q}}(x) + k_{\mathcal{Q}}(x))\, \mathrm{d}\mu_{\mathcal{Q}}(x) < \infty\right\}, \tag{23}$$

The order of $\mathcal{P}$ is D, $t_{\mathcal{P}}(x)$ denotes the sufficient statistics of P_θ, and $k_{\mathcal{P}}(x)$ is a term to adjust/tilt the base measure $\mu_{\mathcal{P}}$. Similarly, the order of $\mathcal{Q}$ is D', $t_{\mathcal{Q}}(x)$ denotes the sufficient statistics of $Q_{\theta'}$, and $k_{\mathcal{Q}}(x)$ is an optional term to adjust the base measure $\mu_{\mathcal{Q}}$. Let p_θ and $q_{\theta'}$ denote the Radon–Nikodym derivatives with respect to the measures $\mu_{\mathcal{P}}$ and $\mu_{\mathcal{Q}}$, respectively:

$$p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu_{\mathcal{P}}} = \exp(\theta^\top t_{\mathcal{P}}(x) - F_{\mathcal{P}}(\theta) + k_{\mathcal{P}}(x)), \tag{24}$$
$$q_{\theta'} = \frac{\mathrm{d}Q_{\theta'}}{\mathrm{d}\mu_{\mathcal{Q}}} = \exp({\theta'}^\top t_{\mathcal{Q}}(x) - F_{\mathcal{Q}}(\theta') + k_{\mathcal{Q}}(x)), \tag{25}$$

where $F_{\mathcal{P}}(\theta)$ and $F_{\mathcal{Q}}(\theta')$ denote the corresponding log-normalizers of $\mathcal{P}$ and $\mathcal{Q}$, respectively.

$$F_{\mathcal{P}}(\theta) = \log\left(\int \exp(\theta^\top t_{\mathcal{P}}(x) + k_{\mathcal{P}}(x))\, \mathrm{d}\mu_{\mathcal{P}}(x)\right), \tag{26}$$
$$F_{\mathcal{Q}}(\theta) = \log\left(\int \exp(\theta^\top t_{\mathcal{Q}}(x) + k_{\mathcal{Q}}(x))\, \mathrm{d}\mu_{\mathcal{Q}}(x)\right). \tag{27}$$

The functions $F_{\mathcal{P}}$ and $F_{\mathcal{Q}}$ are strictly convex and real analytic [1]. Hence, those functions are infinitely many times differentiable on their open natural parameter spaces.

Consider the KLD between $P_\theta \in \mathcal{P}$ and $Q_{\theta'} \in \mathcal{Q}$ such that $\mu_{\mathcal{P}} = \mu_{\mathcal{Q}}$ (and hence $P_\theta \ll Q_{\theta'}$). Then the KLD between P_θ and $Q_{\theta'}$ was first considered in [23]:

$$\begin{aligned} D_{\mathrm{KL}}[P_\theta : Q_{\theta'}] &= E_P\left[\log\left(\frac{\mathrm{d}P_\theta}{\mathrm{d}Q_{\theta'}}\right)\right], \quad (28)\\ &= E_{P_\theta}\left[\left(\theta^\top t_{\mathcal{P}}(x) - {\theta'}^\top t_{\mathcal{Q}}(x) - F_{\mathcal{P}}(\theta) + F_{\mathcal{Q}}(\theta') + k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)\right) \underbrace{\frac{\mathrm{d}\mu_{\mathcal{P}}}{\mathrm{d}\mu_{\mathcal{Q}}}}_{=1}\right],\\ &= F_{\mathcal{Q}}(\theta') - F_{\mathcal{P}}(\theta) + \theta^\top E_{P_\theta}[t_{\mathcal{P}}(x)] - {\theta'}^\top E_{P_\theta}[t_{\mathcal{Q}}(x)] + E_{P_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]. \end{aligned}$$

Recall that the dual parameterization of an exponential family density P_θ is P^η with $\eta = E_{P_\theta}[t_{\mathcal{P}}(x)] = \nabla F_{\mathcal{P}}(\theta)$ [1], and that the Fenchel–Young equality is $F(\theta) + F^*(\eta) = \theta^\top \eta$ for $\eta = \nabla F(\theta)$. Thus the KLD between P_θ and $Q_{\theta'}$ can be rewritten as

$$D_{\mathrm{KL}}[P_\theta : Q_{\theta'}] = F_{\mathcal{Q}}(\theta') + F^*_{\mathcal{P}}(\eta) - {\theta'}^\top E_{P_\theta}[t_{\mathcal{Q}}(x)] + E_{P_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]. \quad (29)$$

This formula was reported in [23] and generalizes the Fenchel–Young divergence [17] obtained when $\mathcal{P} = \mathcal{Q}$ (with $t_{\mathcal{P}}(x) = t_{\mathcal{Q}}(x)$, $k_{\mathcal{P}}(x) = k_{\mathcal{Q}}(x)$, and $F(\theta) = F_{\mathcal{P}}(\theta) = F_{\mathcal{Q}}(\theta)$ and $F^*(\eta) = F^*_{\mathcal{P}}(\eta) = F^*_{\mathcal{Q}}(\eta)$).

The formula of Equation (29) was illustrated in [23] with two examples: the KLD between Laplacian distributions and zero-centered Gaussian distributions, and the KLD between two Weibull distributions. Both these examples use the Lebesgue base measure for $\mu_{\mathcal{P}}$ and $\mu_{\mathcal{Q}}$.

Let us report another example that uses the counting measure as the base measure for $\mu_{\mathcal{P}}$ and $\mu_{\mathcal{Q}}$.

Example 1. *Consider the KLD between a Poisson probability mass function (pmf) and a geometric pmf. The canonical decompositions of the Poisson and geometric pmfs are summarized in Table 1. The KLD between a Poisson pmf p_λ and a geometric pmf q_p is equal to*

$$\begin{aligned} D_{\mathrm{KL}}[P_\lambda : Q_p] &= F_{\mathcal{Q}}(\theta') + F^*_{\mathcal{P}}(\eta) - E_{P_\theta}[t_{\mathcal{Q}}(x)] \cdot \theta' + E_{P_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)], \quad (30)\\ &= -\log p + \lambda \log \lambda - \lambda - \lambda \log(1-p) - E_{P_\lambda}[\log x!] \quad (31) \end{aligned}$$

Since $E_{p_\lambda}[-\log x!] = -\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k \log(k!)}{k!}$, we have

$$D_{\mathrm{KL}}[P_\lambda : Q_p] = -\log p + \lambda \log \frac{\lambda}{1-p} - \lambda - \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k \log(k!)}{k!}. \quad (32)$$

Note that we can calculate the KLD between two geometric distributions Q_{p_1} and Q_{p_2} as

$$\begin{aligned} D_{\mathrm{KL}}[Q_{p_1} : Q_{p_2}] &= B_{F_{\mathcal{Q}}}(\theta(p_2) : \theta(p_1)), \quad (33)\\ &= F_{\mathcal{Q}}(\theta(p_2)) - F_{\mathcal{Q}}(\theta(p_1)) - (\theta(p_2) - \theta(p_1))\eta(p_1), \quad (34) \end{aligned}$$

We obtain:

$$D_{\mathrm{KL}}[Q_{p_1} : Q_{p_2}] = \log\left(\frac{p_1}{p_2}\right) - \left(1 - \frac{1}{p_1}\right) \log \frac{1-p_1}{1-p_2}.$$

Table 1. Canonical decomposition of the Poisson and the geometric discrete exponential families.

Quantity	Poisson Family $\mathcal{P}$	Geometric Family $\mathcal{Q}$
support	$\mathbb{N} \cup \{0\}$	$\mathbb{N} \cup \{0\}$
base measure	counting measure	counting measure
ordinary parameter	rate $\lambda > 0$	success probability $p \in (0,1)$
pmf	$\frac{\lambda^x}{x!}\exp(-\lambda)$	$(1-p)^x p$
sufficient statistic	$t_{\mathcal{P}}(x) = x$	$t_{\mathcal{Q}}(x) = x$
natural parameter	$\theta(\lambda) = \log\lambda$	$\theta(p) = \log(1-p)$
cumulant function	$F_{\mathcal{P}}(\theta) = \exp(\theta)$	$F_{\mathcal{Q}}(\theta) = -\log(1-\exp(\theta))$
	$F_{\mathcal{P}}(\lambda) = \lambda$	$F_{\mathcal{Q}}(p) = -\log(p)$
auxiliary term	$k_{\mathcal{P}}(x) = -\log x!$	$k_{\mathcal{Q}}(x) = 0$
moment $\eta = E[t(x)]$	$\eta = \lambda$	$\eta = \frac{e^\theta}{1-e^\theta} = \frac{1}{p} - 1$
negentropy	$F^*_{\mathcal{P}}(\eta(\lambda)) = \lambda\log\lambda - \lambda$	$F^*_{\mathcal{Q}}(\eta(p)) = \left(1 - \frac{1}{p}\right)\log(1-p) + \log p$
$(F^*(\eta) = \theta\cdot\eta - F(\theta))$		

3. The Duo Fenchel–Young Divergence and Its Corresponding Duo Bregman Divergence

Inspired by formula of Equation (29), we shall define the *duo Fenchel–Young divergence* using a *dominance condition* on a pair $(F_1(\theta), F_2(\theta))$ of strictly convex generators.

Definition 1 (duo Fenchel–Young divergence)**.** *Let $F_1(\theta)$ and $F_2(\theta)$ be two strictly convex functions such that $F_1(\theta) \geq F_2(\theta)$ for any $\theta \in \Theta_{12} = \mathrm{dom}(F_1) \cap \mathrm{dom}(F_2)$. Then the duo Fenchel–Young divergence $Y_{F_1,F_2^*}(\theta,\eta')$ is defined by*

$$Y_{F_1,F_2^*}(\theta,\eta') := F_1(\theta) + F_2^*(\eta') - \theta^\top\eta'. \tag{35}$$

When $F_1(\theta) = F_2(\theta) =: F(\theta)$, we have $F_1^*(\eta) = F_2^*(\eta) =: F^*(\eta)$, and we retrieve the ordinary Fenchel–Young divergence [17]:

$$Y_{F,F^*}(\theta,\eta') := F(\theta) + F^*(\eta') - \theta^\top\eta' \geq 0. \tag{36}$$

Note that in Equation (35), we have $\eta' = \nabla F_2(\theta')$.

Property 1 (Non-negative duo Fenchel–Young divergence)**.** *The duo Fenchel–Young divergence is always non-negative.*

Proof. The proof relies on the reverse dominance property of strictly convex and differentiable conjugate functions:

Lemma 1 (Reverse majorization order of functions by the Legendre–Fenchel transform)**.** *Let $F_1(\theta)$ and $F_2(\theta)$ be two Legendre-type convex functions [14]. Then if $F_1(\theta) \geq F_2(\theta)$ then we have $F_2^*(\eta) \geq F_1^*(\eta)$.*

Proof. This property is graphically illustrated in Figure 3. The reverse dominance property of the Legendre–Fenchel transformation can be checked algebraically as follows:

$$\begin{aligned}
F_1^*(\eta) &= \sup_{\theta\in\Theta}\{\eta^\top\theta - F_1(\theta)\}, && (37)\\
&= \eta^\top\theta_1 - F_1(\theta_1) \qquad (\text{with } \eta = \nabla F_1(\theta_1)), && (38)\\
&\leq \eta^\top\theta_1 - F_2(\theta_1), && (39)\\
&\leq \sup_{\theta\in\Theta}\{\eta^\top\theta - F_2(\theta)\} = F_2^*(\eta). && (40)
\end{aligned}$$

□

Thus we have $F_1^*(\eta) \leq F_2^*(\eta)$ when $F_1(\theta) \geq F_2(\theta)$. Therefore it follows that $Y_{F_1,F_2^*}(\theta,\eta') \geq 0$ since we have

$$\begin{aligned} Y_{F_1,F_2^*}(\theta,\eta') &:= F_1(\theta) + F_2^*(\eta') - \theta^\top \eta', && (41)\\ &\geq F_1(\theta) + F_1^*(\eta') - \theta^\top \eta' = Y_{F_1,F_1^*}(\theta,\eta') \geq 0, && (42) \end{aligned}$$

where Y_{F_1,F_1^*} is the ordinary Fenchel–Young divergence, which is guaranteed to be non-negative from the Fenchel–Young inequality. □

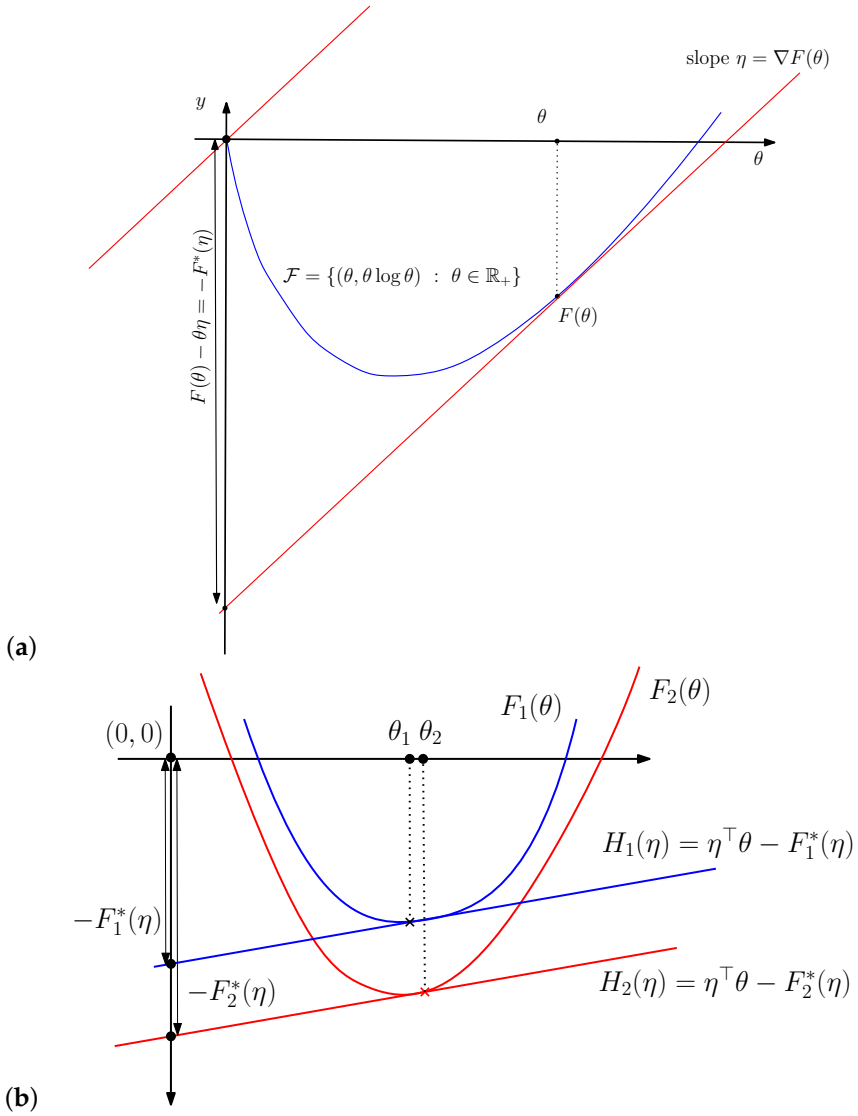


Figure 3. (**a**) Visual illustration of the Legendre–Fenchel transformation: $F^*(\eta)$ is measured as the vertical gap (left long black line with both arrows) between the origin and the hyperplane of the "slope" η tangent at $F(\theta)$ evaluated at $\theta = 0$. (**b**) The Legendre transforms $F_1^*(\eta)$ and $F_1^*(\eta)$ of two functions $F_1(\theta)$ and $F_2(\theta)$ such that $F_1(\theta) \geq F_2(\theta)$ reverse the dominance order: $F_2^*(\eta) \geq F_1^*(\eta)$.

We can express the duo Fenchel–Young divergence using the primal coordinate systems as a generalization of the Bregman divergence to two generators that we term the duo Bregman divergence (see Figure 4) :

$$B_{F_1,F_2}(\theta : \theta') := Y_{F_1,F_2^*}(\theta,\eta') = F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta'), \quad (43)$$

with $\eta' = \nabla F_2(\theta')$.

This generalized Bregman divergence is non-negative when $F_1(\theta) \geq F_2(\theta)$. Indeed, we check that

$$\begin{aligned} B_{F_1,F_2}(\theta : \theta') &= F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta'), & (44) \\ &\geq F_2(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta') = B_{F_2}(\theta : \theta') \geq 0. & (45) \end{aligned}$$

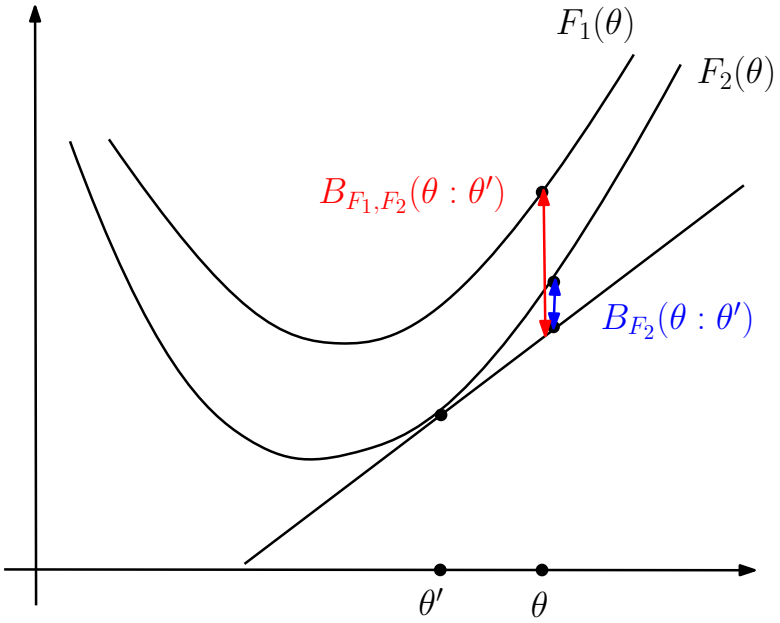


Figure 4. The duo Bregman divergence induced by two strictly convex and differentiable functions F_1 and F_2 such that $F_1(\theta) \geq F_2(\theta)$. We check graphically that $B_{F_1,F_2}(\theta : \theta') \geq B_{F_2}(\theta : \theta')$ (vertical gaps).

Definition 2 (duo Bregman divergence)**.** *Let $F_1(\theta)$ and $F_2(\theta)$ be two strictly convex functions such that $F_1(\theta) \geq F_2(\theta)$ for any $\theta \in \Theta_{12} = \mathrm{dom}(F_1) \cap \mathrm{dom}(F_2)$. Then the generalized Bregman divergence is defined by*

$$B_{F_1,F_2}(\theta : \theta') = F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta') \geq 0. \quad (46)$$

Example 2. *Consider $F_1(\theta) = \frac{a}{2}\theta^2$ for $a > 0$. We have $\eta = a\theta$, $\theta = \frac{\eta}{a}$, and*

$$F_1^*(\eta) = \frac{\eta^2}{a} - \frac{a}{2}\frac{\eta^2}{a^2} = \frac{\eta^2}{2a}. \quad (47)$$

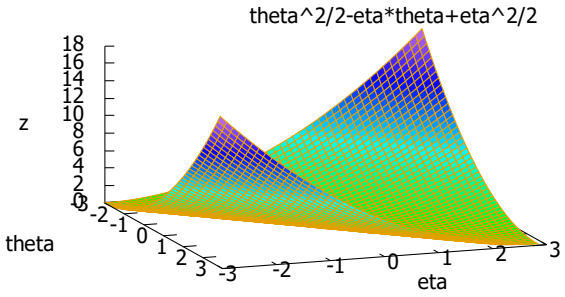
Let $F_2(\theta) = \frac{1}{2}\theta^2$ so that $F_1(\theta) \geq F_2(\theta)$ for $a \geq 1$. We check that $F_1^*(\eta) = \frac{\eta^2}{2a} \leq F_2^*(\eta)$ when $a \geq 1$. The duo Fenchel–Young divergence is

$$Y_{F_1,F_2^*}(\theta, \eta') = \frac{a}{2}\theta^2 + \frac{1}{2}\eta'^2 - \theta\eta' \geq 0, \quad (48)$$

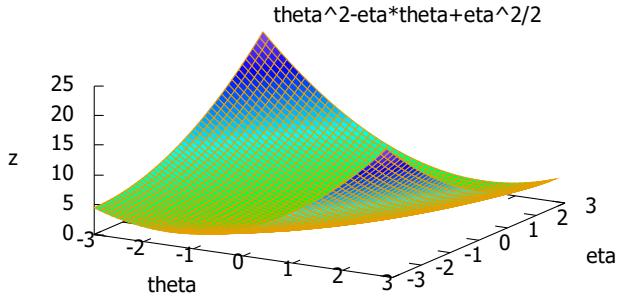
when $a \geq 1$. We can express the duo Fenchel–Young divergence in the primal coordinate systems as

$$B_{F_1,F_2}(\theta, \theta') = \frac{a}{2}\theta^2 + \frac{1}{2}\theta'^2 - \theta\theta'. \quad (49)$$

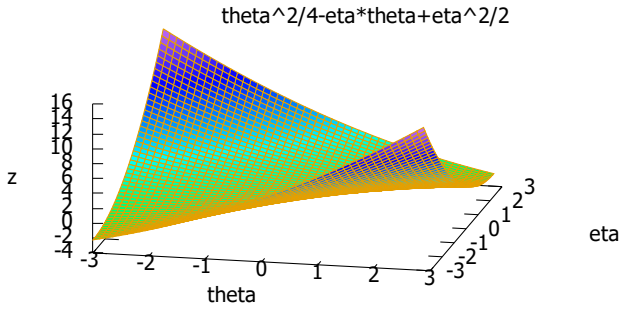
When $a = 1$, $F_1(\theta) = F_2(\theta) = \frac{1}{2}\theta^2 := F(\theta)$, and we obtain $B_F(\theta, \theta') = \frac{1}{2}\|\theta - \theta'\|_2^2$, half the squared Euclidean distance as expected. Figure 5 displays the graph plot of the duo Bregman divergence for several values of a.



(**a**)



(**b**)



(**c**)

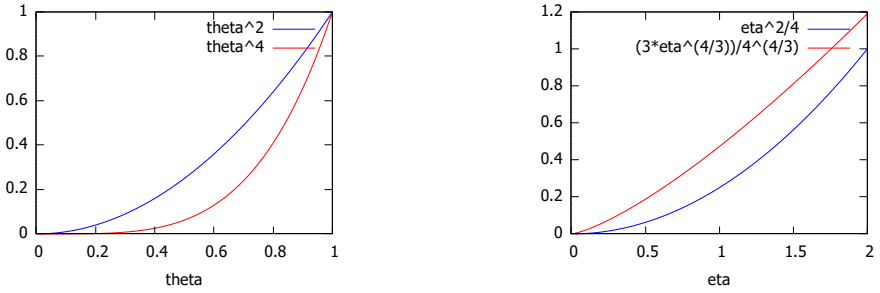
Figure 5. The duo half squared Euclidean distance $D_a^2(\theta : \theta') := \frac{a}{2}\theta^2 + \frac{1}{2}\theta'^2 - \theta\theta'$ is non-negative when $a \geq 1$: (**a**) half squared Euclidean distance ($a = 1$), (**b**) $a = 2$, (**c**) $a = \frac{1}{2}$, which shows that the divergence can be negative then since $a < 1$.

Example 3. *Consider $F_1(\theta) = \theta^2$ and $F_2(\theta) = \theta^4$ on the domain $\Theta = [0, 1]$. We have $F_1(\theta) \geq F_2(\theta)$ for $\theta \in \Theta$. The convex conjugate of $F_1(\eta)$ is $F_1^*(\eta) = \frac{1}{4}\eta^2$. We have*

$$F_2^*(\eta) = \eta^{\frac{4}{3}}\left(\left(\frac{1}{4}\right)^{\frac{1}{3}} - \left(\frac{1}{4}\right)^{\frac{4}{3}}\right) = \frac{3}{4^{\frac{4}{3}}}\eta^{\frac{4}{3}} \tag{50}$$

with $\eta_2(\theta) = 4\theta^3$. Figure 6 plots the convex functions $F_1(\theta)$ and $F_2(\theta)$, and their convex conjugates $F_1^*(\eta)$ and $F_2^*(\eta)$. We observe that $F_1(\theta) \geq F_2(\theta)$ on $\theta \in [0, 1]$ and that $F_1^*(\eta) \leq F_2^*(\eta)$ on $H = [0, 2]$.

We now state a property between dual duo Bregman divergences:



Convex functions $F_1(\theta) \geq F_2(\theta)$ Conjugate functions $F_1^*(\eta) \leq F_2^*(\eta)$

Figure 6. The Legendre transform reverses the dominance ordering: $F_1(\theta) = \theta^2 \geq F_2(\theta) = \theta^4 \Leftrightarrow F_1^*(\eta) \leq F_2^*(\eta)$ for $\theta \in [0,1]$.

Property 2 (Dual duo Fenchel–Young and Bregman divergences)**.** *We have*

$$Y_{F_1,F_2^*}(\theta : \eta') = B_{F_1,F_2}(\theta : \theta') = B_{F_2^*,F_1^*}(\eta' : \eta) = Y_{F_2^*,F_1}(\eta' : \theta) \tag{51}$$

Proof. From the Fenchel–Young equalities of the inequalities, we have $F_1(\theta) = \theta^\top \eta - F_1^*(\eta)$ for $\eta = \nabla F_1(\theta)$ and $F_2(\theta') = {\theta'}^\top \eta' - F_2^*(\eta')$ with $\eta' = \nabla F_2(\theta')$. Thus we have

$$\begin{aligned}
B_{F_1,F_2}(\theta : \theta') &= F_1(\theta) - F_2(\theta') - (\theta - \theta')^\top \nabla F_2(\theta'), && (52)\\
&= \theta^\top \eta - F_1^*(\eta) - {\theta'}^\top \eta' + F_2^*(\eta') - (\theta - \theta')^\top \eta', && (53)\\
&= F_2^*(\eta') - F_1^*(\eta) - (\eta' - \eta)^\top \theta, && (54)\\
&= B_{F_2^*,F_1^*}(\eta' : \eta). && (55)
\end{aligned}$$

Recall that $F_1(\theta) \geq F_2(\theta)$ implies that $F_1^*(\eta) \leq F_2^*(\eta)$ (Lemma 1), $\theta = \nabla F_1^*(\eta)$, and therefore the dual duo Bregman divergence is non-negative:

$$\begin{aligned}
B_{F_2^*,F_1^*}(\eta' : \eta) &= F_2^*(\eta') - F_1^*(\eta) - (\eta' - \eta)^\top \theta,\\
&\geq \underbrace{F_1^*(\eta') - F_1^*(\eta) - (\eta' - \eta)^\top \nabla F_1^*(\eta)}_{B_{F_1^*}(\eta' : \eta) \geq 0}.
\end{aligned}$$

□

4. Kullback–Leibler Divergence between Distributions of Truncated Exponential Families

Let $\mathcal{E}_1 = \{P_\theta : \theta \in \Theta_1\}$ be an exponential family of distributions all dominated by μ with Radon–Nikodym density $p_\theta(x) = \exp(\theta^\top t(x) - F_1(\theta) + k(x))\, \mathrm{d}\mu(x)$ defined on the support $\mathcal{X}_1$. Let $\mathcal{E}_2 = \{Q_\theta : \theta \in \Theta_2\}$ be another exponential family of distributions all dominated by μ with Radon–Nikodym density $q_\theta(x) = \exp(\theta^\top t(x) - F_2(\theta) + k(x))\, \mathrm{d}\mu(x)$ defined on the support $\mathcal{X}_2$ such that $\mathcal{X}_1 \subseteq \mathcal{X}_2$. Let $\tilde{p}_\theta(x) = \exp(\theta^\top t(x) + k(x))\, \mathrm{d}\mu(x)$ be the common unnormalized density so that

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{Z_1(\theta)} \tag{56}$$

and

$$q_\theta(x) = \frac{\tilde{p}_\theta(x)}{Z_2(\theta)} = \frac{Z_1(\theta)}{Z_2(\theta)}\, p_\theta(x), \tag{57}$$

with $Z_1(\theta) = \exp(F_1(\theta))$ and $Z_2(\theta) = \exp(F_2(\theta))$ being the log-normalizer functions of $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively.

We have

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= \int_{\mathcal{X}_1} p_{\theta_1}(x) \log \frac{p_{\theta_1}(x)}{q_{\theta_2}(x)} \mathrm{d}\mu(x), && (58)\\ &= \int_{\mathcal{X}_1} p_{\theta_1}(x) \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \mathrm{d}\mu(x) + \int_{\mathcal{X}_1} p_{\theta_1}(x) \log \left(\frac{Z_2(\theta_2)}{Z_1(\theta_2)} \right) \mathrm{d}\mu(x), && (59)\\ &= D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] + \log Z_2(\theta_2) - \log Z_1(\theta_2). && (60) \end{aligned}$$

Since $D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_{F_1}(\theta_2 : \theta_1)$ and $\log Z_i(\theta) = F_i(\theta)$, we obtain

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= B_{F_1}(\theta_2 : \theta_1) + F_2(\theta_2) - F_1(\theta_2), && (61)\\ &= F_1(\theta_2) - F_1(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F_1(\theta_1) + F_2(\theta_2) - F_1(\theta_2), && (62)\\ &= F_2(\theta_2) - F_1(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F_1(\theta_1) =: B_{F_2,F_1}(\theta_2 : \theta_1). && (63) \end{aligned}$$

Observe that since $\mathcal{X}_1 \subseteq \mathcal{X}2$, we have:

$$F_2(\theta) = \log \int_{\mathcal{X}_2} \tilde{p}_\theta(x) \, \mathrm{d}\mu(x) \geq \log \int_{\mathcal{X}_1} \tilde{p}_\theta(x) \, \mathrm{d}\mu(x) := F_1(\theta). \quad (64)$$

Therefore $\Theta_2 \subseteq \Theta_1$, and the common natural parameter space is $\Theta_{12} = \Theta_1 \cap \Theta_2 = \Theta_2$.

Notice that the reverse Kullback–Leibler divergence $D^*_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = D_{\mathrm{KL}}[q_{\theta_2} : p_{\theta_1}] = +\infty$ since $Q_{\theta_2} \not\ll P_{\theta_1}$.

Theorem 1 (Kullback–Leibler divergence between truncated exponential family densities)**.** *Let $\mathcal{E}_2 = \{q_{\theta_2}\}$ be an exponential family with support $\mathcal{X}_2$, and $\mathcal{E}_1 = \{p_{\theta_1}\}$ a truncated exponential family of $\mathcal{E}_2$ with support $\mathcal{X}_1 \subset \mathcal{X}_2$. Let F_1 and F_2 denote the log-normalizers of $\mathcal{E}_1$ and $\mathcal{E}_2$ and η_1 and η_2 the moment parameters corresponding to the natural parameters θ_1 and θ_2. Then the Kullback–Leibler divergence between a truncated density of $\mathcal{E}_1$ and a density of $\mathcal{E}_2$ is*

$$D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = Y_{F_2,F_1^*}(\theta_2 : \eta_1) = B_{F_2,F_1}(\theta_2 : \theta_1) = B_{F_1^*,F_2^*}(\eta_1 : \eta_2) = Y_{F_1^*,F_2}(\eta_1 : \theta_2). \quad (65)$$

For example, consider the calculation of the KLD between an exponential distribution (view as half a Laplacian distribution, i.e., a truncated Laplacian distribution on the positive real support) and a Laplacian distribution defined on the real line support.

Example 4. *Let $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$ denote the set of positive reals. Let $\mathcal{E}_1 = \{p_\lambda(x) = \lambda \exp(-\lambda x), \lambda \in \mathbb{R}_{++}, x > 0\}$ and $\mathcal{E}_2 = \{q_\lambda(x) = \lambda \exp(-\lambda|x|), \lambda \in \mathbb{R}_{++}, x \in \mathbb{R}\}$ denote the exponential families of exponential distributions and Laplacian distributions, respectively. We have the sufficient statistic $t(x) = -|x|$ and natural parameter $\theta = \lambda$ so that $\tilde{p}_\theta(x) = \exp(-|x|\theta)$. The log-normalizers are $F_1(\theta) = -\log\theta$ and $F_2(\theta) = -\log\theta + \log 2$ (hence $F_2(\theta) \geq F_1(\theta)$). The moment parameter $\eta = \nabla F_1(\theta) = \nabla F_2(\theta) = -\frac{1}{\theta} = -\frac{1}{\lambda}$. Thus using the duo Bregman divergence, we have:*

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= B_{F_2,F_1}(\theta_2 : \theta_1), && (66)\\ &= F_2(\theta_2) - F_1(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F_1(\theta_1), && (67)\\ &= \log 2 + \log \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} - 1. && (68) \end{aligned}$$

Moreover, we can interpret that divergence using the Itakura–Saito divergence [24]:

$$D_{\mathrm{IS}}[\lambda_1 : \lambda_2] := \frac{\lambda_1}{\lambda_2} - \log \frac{\lambda_1}{\lambda_2} - 1 \geq 0. \quad (69)$$

we have

$$D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = D_{\mathrm{IS}}[\lambda_2 : \lambda_1] + \log 2 \geq 0. \quad (70)$$

We check the result using the duo Fenchel–Young divergence:

$$D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = Y_{F_2,F_1^*}(\theta_2 : \eta_1), \tag{71}$$

with $F_1^*(\eta) = -1 + \log\left(-\frac{1}{\eta}\right)$*:*

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= Y_{F_2,F_1^*}(\theta_2 : \eta_1), & (72)\\ &= -\log\lambda_2 + \log 2 - 1 + \log\lambda_1 + \frac{\lambda_2}{\lambda_1}, & (73)\\ &= \log\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} + \log_2 - 1. & (74) \end{aligned}$$

Next, consider the calculation of the KLD between a half-normal distribution and a (full) normal distribution:

Example 5. *Consider* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *to be the scale family of the half standard normal distributions and the scale family of the standard normal distribution, respectively. We have* $\tilde{p}_\theta(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$ *with* $Z_1(\theta) = \sigma\sqrt{\frac{\pi}{2}}$ *and* $Z_2(\theta) = \sigma\sqrt{2\pi}$*. Let the sufficient statistic be* $t(x) = -\frac{x^2}{2}$ *so that the natural parameter is* $\theta = \frac{1}{\sigma^2} \in \mathbb{R}_{++}$*. Here, we have both* $\Theta_1 = \Theta_2 = \mathbb{R}_{++}$*. For this example, we check that* $Z_1(\theta) = \frac{1}{2}\, Z_2(\theta)$*. We have* $F_1(\theta) = -\frac{1}{2}\log\theta + \frac{1}{2}\log\frac{\pi}{2}$ *and* $F_2(\theta) = -\frac{1}{2}\log\theta + \frac{1}{2}\log(2\pi)$ *(with* $F_2(\theta) \geq F_1(\theta)$*). We have* $\eta = -\frac{1}{2\theta} = -\frac{1}{2}\sigma^2$*. The KLD between two half scale normal distributions is*

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] &= B_{F_1}(\theta_2 : \theta_1), & (75)\\ &= \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right). & (76) \end{aligned}$$

Since $F_1(\theta)$ *and* $F_2(\theta)$ *differ only by a constant and the Bregman divergence is invariant under an affine term of its generator, we have*

$$\begin{aligned} D_{\mathrm{KL}}[q_{\theta_1} : q_{\theta_2}] &= B_{F_2}(\theta_2 : \theta_1), & (77)\\ &= B_{F_1}(\theta_2 : \theta_1) = \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right). & (78) \end{aligned}$$

Moreover, we can interpret those Bregman divergences as half of the Itakura–Saito divergence:

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = D_{\mathrm{KL}}[q_{\theta_1} : q_{\theta_2}] = B_{F_2}(\theta_2 : \theta_1) = \frac{1}{2} D_{\mathrm{IS}}[\sigma_1^2 : \sigma_2^2]. \tag{79}$$

It follows that

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] &= B_{F_2,F_1}(\theta_2 : \theta_1) = F_2(\theta_2) - F_1(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F_1(\theta_1), & (80)\\ &= \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \log 4 - 1\right), & (81)\\ &= D_{\mathrm{KL}}[q_{\theta_1} : q_{\theta_2}] + \log 2. & (82) \end{aligned}$$

Since $\log 2 > 0$*, we have* $D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] \geq D_{\mathrm{KL}}[q_{\theta_1} : q_{\theta_2}]$.

Thus the Kullback–Leibler divergence between a truncated density and another density of the same exponential family amounts to calculate a duo Bregman divergence on the reverse parameter order: $D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}] = B_{F_2,F_1}(\theta_2 : \theta_1)$. Let $D_{\mathrm{KL}}^*[p : q] := D_{\mathrm{KL}}[q : p]$ be the reverse Kullback–Leibler divergence. Then $D_{\mathrm{KL}}^*[q_{\theta_2} : p_{\theta_1}] = B_{F_2,F_1}(\theta_2 : \theta_1)$.

Notice that truncated exponential families are also exponential families but those exponential families may be non-steep [25].

Let $\mathcal{E}_1 = \{p_\theta^{a_1,b_1}\}$ and $\mathcal{E}_2 = \{p_\theta^{a_2,b_2}\}$ be two truncated exponential families of the exponential family $\mathcal{E} = \{p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}\}$ with log-normalizer $F(\theta)$ such that

$$p_\theta^{a_i,b_i}(x) = \frac{p_\theta(x)}{Z_{a_i,b_i}(\theta)}, \tag{83}$$

with $Z_{a_i,b_i}(\theta) = \Phi_\theta(b_i) - \Phi_\theta(a_i)$, where $\Phi_\theta(x)$ denotes the CDF of $p_\theta(x)$. Then the log-normalizer of $\mathcal{E}_i$ is $F_i(\theta) = F(\theta) + \log(\Phi_\theta(b_i) - \Phi_\theta(a_i))$ for $i \in \{1,2\}$.

Corollary 1 (Kullback–Leibler divergence between densities of truncated exponential families)**.** *Let $\mathcal{E}_i = \{p_\theta^{a_i,b_i}\}$ be truncated exponential families of the exponential family $\mathcal{E} = \{p_\theta\}$ with support $\mathcal{X}_i = [a_i, b_i] \subset \mathcal{X}$ (where $\mathcal{X}$ denotes the support of $\mathcal{E}$) for $i \in \{1,2\}$. Then the Kullback–Leibler divergence between $p_{\theta_1}^{a_1,b_1}$ and $p_{\theta_2}^{a_2,b_2}$ is infinite if $[a_1,b_1] \not\subset [a_2,b_2]$ and has the following formula when $[a_1,b_1] \subset [a_2,b_2]$:*

$$D_{\mathrm{KL}}[p_{\theta_1}^{a_1,b_1} : p_{\theta_2}^{a_2,b_2}] = D_{\mathrm{KL}}[p_{\theta_1}^{a_1,b_1} : p_{\theta_2}^{a_1,b_1}] + \log \frac{Z_{a_2,b_2}(\theta_2)}{Z_{a_1,b_1}(\theta_2)}. \tag{84}$$

Proof. We have $p_\theta^{a_1,b_1} = \frac{p_\theta}{Z_{a_1,b_1}(\theta)}$ and $p_\theta^{a_2,b_2} = \frac{p_\theta}{Z_{a_2,b_2}(\theta)}$. Therefore $p_\theta^{a_2,b_2} = p_\theta^{a_1,b_1} \frac{Z_{a_1,b_1}(\theta)}{Z_{a_2,b_2}(\theta)}$. Thus we have

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1}^{a_1,b_1} : p_{\theta_2}^{a_2,b_2}] &= \int_{\mathcal{X}_1} p_{\theta_1}^{a_1,b_1}(x) \log \frac{p_{\theta_1}^{a_1,b_1}(x)}{p_{\theta_2}^{a_2,b_2}} \mathrm{d}\mu(x), & (85)\\ &= \int_{\mathcal{X}_1} p_{\theta_1}^{a_1,b_1}(x) \log \frac{p_{\theta_1}^{a_1,b_1}(x)}{p_{\theta_2}^{a_1,b_1}} \mathrm{d}\mu(x) + \log \frac{Z_{a_2,b_2}(\theta_2)}{Z_{a_1,b_1}(\theta_2)}, & (86)\\ &= D_{\mathrm{KL}}[p_{\theta_1}^{a_1,b_1} : p_{\theta_2}^{a_1,b_1}] + \log \frac{Z_{a_2,b_2}(\theta_2)}{Z_{a_1,b_1}(\theta_2)}. & (87) \end{aligned}$$

□

Thus the KLD between truncated exponential family densities $p_{\theta_1}^{a_1,b_1}$ and $p_{\theta_2}^{a_2,b_2}$ amounts to the KLD between the densities with the same truncation parameter with an additive term depending on the log ratio of the mass with respect to the truncated supports evaluated at θ_2. We shall illustrate with two examples the calculation of the KLD between truncated exponential families.

Example 6. *Consider the calculation of the KLD between a truncated exponential distribution $p_{\lambda_1}^{a_1,b_1}$ with support $\mathcal{X}_1 = [a_1,b_1]$ ($b_1 > a_1 \geq 0$) and another truncated exponential distribution $p_{\lambda_2}^{a_2,b_2}$ with support $\mathcal{X}_2 = [a_2,b_2]$ ($b_2 > a_2 \geq 0$). We have $p_\lambda(x) = \lambda \exp(-\lambda x)$ (density of the untruncated exponential family with natural parameter $\theta = \lambda$, sufficient statistic $t(x) = -x$ and log-normalizer $F(\theta) = -\log\theta$), $p_{\lambda_1}^{a_1,b_1} = \frac{1}{Z_{a_1,b_1}(\lambda)} p_{\lambda_1}(x)$, and $p_{\lambda_2}^{a_2,b_2} = \frac{1}{Z_{a_2,b_2}(\lambda)} p_{\lambda_2}(x)$. Let $\Phi_\lambda(x) = 1 - \exp(-\lambda x)$ denote the cumulative distribution function of the exponential distribution. We have $Z_{a,b}(\lambda) = \Phi_b(\lambda) - \Phi_a(\lambda)$ and*

$$F_{a,b}(\lambda) = F(\lambda) + \log(\Phi_b(\lambda) - \Phi_a(\lambda)) = -\log\lambda + \log(e^{-\lambda a} - e^{-\lambda b}). \tag{88}$$

If $[a_1, b_1] \notin [a_2, b_2]$ *then* $D_{\mathrm{KL}}[p_{\lambda_1} : q_{\lambda_2}] = +\infty$. *Otherwise,* $[a_1, b_1] \in [a_2, b_2]$, *and the exponential family* $\{p_\lambda\}$ *is the truncated exponential family* $\{q_\lambda\}$. *Using the computer algebra system Maxima (https://maxima.sourceforge.io/ accessed on 15 March 2022), we find that*

$$-E_{p_\lambda}[x] = \frac{(1+\lambda b)e^{\lambda a} - (1+\lambda a)e^{\lambda b}}{\lambda(e^{\lambda b} - e^{\lambda a})} = F'_{a,b}(\lambda). \tag{89}$$

Thus we have:

$$\begin{aligned} D_{\mathrm{KL}}[p_{\lambda_1}^{a_1,b_1} : q_{\lambda_2}^{a_2,b_2}] &= B_{F_2,F_1}(\theta_2 : \theta_1), & (90)\\ &= F_{a_2,b_2}(\lambda_2) - F_{a_1,b_1}(\lambda_1) - (\lambda_2 - \lambda_1)F'_{a_1,b_1}(\lambda_1), \\ &= \log\frac{\lambda_1}{\lambda_2} + (\lambda_2 - \lambda_1)\, E_{p_{\lambda_1}}[x] + \log\frac{e^{-\lambda_2 a_2} - e^{-\lambda_2 b_2}}{e^{-\lambda_1 a_1} - e^{-\lambda_1 b_1}}. & (91) \end{aligned}$$

When $a_1 = a_2 = 0$ *and* $b_1 = b_2 = +\infty$, *we recover the KLD between two exponential distributions* p_{λ_1} *and* p_{λ_2}:

$$\begin{aligned} D_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] &= B_F(\lambda_2 : \lambda_1), & (92)\\ &= F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)F'(\theta_1), & (93)\\ &= \frac{\lambda_2}{\lambda_1} - \log\frac{\lambda_2}{\lambda_1} - 1 = D_{\mathrm{IS}}[\lambda_2 : \lambda_1]. & (94) \end{aligned}$$

Note that the KLD between two truncated exponential distributions with the same truncation support $\mathcal{X} = [a, b]$ *is*

$$D_{\mathrm{KL}}[p_{\lambda_1}^{a,b} : p_{\lambda_2}^{a,b}] = \log\frac{\lambda_2}{\lambda_1} + \log\frac{\Phi_{\lambda_2}(b) - \Phi_{\lambda_2}(a)}{\Phi_{\lambda_1}(b) - \Phi_{\lambda_1}(a)} + (\lambda_2 - \lambda_1)E_{p_1^{a,b}}[x]. \tag{95}$$

We also check Corollary 1:

$$D_{\mathrm{KL}}[p_{\lambda_1}^{a_1,b_1} : p_{\lambda_2}^{a_2,b_2}] = D_{\mathrm{KL}}[p_{\lambda_1}^{a_1,b_1} : p_{\lambda_2}^{a_1,b_1}] + \log\frac{Z_{a_2,b_2}(\lambda_2)}{Z_{a_1,b_1}(\lambda_2)}, \tag{96}$$

The next example shows how to compute the Kullback–Leibler divergence between two truncated normal distributions:

Example 7. *Let* $N_{a,b}(m, s)$ *denote a truncated normal distribution with support the open interval* (a, b) $(a < b)$ *and probability density function defined by:*

$$p_{m,s}^{a,b}(x) = \frac{1}{Z_{a,b}(m,s)} \exp\left(-\frac{(x-m)^2}{2s^2}\right), \tag{97}$$

where $Z_{a,b}(m, s)$ *is related to the partition function [26] expressed using the cumulative distribution function (CDF)* $\Phi_{m,s}(x)$:

$$Z_{a,b}(m,s) = \sqrt{2\pi}s\,(\Phi_{m,s}(b) - \Phi_{m,s}(a)), \tag{98}$$

with

$$\Phi_{m,s}(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-m}{\sqrt{2}s}\right)\right), \tag{99}$$

where $\mathrm{erf}(x)$ *is the error function:*

$$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t. \tag{100}$$

Thus we have $\mathrm{erf}(x) = 2\,\Phi(\sqrt{2}x) - 1$ *where* $\Phi(x) = \Phi_{0,1}(x)$.

The pdf can also be written as

$$p_{m,s}^{a,b}(x) = \frac{1}{s}\frac{\phi(\frac{x-m}{s})}{\Phi(\frac{b-m}{s}) - \Phi(\frac{a-m}{s})}, \tag{101}$$

where $\phi(x)$ *denotes the standard normal pdf* ($\phi(x) = p_{0,1}^{-\infty,+\infty}(x)$)*:*

$$\phi(x) := \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), \tag{102}$$

and $\Phi(x) = \Phi_{0,1}(x) = \int_{-\infty}^{x}\phi(t)\,\mathrm{d}t$ *is the standard normal CDF. When* $a = -\infty$ *and* $b = +\infty$*, we have* $Z_{-\infty,\infty}(m,s) = \sqrt{2\pi}\,s$ *since* $\Phi(-\infty) = 0$ *and* $\Phi(+\infty) = 1$.

The density $p_{m,s}^{a,b}(x)$ *belongs to an exponential family* $\mathcal{E}_{a,b}$ *with natural parameter* $\theta = \left(\frac{m}{s^2}, -\frac{1}{2s^2}\right)$*, sufficient statistics* $t(x) = (x, x^2)$*, and log-normalizer:*

$$F_{a,b}(\theta) = -\frac{\theta_1^2}{4\theta_2} + \log Z_{a,b}(\theta) \tag{103}$$

The natural parameter space is $\Theta = \mathbb{R}\times\mathbb{R}_{--}$ *where* $\mathbb{R}_{--} = \{x \in \mathbb{R} : x < 0\}$ *denotes the set of negative real numbers.*

The log-normalizer can be expressed using the source parameters (m,s) *(which are not the mean and standard deviation when the support is truncated, hence the notation m and s):*

$$\begin{aligned} F_{a,b}(m,s) &= \frac{m^2}{2s^2} + \log Z_{a,b}(m,s), &(104)\\ &= \frac{m^2}{2s^2} + \frac{1}{2}\log 2\pi s^2 + \log(\Phi_{m,s}(b) - \Phi_{m,s}(a)). &(105)\end{aligned}$$

We shall use the fact that the gradient of the log-normalizer of any exponential family distribution amounts to the expectation of the sufficient statistics [1]:

$$\nabla F_{a,b}(\theta) = E_{p_{m,s}^{a,b}}[t(x)] = \eta. \tag{106}$$

Parameter η *is called the moment or expectation parameter [1].*

The mean $\mu(m,s;a,b) = E_{p_{m,s}^{a,b}}[x] = \frac{\partial}{\partial\theta_1}F_{a,b}(\theta)$ *and the variance* $\sigma^2(m,s;a,b) = E_{p_{m,s}^{a,b}}[x^2] - \mu^2$ *(with* $E_{p_{m,s}^{a,b}}[x^2] = \frac{\partial}{\partial\theta_2}F_{a,b}(\theta)$*) of the truncated normal* $p_{m,s}^{a,b}$ *can be expressed using the following formula [26,27] (page 25):*

$$\begin{aligned} \mu(m,s;a,b) &= m - s\,\frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}, &(107)\\ \sigma^2(m,s;a,b) &= s^2\left(1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\beta) - \phi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right), &(108)\end{aligned}$$

where $\alpha := \frac{a-m}{s}$ *and* $\beta := \frac{b-m}{s}$*. Thus we have the following moment parameter* $\eta = (\eta_1, \eta_2)$ *with*

$$\begin{aligned} \eta_1(m,s;a,b) &= E_{p_{m,s}^{a,b}}[x] = \mu(m,s;a,b), &(109)\\ \eta_2(m,s;a,b) &= E_{p_{m,s}^{a,b}}[x^2] = \sigma^2(m,s;a,b) + \mu^2(m,s;a,b). &(110)\end{aligned}$$

Now consider two truncated normal distributions $p_{m_1,s_1}^{a_1,b_1}$ *and* $p_{m_2,s_2}^{a_2,b_2}$ *with* $[a_1,b_1] \subseteq [a_2,b_2]$ *(otherwise, we have* $D_{\mathrm{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] = +\infty$*). Then the KLD between* $p_{m_1,s_1}^{a_1,b_1}$ *and* $p_{m_2,s_2}^{a_2,b_2}$ *is equivalent to a duo Bregman divergence:*

$$\begin{aligned} D_{\mathrm{KL}}[p_{m_1,s_1}^{a_1,b_1} : p_{m_2,s_2}^{a_2,b_2}] &= F_{m_2,s_2}(\theta_2) - F_{m_1,s_1}(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F_{m_1,s_1}(\theta_1), \\ &= \frac{m_2}{2s_2^2} - \frac{m_1}{2s_1^2} + \log \frac{Z_{a_2,b_2}(m_2,s_2)}{Z_{a_1,b_1}(m_1,s_1)} - \left(\frac{m_2}{s_2^2} - \frac{m_1}{s_1^2}\right)\eta_1(m_1,s_1;a_1,b_1) \\ &\quad - \left(\frac{1}{2s_1^2} - \frac{1}{2s_2^2}\right)\eta_2(m_1,s_1;a_1,b_1). \end{aligned} \tag{111}$$

Note that $F_{m_2,s_2}(\theta) \geq F_{m_1,s_1}(\theta)$.

This formula is valid for (1) the KLD between two truncated normal distributions, or for (2) the KLD between a truncated normal distribution and a (full support) normal distribution. Note that the formula depends on the erf function used in function Φ. *Furthermore, when* $a_1 = a_2 = -\infty$ *and* $b_1 = b_2 = +\infty$, *we recover (3) the KLD between two univariate normal distributions, since* $\log \frac{Z_{a_2,b_2}(m_2,s_2)}{Z_{a_1,b_1}(m_1,s_1)} = \log \frac{\sigma_2}{\sigma_1} = \frac{1}{2}\log \frac{\sigma_2^2}{\sigma_1^2}$:

$$D_{\mathrm{KL}}[p_{m_1,s_1} : p_{m_2,s_2}] = \frac{1}{2}\left(\log \frac{s_2^2}{s_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(m_2 - m_1)^2}{s_2^2} - 1.\right). \tag{112}$$

Note that for full support normal distributions, we have $\mu(m,s;-\infty;+\infty) = m$ *and* $\sigma^2(m,s;-\infty;+\infty) = s^2$.

The entropy of a truncated normal distribution (an exponential family [28]) is $h[p_{m,s}^{a,b}] = -\int_a^b p_{m,s}^{a,b}(x)\log p_{m,s}^{a,b}\mathrm{d}x = -F^*(\eta) = F(\theta) - \theta^\top\eta$. *We find that*

$$h[p_{m,s}^{a,b}] = \log\left(\sqrt{2\pi e}s\left(\Phi(\beta) - \Phi(\alpha)\right)\right) + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2(\Phi(\beta) - \Phi(\alpha))}. \tag{113}$$

When $(a,b) = (-\infty,\infty)$, *we have* $\Phi(\beta) - \Phi(\alpha) = 1$ *and* $\alpha\phi(\alpha) - \beta\phi(\beta) = 0$ *since* $\beta = -\alpha$, $\phi(-x) = \phi(x)$ *(an even function), and* $\lim_{\beta\to+\infty}\beta\phi(\beta) = 0$. *Thus we recover the differential entropy of a normal distribution:* $h[p_{\mu,\sigma}] = \log\left(\sqrt{2\pi e}\sigma\right)$.

5. Bhattacharyya Skewed Divergence Between Truncated Densities of an Exponential Family

The Bhattacharyya α-skewed divergence [29,30] between two densities $p(x)$ and $q(x)$ with respect to μ is defined for a skewing scalar parameter $\alpha \in (0,1)$ as:

$$D_{\mathrm{Bhat},\alpha}[p:q] := -\log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\,\mathrm{d}\mu(x), \tag{114}$$

where $\mathcal{X}$ denotes the support of the distributions. The Bhattacharyya distance is

$$D_{\mathrm{Bhat}}[p,q] = D_{\mathrm{Bhat},\frac{1}{2}}[p:q] = -\log \int_{\mathcal{X}} \sqrt{p(x)q(x)}\,\mathrm{d}\mu(x). \tag{115}$$

The Bhattacharyya distance is not a metric distance since it does not satisfy the triangle inequality. The Bhattacharyya distance is related to the Hellinger distance [31] as follows:

$$D_H[p,q] = \sqrt{\frac{1}{2}\int_{\mathcal{X}}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}\mu(x)} = \sqrt{1 - \exp(-D_{\mathrm{Bhat}}[p,q])}. \tag{116}$$

The Hellinger distance is a metric distance.

Let $I_\alpha[p:q] := \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\,\mathrm{d}\mu(x)$ denote the skewed affinity coefficient so that $D_{\mathrm{Bhat},\alpha}[p:q] = -\log I_\alpha[p:q]$. Since $I_\alpha[p:q] = I_{1-\alpha}[q:p]$, we have $D_{\mathrm{Bhat},\alpha}[p:q] = D_{\mathrm{Bhat},1-\alpha}[q:p]$.

Consider an exponential family $\mathcal{E} = \{p_\theta\}$ with log-normalizer $F(\theta)$. Then it is well-known that the α-skewed Bhattacharyya divergence between two densities of an exponential family amounts to a skewed Jensen divergence [30] (originally called Jensen difference in [32]):

$$D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : p_{\theta_2}] = J_{F,\alpha}(\theta_1 : \theta_2), \tag{117}$$

where the skewed Jensen divergence is defined by

$$J_{F,\alpha}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2). \tag{118}$$

The convexity of the log-normalizer $F(\theta)$ ensures that $J_{F,\alpha}(\theta_1 : \theta_2) \geq 0$. The Jensen divergence can be extended to full real α by rescaling it by $\frac{1}{\alpha(1-\alpha)}$, see [33].

Remark 1. *The Bhattacharyya skewed divergence $D_{\mathrm{Bhat},\alpha}[p : q]$ appears naturally as the negative of the log-normalizer of the exponential family induced by the exponential arc $\{r_\alpha(x)\ \alpha \in (0,1)\}$ linking two densities p and q with $r_\alpha(x) \propto p(x)^\alpha q(x)^{1-\alpha}$. This arc is an exponential family of order 1:*

$$\begin{aligned} r_\alpha(x) &= \exp(\alpha \log p(x) + (1-\alpha)\log q(x) - \log Z_\alpha(p:q)), && (119)\\ &= \exp\left(\alpha \log \frac{p(x)}{q(x)} - F_{pq}(\alpha)\right) q(x). && (120) \end{aligned}$$

The sufficient statistic is $t(x) = \frac{p(x)}{q(x)}$, the natural parameter $\alpha \in (0,1)$, and the log-normalizer $F_{pq}(\alpha) = \log Z_\alpha(p:q) = \log \int p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}\mu(x) = -D_{\mathrm{Bhat},\alpha}[p:q]$. This shows that $D_{\mathrm{Bhat},\alpha}[p:q]$ is concave with respect to α since log-normalizers $F_{pq}(\alpha)$ are always convex. Grünwald called those exponential families the likelihood ratio exponential families [34].

Now, consider calculating $D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : q_{\theta_2}]$ where $p_{\theta_1} \in \mathcal{E}_1$ with $\mathcal{E}_1$ a truncated exponential family of $\mathcal{E}_2$ and $q_{\theta_2} \in \mathcal{E}_2 = \{q_\theta\}$. We have $q_\theta(x) = \frac{Z_1(\theta)}{Z_2(\theta)} p_\theta(x)$, where $Z_1(\theta)$ and $Z_2(\theta)$ are the partition functions of $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively. Thus we have

$$I_\alpha[p_{\theta_1} : q_{\theta_2}] = \left(\frac{Z_1(\theta_2)}{Z_2(\theta_2)}\right)^{1-\alpha} I_\alpha[p_{\theta_1} : p_{\theta_2}], \tag{121}$$

and the α-skewed Bhattacharyya divergence is

$$D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : q_{\theta_2}] = D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : p_{\theta_2}] - (1-\alpha)(F_1(\theta_2) - F_2(\theta_2)). \tag{122}$$

Therefore we obtain

$$\begin{aligned} D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : q_{\theta_2}] &= J_{F_1,\alpha}(\theta_1 : \theta_2) - (1-\alpha)(F_1(\theta_2) - F_2(\theta_2)), && (123)\\ &= \alpha F_1(\theta_1) + (1-\alpha)F_2(\theta_2) - F_1(\alpha\theta_1 + (1-\alpha)\theta_2), && (124)\\ &=: J_{F_1,F_2,\alpha}(\theta_1 : \theta_2). && (125) \end{aligned}$$

We call $J_{F_1,F_2,\alpha}(\theta_1 : \theta_2)$ the duo Jensen divergence. Since $F_2(\theta) \geq F_1(\theta)$, we check that

$$J_{F_1,F_2,\alpha}(\theta_1 : \theta_2) \geq J_{F_1,\alpha}(\theta_1 : \theta_2) \geq 0. \tag{126}$$

Figure 7 illustrates graphically the duo Jensen divergence $J_{F_1,F_2,\alpha}(\theta_1 : \theta_2)$.

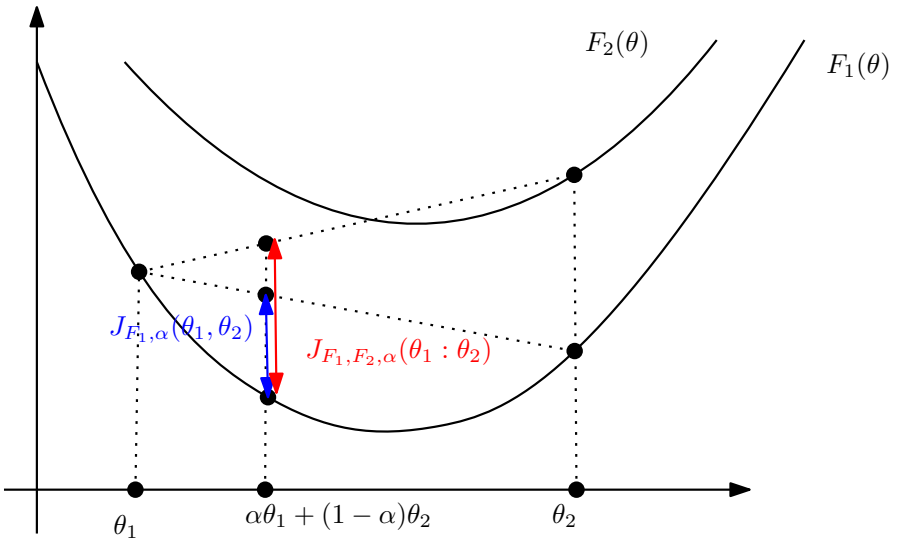


Figure 7. The duo Jensen divergence $J_{F_1,F_2,\alpha}(\theta_1:\theta_2)$ is greater than the Jensen divergence $J_{F_1,\alpha}(\theta_1:\theta_2)$ for $F_2(\theta)\geq F_1(\theta)$.

Theorem 2. *The α-skewed Bhattacharyya divergence for $\alpha\in(0,1)$ between a truncated density of $\mathcal{E}_1$ with log-normalizer $F_1(\theta)$ and another density of an exponential family $\mathcal{E}_2$ with log-normalizer $F_2(\theta)$ amounts to a duo Jensen divergence:*

$$D_{\mathrm{Bhat},\alpha}[p_{\theta_1}:q_{\theta_2}]=J_{F_1,F_2,\alpha}(\theta_1:\theta_2), \tag{127}$$

where $J_{F_1,F_2,\alpha}(\theta_1:\theta_2)$ is the duo skewed Jensen divergence induced by two strictly convex functions $F_1(\theta)$ and $F_2(\theta)$ such that $F_2(\theta)\geq F_1(\theta)$:

$$J_{F_1,F_2,\alpha}(\theta_1:\theta_2)=\alpha F_1(\theta_1)+(1-\alpha)F_2(\theta_2)-F_1(\alpha\theta_1+(1-\alpha)\theta_2). \tag{128}$$

In [30], it is reported that

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1}:p_{\theta_2}] &= B_F(\theta_2:\theta_1), && (129)\\ &= \lim_{\alpha\to 0}\frac{1}{\alpha}J_{F,\alpha}(\theta_2:\theta_1)=\lim_{\alpha\to 0}\frac{1}{\alpha}J_{F,1-\alpha}(\theta_1:\theta_2), && (130)\\ &= \lim_{\alpha\to 0}\frac{1}{\alpha}D_{\mathrm{Bhat},\alpha}[p_{\theta_2}:p_{\theta_1}]=\lim_{\alpha\to 0}\frac{1}{\alpha}D_{\mathrm{Bhat},1-\alpha}[p_{\theta_1}:p_{\theta_2}]. && (131)\end{aligned}$$

Indeed, using the first-order Taylor expansion of

$$F(\theta_1+\alpha(\theta_2-\theta_1))\underset{\alpha\to 0}{\approx}F(\theta_1)+\alpha(\theta_2-\theta_1)^\top\nabla F(\theta_1) \tag{132}$$

when $\alpha\to 0$, we check that we have

$$\begin{aligned} \frac{1}{\alpha}J_{F,\alpha}(\theta_2:\theta_1) &:= \frac{F(\theta_1)+\alpha(F(\theta_2)-F(\theta_1))-F(\theta_1+\alpha(\theta_2-\theta_1))}{\alpha}, && (133)\\ &\underset{\alpha\to 0}{\overset{\text{Equation (132)}}{\approx}} \frac{\cancel{F(\theta_1)}+\alpha(F(\theta_2)-F(\theta_1))-\cancel{F(\theta_1)}-\alpha(\theta_2-\theta_1)^\top\nabla F(\theta_1)}{\alpha}, && (134)\\ &= F(\theta_2)-F(\theta_1)-(\theta_2-\theta_1)^\top\nabla F(\theta_1), && (135)\\ &=: B_F(\theta_2:\theta_1). && (136)\end{aligned}$$

Thus we have $\lim_{\alpha\to 0}\frac{1}{\alpha}J_{F,\alpha}(\theta_2:\theta_1)=B_F(\theta_2:\theta_1)$.

Moreover, we have

$$\lim_{\alpha\to 0}\frac{1}{\alpha}D_{\mathrm{Bhat},1-\alpha}[p:q]=D_{\mathrm{KL}}[p:q]. \tag{137}$$

Similarly, we can prove that

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} J_{F_1,F_2,\alpha}(\theta_1 : \theta_2) = B_{F_2,F_1}(\theta_2 : \theta_1), \tag{138}$$

which can be reinterpreted as

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} D_{\mathrm{Bhat},\alpha}[p_{\theta_1} : q_{\theta_2}] = D_{\mathrm{KL}}[p_{\theta_1} : q_{\theta_2}]. \tag{139}$$

6. Concluding Remarks

We considered the Kullback–Leibler divergence between two parametric densities $p_\theta \in \mathcal{E}_1$ and $q_{\theta'} \in \mathcal{E}_2$ belonging to truncated exponential families [7] $\mathcal{E}_1$ and $\mathcal{E}_2$, and we showed that their KLD is equivalent to a duo Bregman divergence on swapped parameter order (Theorem 1). This result generalizes the study of Azoury and Warmuth [13]. The duo Bregman divergence can be rewritten as a duo Fenchel–Young divergence using mixed natural/moment parameterizations of the exponential family densities (Definition 1). This second result generalizes the approach taken in information geometry [15,35]. We showed how to calculate the Kullback–Leibler divergence between two truncated normal distributions as a duo Bregman divergence. More generally, we proved that the skewed Bhattacharyya distance between two parametric densities of truncated exponential families amounts to a duo Jensen divergence (Theorem 2). We showed asymptotically that scaled duo Jensen divergences tend to duo Bregman divergences generalizing a result of [30,33]. This study of duo divergences induced by pair of generators was motivated by the formula obtained for the Kullback–Leibler divergence between two densities of two different exponential families originally reported in [23] (Equation (29)).

It is interesting to find applications of the duo Fenchel–Young, Bregman, and Jensen divergences beyond the scope of calculating statistical distances between truncated exponential family densities. Note that in [36], the authors exhibit a relationship between densities with nested supports and quasi-convex Bregman divergences. However, those considered parametric densities are not exponential families since their supports depend on the parameter. Recently, Khan and Swaroop [37] used this duo Fenchel–Young divergence in machine learning for knowledge-adaptation priors in the so-called change regularizer task.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author would like to thank the three reviewers for their helpful comments, which led to this improved paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Sundberg, R. *Statistical Modelling by Exponential Families*; Cambridge University Press: Cambridge, UK, 2019; Volume 12.
2. Pitman, E.J.G. *Sufficient Statistics and Intrinsic Accuracy*; Mathematical Proceedings of the cambridge Philosophical Society; Cambridge University Press: Cambridge, UK, 1936; Volume 32, pp. 567–579.
3. Darmois, G. Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris* **1935**, *260*, 85.
4. Koopman, B.O. On distributions admitting a sufficient statistic. *Trans. Am. Math. Soc.* **1936**, *39*, 399–409. [CrossRef]
5. Hiejima, Y. Interpretation of the quasi-likelihood via the tilted exponential family. *J. Jpn. Stat. Soc.* **1997**, *27*, 157–164. [CrossRef]
6. Efron, B.; Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*; Cambridge University Press: Cambridge, UK, 2021; Volume 6.
7. Akahira, M. *Statistical Estimation for Truncated Exponential Families*; Springer: Berlin/Heidelberg, Germany, 2017.

8. Bar-Lev, S.K. Large sample properties of the MLE and MCLE for the natural parameter of a truncated exponential family. *Ann. Inst. Stat. Math.* **1984**, *36*, 217–222. [CrossRef]
9. Shah, A.; Shah, D.; Wornell, G. A Computationally Efficient Method for Learning Exponential Family Distributions. *Adv. Neural Inf. Process. Syst.* **2021**, *34*. Available online: https://proceedings.neurips.cc/paper/2021/hash/84f7e69969dea92a925508f7c1f9579a-Abstract.html (accessed on 15 March 2022).
10. Keener, R.W. *Theoretical Statistics: Topics for a Core Course*; Springer: Berlin/Heidelberg, Germany, 2010.
11. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
12. Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf Beweis der Ergodizitaet von Markoffschen Ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* **1964**, *8*, 85–108.
13. 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]
14. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 2015.
15. Amari, S.I. Differential-geometrical methods in statistics. *Lect. Notes Stat.* **1985**, *28*, 1.
16. 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]
17. Acharyya, S. Learning to Rank in Supervised and Unsupervised Settings Using Convexity and Monotonicity. Ph.D. Thesis, The University of Texas at Austin, Austin, TX, USA, 2013.
18. Blondel, M.; Martins, A.F.; Niculae, V. Learning with Fenchel-Young losses. *J. Mach. Learn. Res.* **2020**, *21*, 1–69.
19. Nielsen, F. An elementary introduction to information geometry. *Entropy* **2020**, *22*, 1100. [CrossRef] [PubMed]
20. Mitroi, F.C.; Niculescu, C.P. *An Extension of Young's Inequality*; Abstract and Applied Analysis; Hindawi: London, UK, 2011; Volume 2011.
21. Jeffreys, H. *The Theory of Probability*; OUP Oxford: Oxford, UK, 1998.
22. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
23. Nielsen, F. On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius. *Entropy* **2021**, *23*, 464. [CrossRef]
24. Itakura, F.; Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968; pp. 280–292.
25. Del Castillo, J. The singly truncated normal distribution: A non-steep exponential family. *Ann. Inst. Stat. Math.* **1994**, *46*, 57–66. [CrossRef]
26. Burkardt, J. *The Truncated Normal Distribution*; Technical Report; Department of Scientific Computing Website, Florida State University: Tallahassee, FL, USA, 2014.
27. Kotz, J.; Balakrishan. *Continuous Univariate Distributions, Volumes I and II*; John Wiley and Sons: Hoboken, NJ, USA, 1994.
28. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3621–3624.
29. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.
30. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
31. Hellinger, E. Neue Begründung der Theorie Quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.* **1909**, *1909*, 210–271. [CrossRef]
32. Rao, C.R. Diversity and dissimilarity coefficients: A unified approach. *Theor. Popul. Biol.* **1982**, *21*, 24–43. [CrossRef]
33. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195. [CrossRef] [PubMed]
34. Grünwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
35. Nielsen, F. The Many Faces of Information Geometry. *Not. Am. Math. Soc.* **2022**, *69*. [CrossRef]
36. Nielsen, F.; Hadjeres, G. *Quasiconvex Jensen Divergences and Quasiconvex Bregman Divergences*; Workshop on Joint Structures and Common Foundations of Statistical Physics, Information Geometry and Inference for Learning; Springer: Berlin/Heidelberg, Germany, 2020; pp. 196–218.
37. Emtiyaz Khan, M.; Swaroop, S. Knowledge-Adaptation Priors. *arXiv* **2021**, arXiv:2106.08769.