Article

Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities

Andrzej Cichocki [1,2,*] **and Shun-ichi Amari** [3]

[1] Riken Brain Science Institute, Laboratory for Advanced Brain Signal Processing, Wako-shi, Japan

[2] Systems Research Institute, Polish Academy of Science, Poland

[3] Riken Brain Science Institute, Laboratory for Mathematical Neuroscience, Wako-shi, Japan

[*] Author to whom correspondence should be addressed; E-Mail: a.cichocki@riken.jp;
Tel.: +81-48467-9668; Fax: +81-48467-9686.

Abstract: In this paper, we extend and overview wide families of Alpha-, Beta- and Gamma-divergences and discuss their fundamental properties. In literature usually only one single asymmetric (Alpha, Beta or Gamma) divergence is considered. We show in this paper that there exist families of such divergences with the same consistent properties. Moreover, we establish links and correspondences among these divergences by applying suitable nonlinear transformations. For example, we can generate the Beta-divergences directly from Alpha-divergences and vice versa. Furthermore, we show that a new wide class of Gamma-divergences can be generated not only from the family of Beta-divergences but also from a family of Alpha-divergences. The paper bridges these divergences and shows also their links to Tsallis and Rényi entropies. Most of these divergences have a natural information theoretic interpretation.

1. Introduction

Many machine learning algorithms for classification and clustering employ a variety of dissimilarity measures. Information theory, convex analysis, and information geometry play key roles in the formulation of such divergences [1–25].

The most popular and often used are: Squared Euclidean distance and Kullback–Leibler divergence. Recently, alternative generalized divergences such as the Csiszár–Morimoto f-divergence and Bregman divergence become attractive alternatives for advanced machine learning algorithms [26–34]. In this paper, we discuss a robust parameterized subclass of the Csiszár–Morimoto and the Bregman divergences: Alpha- and Beta-divergences that may provide more robust solutions with respect to outliers and additive noise and improved accuracy. Moreover, we provide links to new-class of robust Gamma-divergences [35] and extend this class to so called Alpha-Gamma divergences.

Divergences are considered here as (dis)similarity measures. Generally speaking, they measure a quasi-distance or directed difference between two probability distributions $\mathbf{P}$ and $\mathbf{Q}$, which can also be expressed for unconstrained nonnegative multi-way arrays and patterns.

In this paper we assume that $\mathbf{P}$ and $\mathbf{Q}$ are positive measures (densities) not necessary normalized, but should be finite measures. In the special case of normalized densities, we explicitly refer to these as probability densities. If we do not mention explicitly we assume that these measures are continuous. An information divergence is a measure of distance between two probability curves. In this paper, we discuss only one-dimensional probability curves (represented by nonnegative signals or time series). Generalization to two or multidimensional dimensional variables is straightforward. One density $\mathbf{Q}(x)$ is usually known and fixed and another one $\mathbf{P}(\boldsymbol{x})$ is learned or adjusted to achieve a best in some sense similarity to the $\mathbf{Q}(\boldsymbol{x})$. For example, a discrete density $\mathbf{Q}$ corresponds to the observed data and the vector $\mathbf{P}$ to be estimated, or expected data which are subject to constraints imposed on the assumed models. For the Non-negative Matrix Factorization (NMF) problem $\mathbf{Q}$ corresponds to the data matrix $\mathbf{Y}$ and $\mathbf{P}$ corresponds to estimated matrix $\hat{\mathbf{Y}} = \mathbf{A}\mathbf{X}$ (or vice versa) [30].

The distance between two densities is called a metric if the following conditions hold:

1. $D(\mathbf{P} \,||\, \mathbf{Q}) \geq 0$ with equality if and only if $\mathbf{P} = \mathbf{Q}$ (nonnegativity and positive definiteness),

2. $D(\mathbf{P} \,||\, \mathbf{Q}) = D(\mathbf{Q} \,||\, \mathbf{P})$ (symmetry),

3. $D(\mathbf{P} \,||\, \mathbf{Z}) \leq D(\mathbf{P} \,||\, \mathbf{Q}) + D(\mathbf{Q} \,||\, \mathbf{Z})$ (subaddivity/triangle inequality).

Distances which only satisfies Condition 1 are not a metric and are referred to as (asymmetric) divergences.

In many applications, such as image analysis, pattern recognition and statistical machine learning we use the information-theoretic divergences rather than Euclidean squared or l_p-norm distances [28]. Several information divergences such as Kullback–Leibler, Hellinger and Jensen–Shannon divergences are central to estimate similarity between distributions and have long history in information geometry.

The concept of a divergence is not restricted to Euclidean spaces but can be extended to abstract spaces with the help of Radon–Nikodym derivative (see for example [36]). Let $(\mathbf{X}, \mathcal{A}, \mu)$ be a measure space, where μ is a finite or a σ-finite measure on $(\mathbf{X}, \mathcal{A})$ and let assume that $\mathbf{P}$ and $\mathbf{Q}$ are two (probability)

measures on $(\mathbf{X}, \mathcal{A})$ such that $\mathbf{P} << \mu$, $\mathbf{Q} << \mu$ are absolutely continuous with respect to a measure μ, e.g., $\mu = \mathbf{P} + \mathbf{Q}$ and that $p = \frac{d\mathbf{P}}{d\mu}$ and $q = \frac{d\mathbf{Q}}{d\mu}$ the (densities) Radon–Nikodym derivative of $\mathbf{P}$ and $\mathbf{Q}$ with respect to μ. Using such notations the fundamental Kullback–Leibler (KL) divergence between two probabilities distributions can be written as

$$D_{KL}(\mathbf{P} \,||\, \mathbf{Q}) = \int_{\mathbf{X}} p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) \tag{1}$$

which is related to the Shannon entropy

$$H_S(\mathbf{P}) = -\int_{\mathbf{X}} p(\boldsymbol{x}) \log(p(\boldsymbol{x})) d\mu(\boldsymbol{x}) \tag{2}$$

via

$$D_{KL}(\mathbf{P} \,||\, \mathbf{Q}) = V_S(\mathbf{P}, \mathbf{Q}) - H_S(\mathbf{P})$$

where

$$V_S(\mathbf{P}, \mathbf{Q}) = -\int_{\mathbf{X}} p(\boldsymbol{x}) \log(q(\boldsymbol{x})) d\mu(\boldsymbol{x})$$

is the Shannon's cross entropy, provided that integrals exist. (In measure theoretic terms, the integral exists if the measure induced by $\mathbf{P}$ is absolutely continuous with respect to that induced by $\mathbf{Q}$). Here and in the whole paper we assume that all integrals exist.

The Kullback–Leibler divergence has been generalized by using a family of functions called generalized logarithm functions or α-logarithm

$$\log_\alpha(x) = \frac{1}{1-\alpha}(x^{1-\alpha} - 1) \qquad (\text{for } x > 0) \tag{3}$$

which is a power function of x with power $1 - \alpha$, and is the natural logarithm function in the limit $\alpha \to 1$. Often, the power function (3) allows to generate more robust divergences in respect to outliers and consequently better or more flexible performance (see, for example, [37]).

By using this type of extension, we derive and review three series of divergences, the Alpha, Beta- and Gamma- divergences all of which are generalizations of the KL-divergence. Moreover, we show its relation to the Tsallis entropy and the Rényi entropy (see Appendixes A and B). It will be also shown how the Alpha-divergences are derived from the Csiszár–Morimoto f-divergence and the Beta-divergence form the Bregman divergence by using the power functions.

Similarly to work of Zhang [22–24] and Hein and Bousquet [21] one of our motivations is to show the close links and relations among wide class of divergences and provide an elegant way to handle the known divergences and intermediate and new one in the same framework. However, our similarity measures are different from these proposed in [21] and our approach and results are quite different to these presented in [22].

It should be mentioned that there has been previous attempts of unifying divergence functions, (especially related to Alpha-divergence) starting from the work by Zhu and Rohwer [10,11], Amari and Nagaoka [3], Taneya [38,39], Zhang ([22] and Gorban and Judge [40]. In particular, Zhang in ([22], and in subsequent works [23,24] investigated the deep link between information

geometry and various divergence functions mostly related to Alpha-divergence through a unified approach based on convex functions. However, the previous works have not considered explicitly links and relationships among ALL three fundamental classes (Alpha-, Beta, Gamma-) divergences. Moreover, some their basic properties are reviewed and extended.

The scope of the results presented in this paper is vast since the class of generalized (flexible) divergence functions include a large number of useful loss functions containing those based on the relative entropies, generalized Kullback–Leibler or I-divergence, Hellinger distance, Jensen–Shannon divergence, J-divergence, Pearson and Neyman Chi-square divergences, Triangular Discrimination and Arithmetic-Geometric divergence. Moreover, we show that some new divergences can be generated. Especially, we generate a new family of Alpha-Gamma divergences and Itakura–Saito like distances with the invariant scaling property which belongs to a wider class of Beta-divergences. Generally, these new scale-invariant divergences provide extension of the families of Beta- and Gamma- divergences. The discussed in this paper divergence functions are flexible because they allow us to generate a large number of well known and often used particular divergences (for specific values of tuning parameters). Moreover, by adjusting adaptive tuning parameters, we can optimize cost functions for learning algorithms and estimate desired parameters of a model in presence of noise and outliers. In other words, the discussed in this paper divergences, especially Beta- and Gama- divergences are robust in respect to outliers for some values of tuning parameters.

One of important features of the considered family of divergences is that they can give some guidance for the selection and even development of new divergence measures if necessary and allows to unify these divergences under the same framework using the Csiszár-Morimoto and Bregamn divergences and their fundamental properties. Moreover, these families of divergences are generally defined on unnormalized finite measures (not necessary normalized probabilities). This allows us to analyze patterns of different size to be weighted differently, e.g., images with different sizes or documents of different length. Such measures play also an important role in the areas of neural computation, pattern recognition, learning, estimation, inference, and optimization. We have already successfully applied a subset of such divergences as cost functions (possibly with additional constraints and regularization terms) to derive novel multiplicative and additive projected gradient algorithms for nonnegative matrix and tensor factorizations [30,31].

The divergences are closely related to the invariant geometrical properties of the manifold of probability distributions [5–7].

2. Family of Alpha-Divergences

The Alpha-divergences can be derived from the Csiszár–Morimoto f-divergence and as shown recently by Amari using some tricks also from the Bregman divergence [3,6] (see also Appendix A). The Alpha-divergence was proposed by Chernoff [41] and have been extensively investigated and extended by Amari [1,3,5,6] and other researchers. For some modifications and/or extensions see for example works of Liese and Vajda [36], Minka [42], Taneja [38,39,43], Cressie–Read [44], Zhu–Rohwer [10,11] and Zhang [22–25]. One of our motivation to investigate and explore the family of Alpha-divergences is to develop flexible and efficient learning algorithms for nonnegative matrix and tensor factorizations which unify and extend existing algorithms including such algorithms as EMML

(Expectation Maximization Maximum Likelihood) and ISRA (Image Space Reconstruction Algorithm) (see [30,45] and references therein).

2.1. Asymmetric Alpha-Divergences

The basic asymmetric Alpha-divergence can be defined as [3]:

$$D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{\alpha(\alpha-1)} \int \left(p^\alpha(\boldsymbol{x})\, q^{1-\alpha}(\boldsymbol{x}) - \alpha\, p(\boldsymbol{x}) + (\alpha-1)\, q(\boldsymbol{x}) \right) d\mu(\boldsymbol{x}), \qquad \alpha \in \mathbb{R} \backslash \{0,1\} \quad (4)$$

where $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ do not need to be normalized.

The Alpha-divergence can be expressed via a generalized KL divergence as follows

$$D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{GKL}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = -\frac{1}{\alpha} \int \left(p \log_\alpha \left(\frac{q}{p} \right) + p - q \right) d\mu(\boldsymbol{x}), \;\; \alpha \in \mathbb{R} \setminus \{0,1\} \quad (5)$$

For discrete probability measures with mass functions $\mathbf{P} = [p_1, p_2, \ldots, p_n]$ and $\mathbf{Q} = [q_1, q_2, \ldots, q_n]$, the discrete Alpha-divergence is formulated as a separable divergence:

$$D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \sum_{i=1}^n d_A^{(\alpha)}(p_i \,||\, q_i) = \frac{1}{\alpha(\alpha-1)} \sum_{i=1}^n \left(p_i^\alpha\, q_i^{1-\alpha} - \alpha\, p_i + (\alpha-1)\, q_i \right), \quad \alpha \in \mathbb{R} \backslash \{0,1\} \quad (6)$$

where $d_A^{(\alpha)}(p_i \,||\, q_i) = \left(p_i^\alpha q_i^{1-\alpha} - \alpha p_i + (\alpha-1) q_i \right) / \left(\alpha(\alpha-1) \right).$

Note that this form of Alpha-divergence differs slightly from the loss function given by [1], because it was defined only for probability distributions, *i.e.*, under assumptions that $\int p(\boldsymbol{x}) d\mu(\boldsymbol{x}) = 1$ and $\int q(\boldsymbol{x}) d\mu(\boldsymbol{x}) = 1$. It was extended by Zhu and Rohwer in [10,11] (see also Amari and Nagaoka [3]) for positive measures, by incorporating additional terms. These terms are needed to allow de-normalized densities (positive measures), in the same way that the generalized Kullback–Leibler divergences (I-divergence).

Extending the Kullback–Leibler divergence into the family of Alpha-divergence is a crucial step for a unified view of a wide class of divergence functions, since it demonstrates that the nonnegativity of alpha-divergences may be viewed as arising not only from Jensen's inequality but the arithmetic-geometric inequality. As has been pointed out by Jun Zhang this can be fully exploited as a consequence of the inequality of convex functions or more generally by his convex-based approach [22].

For normalized densities: $\bar{p}(\boldsymbol{x}) = p(\boldsymbol{x})/\int p(\boldsymbol{x}) d\mu(\boldsymbol{x})$ and $\bar{q}(\boldsymbol{x}) = q(\boldsymbol{x})/\int q(\boldsymbol{x}) d\mu(\boldsymbol{x})$ the Alpha-divergence simplifies to [1,44]:

$$D_A^{(\alpha)}(\bar{\mathbf{P}} \,||\, \bar{\mathbf{Q}}) = \frac{1}{\alpha(\alpha-1)} \left(\int \bar{p}^\alpha(\boldsymbol{x})\, \bar{q}^{1-\alpha}(\boldsymbol{x}) d\mu(\boldsymbol{x}) - 1 \right), \qquad \alpha \in \mathbb{R} \setminus \{0,1\} \quad (7)$$

and is related to the Tsallis divergence and the Tsallis entropy [46] (see Appendix A):

$$H_T^{(\alpha)}(\bar{\mathbf{P}}) = \frac{1}{1-\alpha} \left(\int \bar{p}^\alpha(\boldsymbol{x})\, d\mu(\boldsymbol{x}) - 1 \right) = -\int \bar{p}^\alpha(\boldsymbol{x}) \log_\alpha \bar{p}(\boldsymbol{x}) d\mu(\boldsymbol{x}) \quad (8)$$

provided integrals on the right exist.

In fact, the Tsallis entropy was first defined by Havrda and Charvat in 1967 [47] and almost forgotten and rediscovered by Tsallis in 1988 [46] in different context (see Appendix A).

Various authors use the parameter α in different ways. For example, using the Amari notation, α_A with $\alpha = (1 - \alpha_A)/2$, the Alpha-divergence takes the following form [1–3,10,11,22–25,44,48,49]:

$$\tilde{D}_A^{(\alpha_A)}(\underline{P} \,||\, \underline{Q}) = \frac{4}{1 - \alpha_A^2} \int \left(\frac{1 - \alpha_A}{2} p + \frac{1 + \alpha_A}{2} q - p^{\frac{1 - \alpha_A}{2}} q^{\frac{1 + \alpha_A}{2}} \right) d\mu(\boldsymbol{x}), \quad \alpha_A \in \mathbb{R}\backslash\{\pm 1\} \quad (9)$$

When α takes values from 0 to 1, α_A takes values from -1 to 1. The duality exists between α_A and $-\alpha_A$, in the sense that $\tilde{D}_A^{(\alpha_A)}(\mathbf{P} \,||\, \mathbf{Q}) = \tilde{D}_A^{(-\alpha_A)}(\mathbf{Q} \,||\, \mathbf{P})$.

In the special cases for $\alpha = 2, 0.5, -1$, we obtain from (4) the well known Pearson Chi-square, Hellinger and inverse Pearson, also called the Neyman Chi-square distances, given respectively by

$$D_A^{(2)}(\mathbf{P} \,||\, \mathbf{Q}) = \quad D_P(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \int \frac{(p(\boldsymbol{x}) - q(\boldsymbol{x}))^2}{q(\boldsymbol{x})} d\mu(\boldsymbol{x})$$

$$D_A^{(1/2)}(\mathbf{P} \,||\, \mathbf{Q}) = \quad 2D_H(\mathbf{P} \,||\, \mathbf{Q}) = 2 \int (\sqrt{p(\boldsymbol{x})} - \sqrt{q(\boldsymbol{x})})^2 d\mu(\boldsymbol{x}) \qquad (10)$$

$$D_A^{(-1)}(\mathbf{P} \,||\, \mathbf{Q}) = \quad D_N(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \int \frac{(p(\boldsymbol{x}) - q(\boldsymbol{x}))^2}{p(\boldsymbol{x})} d\mu(\boldsymbol{x})$$

For the singular values $\alpha = 1$ and $\alpha = 0$ the Alpha-divergences (4) have to be defined as limiting cases respectively for $\alpha \to 1$ and $\alpha \to 0$. When this limit is evaluated (using the L'Hôpital's rule) for $\alpha \to 1$, we obtain the Kullback–Leibler divergence:

$$\begin{aligned} D_{KL}(\mathbf{P} \,||\, \mathbf{Q}) &= \lim_{\alpha \to 1} D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \lim_{\alpha \to 1} \int \frac{p^\alpha q^{1-\alpha} \log p - p^\alpha q^{1-\alpha} \log q - p + q}{2\alpha - 1} d\mu(\boldsymbol{x}) \\ &= \int \left(p \, \log\left(\frac{p}{q}\right) - p + q \right) d\mu(\boldsymbol{x}) \end{aligned} \qquad (11)$$

with the conventions $0/0 = 0$, $0\log(0) = 0$ and $p/0 = \infty$ for $p > 0$.

From the inequality $p \log p \geq p - 1$, it follows that the KL I-divergence is nonnegative and achieves zero if and only if $\mathbf{P} = \mathbf{Q}$.

Similarly, for $\alpha \to 0$, we obtain the reverse Kullback–Leibler divergence:

$$D_{KL}(\mathbf{Q}||\mathbf{P}) = \lim_{\alpha \to 0} D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left(q\log\left(\frac{q}{p}\right) - q + p \right) d\mu(\boldsymbol{x}) \qquad (12)$$

Hence, the Alpha-divergence can be evaluated in a more explicit form as

$$D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \begin{cases} \dfrac{1}{\alpha(\alpha - 1)} \displaystyle\int \left(q(\boldsymbol{x}) \left[\left(\dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^\alpha - 1 \right] - \alpha\,[p(\boldsymbol{x}) - q(\boldsymbol{x})] \right) d\mu(\boldsymbol{x}), & \alpha \neq 0, 1 \\[3mm] \displaystyle\int \left(q(\boldsymbol{x}) \log \dfrac{q(\boldsymbol{x})}{p(\boldsymbol{x})} + p(\boldsymbol{x}) - q(\boldsymbol{x}) \right) d\mu(\boldsymbol{x}), & \alpha = 0 \\[3mm] \displaystyle\int \left(p(\boldsymbol{x}) \log \dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - p(\boldsymbol{x}) + q(\boldsymbol{x}) \right) d\mu(\boldsymbol{x}), & \alpha = 1 \end{cases} \qquad (13)$$

In fact, the Alpha-divergence smoothly connects the I-divergence $D_{KL}(\mathbf{P} \,||\, \mathbf{Q})$ with the reverse I-divergence $D_{KL}(\mathbf{Q} \,||\, \mathbf{P})$ and passes through the Hellinger distance [50]. Moreover, it also

smoothly connects the Pearson Chi-square and Neyman Chi-square divergences and passes through the I-divergences [10,11].

The Alpha-divergence is a special case of Csiszár–Morimoto f-divergence [17,19] proposed later by Ali and Silvey [20]). This class of divergences were also independently defined by Morimoto [51]. The Csiszár–Morimoto f-divergence is associated to any function $f(u)$ that is convex over $(0, \infty)$ and satisfies $f(1) = 0$:

$$D_f(\mathbf{P} \,||\, \mathbf{Q}) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) \tag{14}$$

We define $0f(0/0) = 0$ and $0f(a/0) = \lim_{t \to 0} tf(a/t) = \lim_{u \to \infty} f(u)/u$. Indeed, assuming $f(u) = (u^\alpha - \alpha u + \alpha - 1))/(\alpha^2 - \alpha)$ yields the formula (4).

The Csiszár–Morimoto f-divergence has many beautiful properties [6,17,18,20,22,36,52,53]:

- **Nonnegativity:** The Csiszár–Morimoto f-divergence is always nonnegative, and equal to zero if and only if probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ coincide. This follows immediately from the Jensens inequality (for normalized densities):

$$D_f(\mathbf{P}||\mathbf{Q}) = \int qf\left(\frac{p}{q}\right) d\mu(\boldsymbol{x}) \geq f\left(\int q(\frac{p}{q}) d\mu(\boldsymbol{x})\right) = f(1) = 0. \tag{15}$$

- **Generalized entropy:** It corresponds to a generalized f-entropy of the form:

$$H_f(\mathbf{P}) = -\int f(p(\boldsymbol{x})) d\mu(\boldsymbol{x}) \tag{16}$$

for which the Shannon entropy is a special case for $f(p) = p \log(p)$. Note that H_f is concave while f is convex.

- **Convexity:** For any $0 \leq \lambda \leq 1$

$$D_f(\lambda \mathbf{P}_1 + (1 - \lambda)\mathbf{P}_2 \,||\, \lambda \mathbf{Q}_1 + (1 - \lambda)\mathbf{Q}_2) \leq \lambda D_f(\mathbf{P}_1||\mathbf{Q}_1) + (1 - \lambda)D_f(\mathbf{P}_2||\mathbf{Q}_2). \tag{17}$$

- **Scaling:** For any positive constant $c > 0$ we have

$$cD_f(\mathbf{P}||\mathbf{Q}) = D_{cf}(\mathbf{P}||\mathbf{Q}) \tag{18}$$

- **Invariance:** The f-divergence is invariant to bijective transformations [6,20]. This means that, when $\boldsymbol{x}$ is transformed to $\boldsymbol{y}$ bijectively by

$$\boldsymbol{y} = \boldsymbol{k}(\boldsymbol{x}) \tag{19}$$

probability distribution $\mathbf{P}(\boldsymbol{x})$ changes to $\tilde{\mathbf{P}}(\boldsymbol{y})$. However,

$$D_f(\mathbf{P}||\mathbf{Q}) = D_f(\tilde{\mathbf{P}}||\tilde{\mathbf{Q}}) \tag{20}$$

Additionally,

$$D_f(\mathbf{P}||\mathbf{Q}) = D_{\tilde{f}}(\mathbf{P}||\mathbf{Q}) \tag{21}$$

for $\tilde{f}(u) = f(u) - c(u-1)$ for an arbitrary constant c, and

$$D_f(\mathbf{P}||\mathbf{Q}) = D_{f^*}(\mathbf{Q}||\mathbf{P}) \tag{22}$$

where $f^*(u) = uf(1/u)$ is called a conjugate function.

- **Symmetricity:** For an arbitrary Csiszár–Morimoto f-divergence, it is possible to construct a symmetric divergence for $f_{sym}(u) = f(u) + f^*(u)$.

- **Boundedness** The Csiszár–Morimoto f-divergence for positive measures (densities) is bounded (if limit exists and it is finite) [28,36]

$$0 \leq D_f(\mathbf{P}||\mathbf{Q}) \leq \lim_{u \to 0^+} \left\{ f(u) + uf(\frac{1}{u}) \right\} \tag{23}$$

Furthermore [54],

$$0 \leq D_f(\mathbf{P}||\mathbf{Q}) \leq \int (p-q)f'(\frac{p}{q})d\mu(\boldsymbol{x}) \tag{24}$$

Using fundamental properties of the Csiszár–Morimoto f-divergence we can establish basic properties of the Alpha-divergences [3,6,22,40,42].

The Alpha-divergence (4) has the following basic properties:

1. **Convexity:** $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q})$ is convex with respect to both $\mathbf{P}$ and $\mathbf{Q}$.

2. **Strict Positivity:** $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q}) \geq 0$ and $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q}) = 0$ if and only if $\mathbf{P} = \mathbf{Q}$.

3. **Continuity:** The Alpha-divergence is continuous function of real variable α in the whole range including singularities.

4. **Duality:** $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q}) = D_A^{(1-\alpha)}(\mathbf{Q}||\mathbf{P})$.

5. **Exclusive/Inclusive Properties:** [42]

   - For $\alpha \to -\infty$, the estimation of $q(\boldsymbol{x})$ that approximates $p(\boldsymbol{x})$ is exclusive, that is $q(\boldsymbol{x}) \leq p(\boldsymbol{x})$ for all $\boldsymbol{x}$. This means that the minimization of $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q})$ with respect to $q(\boldsymbol{x})$ will force $q(\boldsymbol{x})$ to be exclusive approximation, *i.e.*, the mass of $q(\boldsymbol{x})$ will lie within $p(\boldsymbol{x})$ (see detail and graphical illustrations in [42]).

   - For $\alpha \to \infty$, the estimation of $q(\boldsymbol{x})$ that approximates $p(\boldsymbol{x})$ is inclusive, that is $q(\boldsymbol{x}) \geq p(\boldsymbol{x})$ for all $\boldsymbol{x}$. In other words, the mass of $q(\boldsymbol{x})$ includes all the mass of $p(\boldsymbol{x})$.

6. **Zero-forcing and zero-avoiding properties:** [42]

   Here, we treat the case where $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are not necessary mutually absolutely continuous. In such a case the divergence may diverges to ∞. However, the following two properties hold:

   - For $\alpha \leq 0$ the estimation of $q(\boldsymbol{x})$ that approximates $p(\boldsymbol{x})$ is zero-forcing (coercive), that is, $p(\boldsymbol{x}) = 0$ forces $q(\boldsymbol{x}) = 0$.

- For $\alpha \geq 1$ the estimation of $q(\boldsymbol{x})$ that approximates $p(\boldsymbol{x})$ is zero-avoiding, that is, $p(\boldsymbol{x}) > 0$ implies $q(\boldsymbol{x}) > 0$.

One of the most important property of the Alpha-divergence is that it is a convex function with respect to $\mathbf{P}$ and $\mathbf{Q}$ and has a unique minimum for $\mathbf{P} = \mathbf{Q}$ (see e.g. [22]).

2.2. Alpha-Rényi Divergence

It is interesting to note that the Alpha-divergence is closely related to the Rényi divergence. We define an Alpha-Rényi divergence as

$$
\begin{aligned}
D_{AR}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) &= \frac{1}{\alpha(\alpha - 1)} \log\left(1 + \alpha(\alpha - 1) D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q})\right) \\
&= \frac{1}{\alpha(\alpha - 1)} \log\left(\int \left(p^\alpha q^{1-\alpha} - \alpha p + (\alpha - 1)q\right) d\mu(\boldsymbol{x}) + 1\right), \quad \alpha \in \mathbb{R} \setminus \{0, 1\}
\end{aligned}
\tag{25}
$$

For $\alpha = 0$ and $\alpha = 1$ the Alpha-Rényi divergence simplifies to the Kullback–Leibler divergences:

$$
\begin{aligned}
D_{KL}(\mathbf{P} \,||\, \mathbf{Q}) = \lim_{\alpha \to 1} D_{AR}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) &= \lim_{\alpha \to 1} \frac{1}{2\alpha - 1} \frac{\int (p^\alpha q^{1-\alpha}(\log p - \log q) - p + q) d\mu(\boldsymbol{x})}{\int (p^\alpha q^{1-\alpha} - \alpha p - (\alpha - 1)q) d\mu(\boldsymbol{x}) + 1} \\
&= \int \left(p \, \log\left(\frac{p}{q}\right) - p + q\right) d\mu(\boldsymbol{x})
\end{aligned}
\tag{26}
$$

and

$$
D_{KL}(\mathbf{Q}||\mathbf{P}) = \lim_{\alpha \to 0} D_{AR}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left(q \log\left(\frac{q}{p}\right) - q + p\right) d\mu(\boldsymbol{x})
\tag{27}
$$

We define the Alpha-Rényi divergence for normalized probability densities as follows:

$$
D_{AR}^{(\alpha)}(\bar{\mathbf{P}} \,||\, \bar{\mathbf{Q}}) = \frac{1}{\alpha(\alpha - 1)} \log\left(\int \bar{p}^\alpha \bar{q}^{1-\alpha} d\mu(\boldsymbol{x})\right) = \frac{1}{\alpha - 1} \log\left(\int \bar{q}\left(\frac{\bar{p}}{\bar{q}}\right)^\alpha d\mu(\boldsymbol{x})\right)^{\frac{1}{\alpha}}
\tag{28}
$$

which corresponds to the Rényi entropy [55–57] (see Appendix B)

$$
H_R^{(\alpha)}(\bar{\mathbf{P}}) = -\frac{1}{\alpha - 1} \log\left(\int \bar{p}^\alpha(\boldsymbol{x}) d\mu(\boldsymbol{x})\right)
\tag{29}
$$

Note that we used different scaling factor than in the original Rényi divergence [56,57]. In general, the Alpha-Rényi divergence make sense only for the (normalized) probability densities (since, otherwise the function (25) can be complex-valued for some positive non-normalized measures).

Furthermore, the Alpha-divergence is convex with respect to positive densities p and q for any parameter of $\alpha \in \mathbb{R}$, while the Alpha-Rényi divergence (28) is convex jointly in p and q for $\alpha \in [0, 1]$ and it is not convex for $\alpha > 1$. Convexity implies that $D_{AR}^{(\alpha)}$ is increasing in α when p and q are fixed. Actually, the Rényi divergence is increasing in α on the whole set $(0, \infty)$ [58–60].

2.3. Extended Family of Alpha-Divergences

There are several ways to extend the asymmetric Alpha-divergence. For example, instead of q, we can take the average of q and p, under the assumption that if p and q are similar, they should be "close" to their average, so we can define the modified Alpha-divergence as

$$D_{Am1}^{(\alpha)}(\mathbf{P} \,||\, \widetilde{\mathbf{Q}}) = \frac{1}{\alpha(\alpha-1)} \int \left(\widetilde{q} \left[\left(\frac{p}{\widetilde{q}} \right)^{\alpha} - 1 \right] - \alpha \, (p - \widetilde{q}) \right) d\mu(\boldsymbol{x}) \tag{30}$$

whereas an adjoint Alpha-divergence is given by

$$D_{Am2}^{(\alpha)}(\widetilde{\mathbf{Q}} \,||\, \mathbf{P}) = \frac{1}{\alpha(\alpha-1)} \int \left(p \left[\left(\frac{\widetilde{q}}{p} \right)^{\alpha} - 1 \right] + \alpha \, (p - \widetilde{q}) \right) d\mu(\boldsymbol{x}) \tag{31}$$

where $\widetilde{\mathbf{Q}} = (\mathbf{P} + \mathbf{Q})/2$ and $\widetilde{q} = (p + q)/2$.

For the singular values $\alpha = 1$ and $\alpha = 0$, the Alpha-divergences (30) can be evaluated as

$$\lim_{\alpha \to 0} D_{Am1}^{(\alpha)}(\mathbf{P} \,||\, \widetilde{\mathbf{Q}}) = \int \left(\widetilde{q} \, \log \left(\frac{\widetilde{q}}{p} \right) + p - \widetilde{q} \right) d\mu(\boldsymbol{x}) = \int \left(\frac{p+q}{2} \, \log \left(\frac{p+q}{2p} \right) + \frac{p-q}{2} \right) d\mu(\boldsymbol{x})$$

and

$$\lim_{\alpha \to 1} D_{Am1}^{(\alpha)}(\mathbf{P} \,||\, \widetilde{\mathbf{Q}}) = \int \left(p \log \left(\frac{p}{\widetilde{q}} \right) - p + \widetilde{q} \right) d\mu(\boldsymbol{x}) = \int \left(p \log \left(\frac{2p}{p+q} \right) + \frac{q-p}{2} \right) d\mu(\boldsymbol{x})$$

As examples, we consider the following prominent cases for (31):

1. Triangular Discrimination (TD) [61]

$$D_{Am2}^{(-1)}(\widetilde{\mathbf{Q}} \,||\, \mathbf{P}) = \frac{1}{4} D_{TD}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{4} \int \frac{(p-q)^2}{p+q} d\mu(\boldsymbol{x}) \tag{32}$$

2. Relative Jensen–Shannon divergence [62–64]

$$\lim_{\alpha \to 0} D_{Am2}^{(\alpha)}(\widetilde{\mathbf{Q}} \,||\, \mathbf{P}) = D_{RJS}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left(p \log \left(\frac{2p}{p+q} \right) - p + q \right) d\mu(\boldsymbol{x}) \tag{33}$$

3. Relative Arithmetic-Geometric divergence proposed by Taneya [38,39,43]

$$\lim_{\alpha \to 1} D_{Am2}^{(\alpha)}(\widetilde{\mathbf{Q}} \,||\, \mathbf{P}) = \frac{1}{2} D_{RAG}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left((p+q) \log \left(\frac{p+q}{2p} \right) + p - q \right) d\mu(\boldsymbol{x}) \tag{34}$$

3. Neyman Chi-square divergence

$$D_{Am2}^{(2)}(\widetilde{\mathbf{Q}} \,||\, \mathbf{P}) = \frac{1}{8} D_{\chi^2}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{8} \int \frac{(p-q)^2}{p} d\mu(\boldsymbol{x}) \tag{35}$$

The asymmetric Alpha-divergences can be expressed formally as the Csiszár–Morimoto f-divergence, as shown in Table 1 [30].

Table 1. Asymmetric Alpha-divergences and associated convex Csiszár-Morimoto functions [30].

Divergence $D_A^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) = \int q f^{(\alpha)}\left(\frac{p}{q}\right) d\mu(\boldsymbol{x})$	Csiszár function $f^{(\alpha)}(u), \quad u = p/q$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(q\left[\left(\dfrac{p}{q}\right)^{\alpha} - 1\right] - \alpha(p-q)\right) d\mu(\boldsymbol{x}), \\ \int \left(q\log\dfrac{q}{p} + p - q\right) d\mu(\boldsymbol{x}), \\ \int \left(p\log\dfrac{p}{q} - p + q\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{u^{\alpha} - 1 - \alpha(u-1)}{\alpha(\alpha-1)}, & \alpha \neq 0, 1 \\[2mm] u - 1 - \log u, & \alpha = 0, \\[2mm] 1 - u + u\log u, & \alpha = 1. \end{cases}$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p\left[\left(\dfrac{q}{p}\right)^{\alpha} - 1\right] + \alpha(p-q)\right) d\mu(\boldsymbol{x}), \\ \int \left(p\log\dfrac{p}{q} - p + q\right) d\mu(\boldsymbol{x}), \\ \int \left(q\log\dfrac{q}{p} + p - q\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{u^{1-\alpha} + (\alpha-1)u - \alpha}{\alpha(\alpha-1)}, & \alpha \neq 0, 1, \\[2mm] 1 - u + u\log u, & \alpha = 0 \\[2mm] u - 1 - \log u, & \alpha = 1. \end{cases}$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(q\left(\dfrac{p+q}{2q}\right)^{\alpha} - q - \alpha\dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \\ \int \left(q\log\left(\dfrac{2q}{p+q}\right) + \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \\ \int \left(\dfrac{p+q}{2}\log\left(\dfrac{p+q}{2q}\right) - \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{\left(\dfrac{u+1}{2}\right)^{\alpha} - 1 - \alpha\left(\dfrac{u-1}{2}\right)}{\alpha(\alpha-1)}, & \alpha \neq 0, 1, \\[2mm] \dfrac{u-1}{2} + \log\left(\dfrac{2}{u+1}\right), & \alpha = 0, \\[2mm] \dfrac{1-u}{2} + \dfrac{u+1}{2}\log\left(\dfrac{u+1}{2}\right), & \alpha = 1. \end{cases}$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p\left(\dfrac{p+q}{2p}\right)^{\alpha} - p + \alpha\dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \\ \int \left(p\log\left(\dfrac{2p}{p+q}\right) - \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \\ \int \left(\dfrac{p+q}{2}\log\left(\dfrac{p+q}{2p}\right) + \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{u\left(\dfrac{u+1}{2u}\right)^{\alpha} - u - \alpha\left(\dfrac{1-u}{2}\right)}{\alpha(\alpha-1)}, & \alpha \neq 0, 1, \\[2mm] \dfrac{1-u}{2} - u\log\left(\dfrac{u+1}{2u}\right), & \alpha = 0, \\[2mm] \dfrac{u-1}{2} + \left(\dfrac{u+1}{2}\right)\log\left(\dfrac{u+1}{2u}\right), & \alpha = 1. \end{cases}$
$\begin{cases} \dfrac{1}{\alpha-1} \int (p-q)\left[\left(\dfrac{p+q}{2q}\right)^{\alpha-1} - 1\right] d\mu(\boldsymbol{x}), \\ \int (p-q)\log\left(\dfrac{p+q}{2q}\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{(u-1)\left[\left(\dfrac{u+1}{2}\right)^{\alpha-1} - 1\right]}{\alpha-1}, & \alpha \neq 1, \\[2mm] (u-1)\log\left(\dfrac{u+1}{2}\right), & \alpha = 1. \end{cases}$
$\begin{cases} \dfrac{1}{\alpha-1} \int (q-p)\left[\left(\dfrac{p+q}{2p}\right)^{\alpha-1} - 1\right] d\mu(\boldsymbol{x}), \\ \int (q-p)\log\left(\dfrac{p+q}{2p}\right) d\mu(\boldsymbol{x}). \end{cases}$	$\begin{cases} \dfrac{(1-u)\left[\left(\dfrac{u+1}{2u}\right)^{\alpha-1} - 1\right]}{\alpha-1}, & \alpha \neq 1, \\[2mm] (1-u)\log\left(\dfrac{u+1}{2u}\right), & \alpha = 1. \end{cases}$

2.4. *Symmetrized Alpha-Divergences*

The basic Alpha-divergence is asymmetric, that is, $D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) \neq D_A^{(\alpha)}(\mathbf{Q} \,||\, \mathbf{P})$.
Generally, there are two ways to symmetrize divergences: Type-1

$$D_{S1}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \left[D_A(\mathbf{P} \,||\, \mathbf{Q}) + D_A(\mathbf{Q} \,||\, \mathbf{P}) \right] \tag{36}$$

and Type 2

$$D_{S2}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \left[D_A \left(\mathbf{P} \,||\, \frac{\mathbf{P} + \mathbf{Q}}{2} \right) + D_A \left(\mathbf{Q} \,||\, \frac{\mathbf{P} + \mathbf{Q}}{2} \right) \right] \tag{37}$$

The symmetric Alpha-divergence (Type-1) can be defined as (we will omit scaling factor $1/2$ for simplicity)

$$D_{AS1}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) + D_A^{(\alpha)}(\mathbf{Q} \,||\, \mathbf{P}) = \int \frac{(p^\alpha - q^\alpha)(p^{1-\alpha} - q^{1-\alpha})}{\alpha(1 - \alpha)} \, d\mu(\boldsymbol{x}) \tag{38}$$

As special cases, we obtain several well-known symmetric divergences:

1. Symmetric Chi-Squared divergence [54]

$$D_{AS1}^{(-1)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{AS1}^{(2)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} D_{\chi^2}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \int \frac{(p - q)^2 (p + q)}{pq} d\mu(\boldsymbol{x}) \tag{39}$$

2. Symmetrized KL divergence, called also J-divergence corresponding to Jeffreys entropy maximization [65,66]

$$\lim_{\alpha \to 0} D_{AS1}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \lim_{\alpha \to 1} D_{AS1}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = D_J(\mathbf{P} \,||\, \mathbf{Q}) = \int (p - q) \log \left(\frac{p}{q} \right) d\mu(\boldsymbol{x}) \tag{40}$$

An alternative wide class of symmetric divergences can be described by the following symmetric Alpha-divergence (Type-2):

$$\begin{aligned} D_{AS2}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) \ &= D_A^{(\alpha)} \left(\mathbf{P} \,||\, \frac{\mathbf{P} + \mathbf{Q}}{2} \right) + D_A^{(\alpha)} \left(\mathbf{Q} \,||\, \frac{\mathbf{P} + \mathbf{Q}}{2} \right) \\ &= \frac{1}{\alpha(\alpha - 1)} \int \left((p^{1-\alpha} + q^{1-\alpha}) \left(\frac{p + q}{2} \right)^\alpha - (p + q) \right) d\mu(\boldsymbol{x}) \end{aligned}$$

The above measure admits the following prominent cases

1. Triangular Discrimination [30,38]

$$D_{AS2}^{(-1)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} D_{TD}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} \int \frac{(p - q)^2}{p + q} d\mu(\boldsymbol{x}) \tag{41}$$

2. Symmetric Jensen–Shannon divergence [62,64]

$$\lim_{\alpha \to 0} D_{AS2}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{JS}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left(p \log \left(\frac{2p}{p + q} \right) + q \log \left(\frac{2q}{p + q} \right) \right) d\mu(\boldsymbol{x}) \tag{42}$$

It is worth mentioning, that the Jensen–Shannon divergence is a symmetrized and smoothed variant of the Kullback–Leibler divergence, *i.e.*, it can be interpreted as the average of the Kullback–Leibler divergences to the average distribution. For the normalized probability densities the Jensen–Shannon divergence is related to the Shannon entropy in the following sense:

$$D_{JS} = H_S((\mathbf{P} + \mathbf{Q})/2) - (H_S(\mathbf{P}) + H_S(\mathbf{Q}))/2 \tag{43}$$

where $H_S(\mathbf{P}) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\mu(\boldsymbol{x})$

3. Arithmetic-Geometric divergence [39]

$$\lim_{\alpha \to 1} D_{AS2}^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) = \int (p + q) \log \left(\frac{p + q}{2\sqrt{pq}} \right) d\mu(\boldsymbol{x}) \tag{44}$$

4. Symmetric Chi-square divergence [54]

$$D_{AS2}^{(2)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{8} D_{\chi^2}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{8} \int \frac{(p - q)^2 (p + q)}{pq} d\mu(\boldsymbol{x}) \tag{45}$$

The above Alpha-divergence is symmetric in its arguments $\mathbf{P}$ and $\mathbf{Q}$, and it is well-defined even if $\mathbf{P}$ and $\mathbf{Q}$ are not absolutely continuous. For example, for discrete $D_{AS2}^{(\alpha)}(\mathbf{P}||\mathbf{Q})$ is well-defined even if, for some indices p_i, it vanishes without vanishing q_i or if q_i vanishes without vanishing p_i [54]. It is also lower- and upper-bounded, for example, the Jensen–Shannon divergence is bounded between 0 and 2 [36].

3. Family of Beta-Divergences

The basic Beta-divergence was introduced by Basu *et al.* [67] and Minami and Eguchi [15] and many researchers investigated their applications including [8,13,30–34,37,37,68–72], and references therein. The main motivation to investigate the beta divergence, at least from the practical point of view, is to develop highly robust in respect to outliers learning algorithms for clustering, feature extraction, classification and blind source separation. Until now the Beta-divergence has been successfully applied for robust PCA (Principal Component Analysis) and clustering [71], robust ICA (Independent Component Analysis) [15,68,69], and robust NMF/NTF [30,70,73–76].

First, let us define the basic asymmetric Beta-divergence between two unnormalized density functions by

$$D_B^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = \int \left(p(\boldsymbol{x}) \frac{p^{\beta-1}(\boldsymbol{x}) - q^{\beta-1}(\boldsymbol{x})}{\beta - 1} - \frac{p^\beta(\boldsymbol{x}) - q^\beta(\boldsymbol{x})}{\beta} \right) d\mu(\boldsymbol{x}), \quad \beta \in \mathbb{R} \setminus \{0, 1\} \tag{46}$$

where β is a real number and, for $\beta = 0, 1$, is defined by continuity (see below for more explanation).

For discrete probability measures with mass functions $\mathbf{P} = [p_1, p_2, \dots, p_n]$ and $\mathbf{Q} = [q_1, q_2, \dots, q_n]$ the discrete Beta-divergence is defined as

$$D_B^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = \sum_{i=1}^n d_B^{(\beta)}(p_i \,||\, q_i) = \sum_{i=1}^n \left(p_i \frac{p_i^{\beta-1} - q_i^{\beta-1}}{\beta - 1} - \frac{p_i^\beta - q_i^\beta}{\beta} \right) \quad \beta \in \mathbb{R} \setminus \{0, 1\} \tag{47}$$

The Beta-divergence can be expressed via a generalized KL divergence as follows

$$D_B^{(\beta)}(\mathbf{P}\,||\,\mathbf{Q}) = D_{GKL}^{(\beta-1)}(\mathbf{P}^\beta\,||\,\mathbf{Q}^\beta) = -\frac{1}{\beta}\int\left(p^\beta\log_{(\frac{1}{\beta})}\left(\frac{q^\beta}{p^\beta}\right) + p^\beta - q^\beta\right)d\mu(\boldsymbol{x}) \quad \beta\in\mathbb{R}\backslash\{0,1\} \quad (48)$$

The above representation of the Beta-divergence indicates why it is robust to outliers for some values of the tuning parameter β and therefore, it is often better suited than others for some specific applications. For example, in sound processing, the speech power spectra can be modeled by exponential family densities of the form, whose for $\beta = 0$ the Beta-divergence is no less than the Itakura–Saito distance (called also Itakura–Saito divergence or Itakura–Saito distortion measure or Burg cross entropy) [12,13,30,76–79]. In fact, the Beta-divergence has to be defined in limiting case for $\beta\to 0$ as the Itakura–Saito distance:

$$D_{IS}(\mathbf{P}\,||\,\mathbf{Q}) = \lim_{\beta\to 0}D_B^{(\beta)}(\mathbf{P}\,||\,\mathbf{Q}) = \int\left(\log\frac{q}{p} + \frac{p}{q} - 1\right)d\mu(\boldsymbol{x}) \quad (49)$$

The Itakura and Saito distance was derived from the maximum likelihood (ML) estimation of speech spectra [77]. It was used as a measure of the distortion or goodness of fit between two spectra and is often used as a standard measure in the speech processing community due to the good perceptual properties of the reconstructed signals since it is scale invariant, Due to scale invariance low energy components of p have the same relative importance as high energy ones. This is especially important in the scenario in which the coefficients of p have a large dynamic range, such as in short-term audio spectra [30,76,79].

It is also interesting to note that, for $\beta = 2$, we obtain the standard squared Euclidean (L_2-norm) distance, while for the singular case $\beta = 1$, we obtain the KL I-divergence:

$$D_{KL}(\mathbf{P}\,||\,\mathbf{Q}) = \lim_{\beta\to 1}D_B^{(\beta)}(\mathbf{P}\,||\,\mathbf{Q}) = \int\left(p\,\log\frac{p}{q} - p + q\right)d\mu(\boldsymbol{x}) \quad (50)$$

Note, that we used here, the following formulas $\lim_{\beta\to 0}\frac{p^\beta-q^\beta}{\beta} = \log(p/q)$ and $\lim_{\beta\to 0}\frac{p^\beta-1}{\beta} = \log p$

Hence, the Beta-divergence can be represented in a more explicit form:

$$D_B^{(\beta)}(\mathbf{P}\,||\,\mathbf{Q}) = \begin{cases} \dfrac{1}{\beta(\beta-1)}\displaystyle\int\left(p^\beta(\boldsymbol{x}) + (\beta-1)q^\beta(\boldsymbol{x}) - \beta p(\boldsymbol{x})q^{\beta-1}(\boldsymbol{x})\right)d\mu(\boldsymbol{x}), & \beta\neq 0,1 \\ \displaystyle\int\left(p(\boldsymbol{x})\,\log(\dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}) - p(\boldsymbol{x}) + q(\boldsymbol{x})\right)d\mu(\boldsymbol{x}), & \beta = 1 \\ \displaystyle\int\left(\log(\dfrac{q(\boldsymbol{x})}{p(\boldsymbol{x})}) + \dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right)d\mu(\boldsymbol{x}), & \beta = 0 \end{cases} \quad (51)$$

We have shown that the basic Beta-divergence smoothly connects the Itakura–Saito distance and the squared Euclidean L_2-norm distance and passes through the KL I-divergence $D_{KL}(\mathbf{P}\,||\,\mathbf{Q})$. Such a parameterized connection is impossible in the family of the Alpha-divergences.

The choice of the tuning parameter β depends on the statistical distribution of data sets. For example, the optimal choice of the parameter β for the normal distribution is $\beta = 2$, for the gamma distribution it is $\beta = 0$, for the Poisson distribution $\beta = 1$, and for the compound Poisson distribution $\beta\in(1,2)$ [15,31–34,68,69].

It is important to note that the Beta divergence can be derived from the Bregman divergence. The Bregman divergence is a pseudo-distance for measuring discrepancy between two values of density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ [9,16,80]:

$$d_\Phi(p||q) = \Phi(p) - \Phi(q) - (p - q)\Phi'(q) \tag{52}$$

where $\Phi(t)$ is strictly convex real-valued function and $\Phi'(q)$ is the derivative with respect to q. The total discrepancy between two functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is given by

$$D_\Phi(\mathbf{P}||\mathbf{Q}) = \int \left[\Phi(p(\boldsymbol{x})) - \Phi(q(\boldsymbol{x})) - (p(\boldsymbol{x}) - q(\boldsymbol{x}))\Phi'(q(\boldsymbol{x})) \right] d\mu(\boldsymbol{x}) \tag{53}$$

and it corresponds to the Φ-entropy of continuous probability measure $p(\boldsymbol{x}) \geq 0$ defined by

$$H_\Phi(\mathbf{P}) = - \int \Phi(p(x)) d\mu(\boldsymbol{x}) \tag{54}$$

Remark: The concept of divergence and entropy are closely related. Let $\mathbf{Q}_0$ be a uniform distribution for which

$$\mathbf{Q}_0(\boldsymbol{x}) = const \tag{55}$$

(When $\boldsymbol{x}$ is an infinite space $\mathbb{R}^n$ this might not be a probability distribution but is a measure). Then,

$$H(\mathbf{P}) = -D(\mathbf{P}||\mathbf{Q}_0) + const \tag{56}$$

is regarded as the related entropy. This the negative of the divergence form $\mathbf{P}$ to the uniform distribution. On the other hand, given a concave entropy $H(\mathbf{P})$, we can define the related divergence as the Bregman divergence derived from a convex function $\Phi(\mathbf{P}) = -H(\mathbf{P})$.

If $\boldsymbol{x}$ takes discrete values on a certain space, the separable Bregman divergence is defined as $D_\Phi(\mathbf{P}||\mathbf{Q}) = \sum_{i=1} d_\Phi(p_i||q_i) = \sum_{i=1}^n [\Phi(p_i) - \Phi(q_i) - (p_i - q_i)\Phi'(q_i)]$, where $\Phi'(q)$ denotes derivative with respect to q. In a general (nonseparable) case for two vectors $\mathbf{P}$ and $\mathbf{Q}$, the Bregman divergence is defined as $D_\Phi(\mathbf{P}||\mathbf{Q}) = \Phi(\mathbf{P}) - \Phi(\mathbf{Q}) - (\mathbf{P} - \mathbf{Q})^T \nabla\Phi(\mathbf{Q})$, where $\nabla\Phi(\mathbf{Q})$ is the gradient of Φ evaluated at $\mathbf{Q}$.

Note that $D_\Phi(\mathbf{P}||\mathbf{Q})$ equals the tail of the first-order Taylor expansion of $\Phi(\mathbf{P})$ at $\mathbf{Q}$. Bregman divergences include many prominent dissimilarity measures like the squared Euclidean distance, the Mahalanobis distance, the generalized Kullback–Leibler divergence and the Itakura–Saito distance.

It is easy to check that the Beta-divergence can be generated from the Bregman divergence using the following strictly convex continuous function [36,81]

$$\Phi(t) = \begin{cases} \dfrac{1}{\beta(\beta - 1)} \left(t^\beta - \beta t + \beta - 1 \right), & \beta \neq 0, 1 \\ t \log(t) - t + 1, & \beta = 1 \\ t - \log(t) - 1, & \beta = 0 \end{cases} \tag{57}$$

It is also interesting to note that the same generating function $f(u) = \Phi(t)_{|t=u}$ (with $\alpha = \beta$) can be used to generate the Alpha-divergence using the Csiszár–Morimoto f-divergence $D_f(\mathbf{P}||\mathbf{Q}) = \int q f(p/q) d\mu(\boldsymbol{x})$.

Furthermore, the Beta-divergence can be generated by a generalized f-divergence:

$$\widetilde{D}_f(\mathbf{P}||\mathbf{Q}) = \int q^\beta f(p/q)\, d\mu(\boldsymbol{x}) = \int q^\beta \tilde{f}\left(\frac{p^\beta}{q^\beta}\right)\, d\mu(\boldsymbol{x}) \tag{58}$$

where $f(u) = \Phi(t)_{|t=u}$ with $u = p/q$ and

$$\tilde{f}(\tilde{u}) = \frac{1}{1-\beta}\left(\tilde{u}^{\frac{1}{\beta}} - \frac{1}{\beta}\tilde{u} + \frac{1}{\beta} - 1\right) \tag{59}$$

is convex generating function with $\tilde{u} = p^\beta/q^\beta$.

The links between the Bregman and Beta-divergences are important, since the many well known fundamental properties of the Bregman divergence are also valid for the Beta-divergence [28,82]:

1. **Convexity:** The Bregman divergence $D_\Phi(\mathbf{P}||\mathbf{Q})$ is always convex in the first argument $\mathbf{P}$, but is often not in the second argument $\mathbf{Q}$.

2. **Nonnegativity:** The Bregman divergence is nonnegative $D_\Phi(\mathbf{P}||\mathbf{Q}) \geq 0$ with zero $\mathbf{P} = \mathbf{Q}$.

3. **Linearity:** Any positive linear combination of Bregman divergences is also a Bregman divergence, *i.e.*,

$$D_{c_1\Phi_1+c_2\Phi_2}(\mathbf{P}||\mathbf{Q}) = c_1 D_{\Phi_1}(\mathbf{P}||\mathbf{Q}) + c_2 D_{\Phi_2}(\mathbf{P}||\mathbf{Q}) \tag{60}$$

where c_1, c_2 are positive constants and Φ_1, Φ_2 are strictly convex functions.

4. **Invariance:** The functional Bregman divergence is invariant under affine transforms $\Gamma(\mathbf{Q}) = \Phi(\mathbf{Q}) + \int a(x - x')q(x')dx' + c$ for positive measures $\mathbf{P}$ and $\mathbf{Q}$ to linear and arbitrary constant terms [28,82], *i.e.*,

$$D_\Gamma(\mathbf{P}||\mathbf{Q}) = D_\Phi(\mathbf{P}||\mathbf{Q}) \tag{61}$$

5. **The three-point property generalizes the "Law of Cosines":**

$$D_\Phi(\mathbf{P}\,||\,\mathbf{Q}) = D_\Phi(\mathbf{P}\,||\,\mathbf{Z}) + D_\Phi(\mathbf{Z}\,||\,\mathbf{Q}) - (\mathbf{P} - \mathbf{Z})^T\left(\frac{\delta}{\delta\mathbf{Q}}\Phi(\mathbf{Q}) - \frac{\delta}{\delta\mathbf{Z}}\Phi(\mathbf{Z})\right) \tag{62}$$

6. **Generalized Pythagoras Theorem:**

$$D_\Phi(\mathbf{P}\,||\,\mathbf{Q}) \geq D_\Phi(\mathbf{P}\,||\,P_\Omega(\mathbf{Q})) + D_\Phi(P_\Omega(\mathbf{Q})\,||\,\mathbf{Q}) \tag{63}$$

where $P_\Omega(\mathbf{Q}) = \arg\min_{\boldsymbol{\omega}\in\Omega} D_\Phi(\boldsymbol{\omega}||\mathbf{Q})$ is the Bregman projection onto the convex set Ω and $\mathbf{P} \in \Omega$. When Ω is an affine set then it holds with equality. This is proved to be the generalized Pythagorean relation in terms of information geometry.

For the Beta-divergence (46) the first and second-order Fréchet derivative with respect to $\mathbf{Q}$ are given by [28,76]

$$\frac{\delta D_B^{(\beta)}}{\delta q} = q^{\beta-2}(q - p), \qquad \frac{\delta^2 D_B^{(\beta)}}{\delta q^2} = q^{\beta-3}((\beta-1)q - (\beta-2)p) \tag{64}$$

Hence, we conclude that the Beta-divergence has a single global minimum equal to zero for $\mathbf{P} = \mathbf{Q}$ and increases with $|p - q|$. Moreover, the Beta divergence is strictly convex for $q(\boldsymbol{x}) > 0$ only for $\beta \in [1, 2]$. For $\beta = 0$ (Itakura–Saito distance), it is convex if $\int q^{-3}(2p - q)d\mu(\boldsymbol{x}) \geq 0$ *i.e.*, if $q/p \leq 2$ [78].

3.1. *Generation of Family of Beta-divergences Directly from Family of Alpha-Divergences*

It should be noted that in the original works [15,67–69] they considered only the Beta-divergence function for $\beta \geq 1$. Moreover, they did not consider the whole range of non-positive values for parameter β, especially $\beta = 0$, for which we have the important Itakura–Saito distance. Furthermore, similar to the Alpha-divergences there exist an associated family of Beta-divergences and as special cases a family of generalized Itakura–Saito like distances. The fundamental question arises: How to generate a whole family of Beta-divergences or what is the relationships or correspondences between the Alpha- and Beta-divergences. In fact, on the basis of our considerations above, it is easy to find that the complete set of Beta-divergences can be obtained from the Alpha-divergences and conversely the Alpha-divergences, can obtained directly from Beta-divergences.

In order to obtain a Beta-divergence from the corresponding (associated) Alpha-divergence, we need to apply the following nonlinear transformations:

$$p \to p^{\beta}, \qquad q \to q^{\beta} \quad \text{and} \quad \alpha = \beta^{-1} \tag{65}$$

For example, using these transformations (substitutions) for a basic asymmetric Alpha-divergence (4) and assuming that $\alpha = (\beta)^{-1}$, we obtain the following divergence

$$D_A^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = \beta^2 \int \left(\frac{-\beta p\, q^{\beta-1} + p^{\beta} + (\beta - 1)\, q^{\beta}}{\beta(\beta - 1)} \right) d\mu(\boldsymbol{x}) \tag{66}$$

Observe that, by ignoring the scaling factor β^2, we obtain the basic asymmetric Beta-divergence defined by Equation (46).

In fact, there exists the same link between the whole family of Alpha-divergences and the family of Beta-divergences (see Table 2).

For example, we can derive a symmetric Beta-divergence from the symmetric Alpha-divergence (Type-1) (38):

$$
\begin{aligned}
D_{BS1}^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = \quad & D_B^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) + D_B^{(\beta)}(\mathbf{Q} \,||\, \mathbf{P}) \\
= \quad & \frac{1}{\beta - 1} \int \left((p - q)(p^{\beta-1} - q^{\beta-1}) \right) d\mu(\boldsymbol{x})
\end{aligned}
$$

It is interesting to note that, in special cases, we obtain:
Symmetric KL of J-divergence [65]:

$$D_{BS1}^{(1)} = \lim_{\beta \to 1} D_{BS1}^{(\beta)} = \int (p - q) \log\left(\frac{p}{q} \right) d\mu(\boldsymbol{x}) \tag{67}$$

and symmetric Chi-square divergence [54]

$$D_{BS1}^{(0)} = \lim_{\beta \to 0} D_{BS1}^{(\beta)} = \int \frac{(p - q)^2}{pq} d\mu(\boldsymbol{x}) \tag{68}$$

Analogously, from the symmetric Alpha-divergence (Type-2), we obtain

$$D_{BS2}^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{\beta - 1} \int \left(p^{\beta} + q^{\beta} - (p^{\beta-1} + q^{\beta-1})\left(\frac{p^{\beta} + q^{\beta}}{2} \right)^{\frac{1}{\beta}} \right) d\mu(\boldsymbol{x}) \tag{69}$$

Table 2. Family of Alpha-divergences and corresponding Beta-divergences. We applied the following transformations $p \to p^\beta$, $q \to q^\beta$, $\alpha = 1/\beta$. Note that $D_A^{(1)}(\mathbf{P}||\mathbf{Q}) = D_B^{(1)}(\mathbf{P}||\mathbf{Q})$ and they represents for $\alpha = \beta = 1$ extended family of KL divergences. Furthermore, Beta-divergences for $\beta = 0$ describe the family of generalized (extended) Itakura–Saito like distances.

Alpha-divergence $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q})$	Beta-divergence $D_B^{(\beta)}(\mathbf{P}||\mathbf{Q})$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p^\alpha q^{1-\alpha} - \alpha p + (\alpha-1) q\right) d\mu(\boldsymbol{x}), \\ \int \left(q \log\left(\dfrac{q}{p}\right) + p - q\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\ \int \left(p \log\left(\dfrac{p}{q}\right) - p + q\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$	$\begin{cases} \dfrac{1}{\beta(\beta-1)} \int \left(p^\beta + (\beta-1) q^\beta - \beta\, p\, q^{\beta-1}\right) d\mu(\boldsymbol{x}), \\ \int \left(\log\left(\dfrac{q}{p}\right) + \dfrac{p}{q} - 1\right) d\mu(\boldsymbol{x}), \quad \beta = 0 \\ \int \left(p \log\left(\dfrac{p}{q}\right) - p + q\right) d\mu(\boldsymbol{x}), \quad \beta = 1 \end{cases}$
$\begin{cases} \dfrac{\int \left((\frac{p+q}{2})^\alpha q^{1-\alpha} - \frac{\alpha}{2} p + (\frac{\alpha}{2}-1) q\right) d\mu(\boldsymbol{x})}{\alpha(\alpha-1)}, \\ \int \left(q \log\left(\dfrac{2q}{p+q}\right) + \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\ \int \left(\dfrac{p+q}{2} \log\left(\dfrac{p+q}{2q}\right) - \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$	$\begin{cases} \dfrac{\int \left(p^\beta + (2\beta-1)q^\beta - 2\beta q^{\beta-1}\left(\frac{p^\beta + q^\beta}{2}\right)^{\frac{1}{\beta}}\right) d\mu(\boldsymbol{x})}{\beta(\beta-1)}, \\ \int \left(\log(\dfrac{q}{p}) + 2(\sqrt{\dfrac{p}{q}} - 1)\right) d\mu(\boldsymbol{x}), \quad \beta = 0 \\ \int \left(\dfrac{p+q}{2} \log\left(\dfrac{p_1+q}{2q}\right) - \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \quad \beta = 1 \end{cases}$
$\begin{cases} \dfrac{1}{\alpha-1} \int (p-q)\left[\left(\dfrac{p+q}{2q}\right)^{\alpha-1} - 1\right] d\mu(\boldsymbol{x}), \\ \int \dfrac{(p-q)^2}{p+q} d\mu(\boldsymbol{x}), \quad \alpha = 0 \\ \int (p-q) \log\left(\dfrac{p+q}{2q}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$	$\begin{cases} \dfrac{1}{\beta(\beta-1)} \int (p^\beta - q^\beta)\left(1 - \dfrac{2^{\frac{\beta-1}{\beta}} q^{\beta-1}}{(p^\beta + q^\beta)^{\frac{\beta-1}{\beta}}}\right) d\mu(\boldsymbol{x}), \\ \int \left(\sqrt{\dfrac{p}{q}} - 1\right) \log\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad \beta = 0 \\ \int (p-q) \log\left(\dfrac{p+q}{2q}\right) d\mu(\boldsymbol{x}), \quad \beta = 1 \end{cases}$
$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p^\alpha q^{1-\alpha} + p^{1-\alpha} q^\alpha - p - q\right) d\mu(\boldsymbol{x}), \\ \int (p-q) \log\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad \alpha = 0, 1 \end{cases}$	$\begin{cases} \dfrac{1}{\beta-1} \int \left(p^\beta + q^\beta - pq^{\beta-1} - p^{\beta-1}q\right) d\mu(\boldsymbol{x}), \\ \int \dfrac{(p-q)^2}{pq} d\mu(\boldsymbol{x}), \quad \beta = 0 \\ \int (p-q) \log\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad \beta = 1 \end{cases}$
$\begin{cases} \dfrac{1}{1-\alpha} \int \left(\dfrac{p+q}{2} - \left(\dfrac{p^\alpha + q^\alpha}{2}\right)^{\frac{1}{\alpha}}\right) d\mu(\boldsymbol{x}), \\ \int \left(\dfrac{(\sqrt{p} - \sqrt{q})^2}{2}\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\ H_S\left(\dfrac{\mathbf{P}+\mathbf{Q}}{2}\right) - \dfrac{H_S(\mathbf{P}) + H_S(\mathbf{Q})}{2}, \quad \alpha = 1 \end{cases}$	$\begin{cases} \dfrac{1}{\beta(\beta-1)} \int \left(\dfrac{p^\beta + q^\beta}{2} - \left(\dfrac{p+q}{2}\right)^\beta\right) d\mu(\boldsymbol{x}), \\ \int \log\left(\dfrac{p+q}{2\sqrt{pq}}\right) d\mu(\boldsymbol{x}), \quad \beta = 0 \\ H_S\left(\dfrac{\mathbf{P}+\mathbf{Q}}{2}\right) - \dfrac{H_S(\mathbf{P}) + H_S(\mathbf{Q})}{2}, \quad \beta = 1 \end{cases}$
$\begin{cases} \dfrac{\int \left((p^{1-\alpha} + q^{1-\alpha})\left(\frac{p+q}{2}\right)^\alpha - p - q\right) d\mu(\boldsymbol{x})}{\alpha(\alpha-1)}, \\ \int \left(p \log\left(\dfrac{2p}{p+q}\right) + q \log\left(\dfrac{2q}{p+q}\right)\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\ \int (p+q) \log\left(\dfrac{p+q}{2\sqrt{pq}}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$	$\begin{cases} \dfrac{1}{\beta-1} \int \left(p^\beta + q^\beta - (p^{\beta-1} + q^{\beta-1})\left(\dfrac{p^\beta + q^\beta}{2}\right)^{\frac{1}{\beta}}\right) d\mu(\boldsymbol{x}), \\ \int \left(\sqrt{\dfrac{p}{q}} + \sqrt{\dfrac{q}{p}} - 2\right) d\mu(\boldsymbol{x}), \quad \beta = 0 \\ \int (p+q) \log\left(\dfrac{p+q}{2\sqrt{pq}}\right) d\mu(\boldsymbol{x}), \quad \beta = 1 \end{cases}$

It should be noted that in special cases, we obtain:
The Arithmetic-Geometric divergence [38,39]:

$$D_{BS2}^{(1)} = \lim_{\beta \to 1} D_{BS2}^{(\beta)} = \int (p+q) \log \left(\frac{p+q}{2\sqrt{pq}} \right) d\mu(\boldsymbol{x}), \tag{70}$$

and a symmetrized Itakura–Saito distance (called also the COSH distance) [12,13]:

$$D_{BS2}^{(0)} = \lim_{\beta \to 0} D_{BS2}^{(\beta)} = \int \left(\sqrt{\frac{p}{q}} + \sqrt{\frac{q}{p}} - 2 \right) d\mu(\boldsymbol{x}) = \int \frac{(\sqrt{p} - \sqrt{q})^2}{\sqrt{pq}} d\mu(\boldsymbol{x}) \tag{71}$$

4. Family of Gamma-Divergences Generated from Beta- and Alpha-Divergences

A basic asymmetric Gamma-divergence has been proposed very recently by Fujisawa and Eguchi [35] as a very robust similarity measure with respect to outliers:

$$D_G^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{\gamma(\gamma-1)} \log \left(\frac{\left(\int p^\gamma(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right) \left(\int q^\gamma(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{\gamma-1}}{\left(\int p(\boldsymbol{x}) \, q^{\gamma-1}(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^\gamma} \right) \tag{72}$$

The Gamma-divergence employs the nonlinear transformation ($\log$) for cumulative patterns and the terms p, q are not separable. The main motivation for employing the Gamma divergence is that it allows "super" robust estimation of some parameters in presence of outlier. In fact, the authors demonstrated that the bias caused by outliers can become sufficiently small even in the case of very heavy contamination and that some contamination can be naturally and automatically neglected [35,37].

In this paper, we show that we can formulate the whole family of Gamma-divergences generated directly from Alpha- and also Beta-divergences. In order to obtain a robust Gamma-divergence from an Alpha- or Beta-divergence, we use the following transformations (see also Table 3):

$$c_0 \int p^{c_1}(\boldsymbol{x}) \, q^{c_2}(\boldsymbol{x}) \, d\mu(\boldsymbol{x}) \to \log \left(\int p^{c_1}(\boldsymbol{x}) \, q^{c_2}(\boldsymbol{x}) \, d\mu(\boldsymbol{x}) \right)^{c_0} \tag{73}$$

where c_0, c_1 and c_2 are real constants and $\gamma = \alpha$.

Applying the above transformation to all monomials to the basic Alpha-divergence (4), we obtain a new divergence referred to as here the Alpha-Gamma-divergence:

$$
\begin{aligned}
D_{AG}^{(\gamma)}(\mathbf{P}||\mathbf{Q}) &= \log \left(\int p^\gamma(\boldsymbol{x}) q^{1-\gamma}(\boldsymbol{x}) \, d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma(\gamma-1)}} - \log \left(\int p(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma-1}} + \log \left(\int q(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma}} \\
&= \frac{1}{\gamma(\gamma-1)} \log \left(\frac{\int p^\gamma(\boldsymbol{x}) q^{1-\gamma}(\boldsymbol{x}) d\mu(\boldsymbol{x})}{\left(\int p(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^\gamma \left(\int q(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{1-\gamma}} \right)
\end{aligned}
$$

The asymmetric Alpha-Gamma-divergence has the following important properties:

1. $D_{AG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) \geq 0$. The equality holds if and only if $\mathbf{P} = c\mathbf{Q}$ for a positive constant c.

2. It is scale invariant for any value of γ, that is, $D_{AG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{AG}^{(\gamma)}(c_1\mathbf{P} \,||\, c_2\mathbf{Q})$, for arbitrary positive scaling constants c_1, c_2.

3. The Alpha-Gamma divergence is equivalent to the normalized Alpha-Rényi divergence (25), *i.e.*,

$$
\begin{aligned}
D_{AG}^{(\gamma)}(\mathbf{P}||\mathbf{Q}) \ &= \frac{1}{\gamma(\gamma-1)} \log \left(\frac{\int p^\gamma(\boldsymbol{x})\, q^{1-\gamma}(\boldsymbol{x}) d\mu(\boldsymbol{x})}{\left(\int p(\boldsymbol{x}) d\mu(\boldsymbol{x})\right)^\gamma \left(\int q(\boldsymbol{x}) d\mu(\boldsymbol{x})\right)^{1-\gamma}} \right) \\
&= \frac{1}{\gamma(\gamma-1)} \log \left(\int \bar{p}^\gamma(\boldsymbol{x})\, \bar{q}^{1-\gamma}(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right) \\
&= \frac{1}{\gamma-1} \log \left(\int \bar{q}(\boldsymbol{x}) \left(\frac{\bar{p}(\boldsymbol{x})}{\bar{q}(\boldsymbol{x})}\right)^\gamma d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma}} = D_{AR}^{(\gamma)}(\bar{\mathbf{P}}||\bar{\mathbf{Q}})
\end{aligned}
$$

for $\alpha = \gamma$ and normalized densities $\bar{p}(\boldsymbol{x}) = p(\boldsymbol{x})/\int (p(\boldsymbol{x}) d\mu(\boldsymbol{x})$ and $\bar{q}(\boldsymbol{x}) = q(\boldsymbol{x})/\int (q(\boldsymbol{x}) d\mu(\boldsymbol{x})$.

4. It can be expressed via generalized weighted mean:

$$
D_{AG}^{(\gamma)}(\mathbf{P}||\mathbf{Q}) = \frac{1}{\gamma-1} \log \left(\overline{M}_\gamma \left\{ \bar{q}; \frac{\bar{p}}{\bar{q}} \right\} \right) \tag{74}
$$

where the weighted mean is defined as $\overline{M}_\gamma \left\{ \bar{q}; \frac{\bar{p}}{\bar{q}} \right\} = \left(\int \bar{q}(\boldsymbol{x}) \left(\frac{\bar{p}(\boldsymbol{x})}{\bar{q}(\boldsymbol{x})}\right)^\gamma d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma}}$.

5. As $\gamma \to 0$, the Alpha-Gamma-divergence becomes the Kullback–Leibler divergence:

$$
\lim_{\gamma \to 0} D_{AG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = D_{KL}(\bar{\mathbf{P}}\,||\,\bar{\mathbf{Q}}) = \int \bar{p}(\boldsymbol{x}) \log \frac{\bar{p}(\boldsymbol{x})}{\bar{q}(\boldsymbol{x})} d\mu(\boldsymbol{x}) \tag{75}
$$

6. For $\gamma \to 1$, the Alpha-Gamma-divergence can be expressed by the reverse Kullback–Leibler divergence:

$$
\lim_{\gamma \to 1} D_{AG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = D_{KL}(\bar{\mathbf{Q}}\,||\,\bar{\mathbf{P}}) = \int \bar{q}(\boldsymbol{x}) \log \frac{\bar{q}(\boldsymbol{x})}{\bar{p}(\boldsymbol{x})} d\mu(\boldsymbol{x}) \tag{76}
$$

In a similar way, we can generate the whole family of Alpha-Gamma-divergences from the family of Alpha-divergences, which are summarized in Table 3.

It is interesting to note that using the above transformations (73) with $\gamma = \beta$, we can generate another family of Gamma divergences, referred to as Beta-Gamma divergences.

In particular, using the nonlinear transformations (73) for the basic asymmetric Beta-divergence (46), we obtain the Gamma-divergence (72) [35] referred to as here a Beta-Gamma-divergence ($D_G^{(\gamma)}(\mathbf{P}||\mathbf{Q}) = D_{BG}^{(\gamma)}(\mathbf{P}||\mathbf{Q})$)

$$
\begin{aligned}
D_{BG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) \ &= \frac{1}{\gamma(\gamma-1)} \left[\log(\int p^\gamma d\mu(\boldsymbol{x})) + (\gamma-1)\log(\int q^\gamma d\mu(\boldsymbol{x})) - \gamma\log(\int p\, q^{\gamma-1} d\mu(\boldsymbol{x})) \right] \\
&= \log \left(\frac{\left(\int p\, q^{\gamma-1} d\mu(\boldsymbol{x})\right)^{\frac{1}{1-\gamma}}}{\left(\int p^\gamma d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma(1-\gamma)}} \left(\int q^\gamma d\mu(\boldsymbol{x})\right)^{\frac{-1}{\gamma}}} \right) \\
&= \frac{1}{1-\gamma} \log \left(\int \widetilde{q}^\gamma(\boldsymbol{x}) \left(\frac{\widetilde{p}(\boldsymbol{x})}{\widetilde{q}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) \right)
\end{aligned}
$$

where

$$\widetilde{p}(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\left(\int p^\gamma(\boldsymbol{x})d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}}, \qquad \widetilde{q}(\boldsymbol{x}) = \frac{q(\boldsymbol{x})}{\left(\int q^\gamma(\boldsymbol{x})d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}} \tag{77}$$

Analogously, for discrete densities we can express the Beta-Gamma-divergence via generalized power means also known as the power mean or Hölder means as follows

$$D_{BG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = -\log \frac{\left(\sum_{i=1}^{n} p_i\, q_i^{\gamma-1}\right)^{\frac{1}{\gamma-1}}}{\left(\sum_{i=1}^{n} p_i^\gamma\right)^{\frac{1}{\gamma(\gamma-1)}} \left(\sum_{i=1}^{n} q_i^\gamma\right)^{\frac{1}{\gamma}}} \tag{78}$$

Hence,

$$D_{BG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = -\log \frac{\left(\sum_{i=1}^{n} \frac{p_i}{q_i}\, q_i^\gamma\right)^{\frac{1}{\gamma-1}}}{\left[\left(\sum_{i=1}^{n} p_i^\gamma\right)^{\frac{1}{\gamma}}\right]^{\frac{1}{\gamma-1}} \left(\sum_{i=1}^{n} q_i^\gamma\right)^{\frac{1}{\gamma}}} = -\log \left(\frac{\frac{1}{n}\sum_{i=1}^{n} \frac{p_i}{q_i}\, q_i^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n} p_i^\gamma\right)^{\frac{1}{\gamma}} \left[\left(\frac{1}{n}\sum_{i=1}^{n} q_i^\gamma\right)^{\frac{1}{\gamma}}\right]^{\gamma-1}}\right)^{\frac{1}{\gamma-1}}$$

and finally

$$D_{BG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = \frac{1}{1-\gamma}\log \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{p_i}{q_i}\, q_i^\gamma}{M_\gamma\{p_i\}\,[M_\gamma\{q_i\}]^{\gamma-1}} \tag{79}$$

where the (generalized) power mean of the order-γ is defined as

$$M_\gamma\{p_i\} = \left(\frac{1}{n}\sum_{i=1}^{n} p_i^\gamma\right)^{\frac{1}{\gamma}} \tag{80}$$

In the special cases, we obtain standard harmonic mean ($\gamma = -1$), geometric mean ($\gamma = 0$), arithmetic mean($\gamma = 1$), and squared root mean $\gamma = 2$ with the following relations:

$$M_{-\infty}\{p_i\} \le M_{-1}(\{p_i\}) \le M_0\{p_i\} \le M_1\{p_i\} \le M_2\{p_i\} \le M_\infty\{p_i\}, \tag{81}$$

with $M_0\{p_i\} = \lim_{\gamma\to 0} M_\gamma\{p_i\} = (\prod_{i=1}^{n} p_i)^{1/n}$, $M_{-\infty}\{p_i\} = \min_i\{p_i\}$ and $M_\infty\{p_i\} = \max_i\{p_i\}$.

Table 3. Family of Alpha-divergences and corresponding robust Alpha-Gamma-divergences; $\bar{p}(\boldsymbol{x}) = p(\boldsymbol{x})/(\int p(\boldsymbol{x})d\mu(\boldsymbol{x}))$, $\quad \bar{q}(\boldsymbol{x}) = q(\boldsymbol{x})/(\int q(\boldsymbol{x})d\mu(\boldsymbol{x}))$. For $\gamma = 0, 1$, we obtained a generalized robust KL divergences. Note that Gamma divergences are expressed compactly via generalized power means. (see also Table 4 for more direct representations).

Alpha-divergence $D_A^{(\alpha)}(\mathbf{P}||\mathbf{Q})$ Robust Alpha-Gamma-divergence $D_{AG}^{(\gamma)}(c\mathbf{P}||c\mathbf{Q})$

$$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p^\alpha q^{1-\alpha} - \alpha p + (\alpha-1)q\right) d\mu(\boldsymbol{x}), \\[2mm] \int \left(q\log\left(\dfrac{q}{p}\right) + p - q\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\[2mm] \int \left(p\log\left(\dfrac{p}{q}\right) - p + q\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\gamma-1} \log\left(\int \bar{q}\left(\dfrac{\bar{p}}{\bar{q}}\right)^\gamma d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}, \\[2mm] \int \bar{q}\log\left(\dfrac{\bar{q}}{\bar{p}}\right) d\mu(\boldsymbol{x}), \quad \gamma = 0 \\[2mm] \int \bar{p}\log\left(\dfrac{\bar{p}}{\bar{q}}\right) d\mu(\boldsymbol{x}), \quad \gamma = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p^{1-\alpha} q^{\alpha} + (\alpha-1)p - \alpha q\right) d\mu(\boldsymbol{x}), \\[2mm] \int \left(p\log\left(\dfrac{p}{q}\right) - p + q\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\[2mm] \int \left(q\log\left(\dfrac{q}{p}\right) + p - q\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\gamma-1} \log\left(\int \bar{p}\left(\dfrac{\bar{q}}{\bar{p}}\right)^\gamma d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}, \\[2mm] \int \bar{p}\log\left(\dfrac{\bar{p}}{\bar{q}}\right) d\mu(\boldsymbol{x}), \quad \gamma = 0 \\[2mm] \int \bar{q}\log\left(\dfrac{\bar{q}}{\bar{p}}\right) d\mu(\boldsymbol{x}), \quad \gamma = 1 \end{cases}$$

$$\begin{cases} \dfrac{\int \left(\left(\dfrac{p+q}{2}\right)^\alpha q^{1-\alpha} - \dfrac{\alpha}{2}p - \left(1-\dfrac{\alpha}{2}\right)q\right) d\mu(\boldsymbol{x})}{\alpha(\alpha-1)}, \\[2mm] \int \left(q\log\left(\dfrac{2q}{p+q}\right) + \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\[2mm] \int \left(\dfrac{p+q}{2}\log\left(\dfrac{p+q}{2q}\right) - \dfrac{p-q}{2}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\gamma-1} \log\left[\left(\dfrac{\int q\,d\mu(\boldsymbol{x})}{\int p\,d\mu(\boldsymbol{x})}\right)^{\frac{1}{2}} \left(\int \bar{q}\left(\dfrac{p+q}{2q}\right)^\gamma d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}\right] \\[2mm] \int \left(\bar{q}\log\left(\dfrac{2q}{p+q}\right)\right) d\mu(\boldsymbol{x}), \quad \gamma = 0 \\[2mm] \int \left(\dfrac{\bar{p}+\bar{q}}{2}\log\left(\dfrac{p+q}{2q}\right)\right) d\mu(\boldsymbol{x}), \quad \gamma = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\alpha-1} \int \left[\left(\dfrac{p+q}{2q}\right)^{\alpha-1} - 1\right](p-q)\,d\mu(\boldsymbol{x}), \\[2mm] \int (p-q)\log\left(\dfrac{p+q}{2q}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$$

$$\begin{cases} \log\left(\dfrac{\int (\bar{p}\left(\dfrac{p+q}{2q}\right)^{\gamma-1})d\mu(\boldsymbol{x})}{\int (\bar{q}\left(\dfrac{p+q}{2q}\right)^{\gamma-1})d\mu(\boldsymbol{x})}\right)^{\frac{1}{\gamma-1}} \\[2mm] \int (\bar{p}-\bar{q})\log\left(\dfrac{p+q}{2q}\right) d\mu(\boldsymbol{x}), \quad \gamma = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\alpha(\alpha-1)} \int \left(p^\alpha q^{1-\alpha} + p^{1-\alpha} q^\alpha - p - q\right) d\mu(\boldsymbol{x}), \\[2mm] \int (p-q)\log\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad \alpha = 0, 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\gamma-1} \log\left[\int \bar{q}\left(\dfrac{p}{q}\right)^\gamma d\mu(\boldsymbol{x}) \int \bar{p}\left(\dfrac{q}{p}\right)^\gamma d\mu(\boldsymbol{x})\right]^{\frac{1}{\gamma}} \\[2mm] \int (\bar{p}-\bar{q})\log\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad \gamma = 0, 1 \end{cases}$$

$$\begin{cases} \dfrac{\int \left((p^{1-\alpha} + q^{1-\alpha})\left(\dfrac{p+q}{2}\right)^\alpha - p - q\right) d\mu(\boldsymbol{x})}{\alpha(\alpha-1)}, \\[2mm] \int \left(p\log\left(\dfrac{2p}{p+q}\right) + q\log\left(\dfrac{2q}{p+q}\right)\right) d\mu(\boldsymbol{x}), \quad \alpha = 0 \\[2mm] \int (p+q)\log\left(\dfrac{p+q}{2\sqrt{pq}}\right) d\mu(\boldsymbol{x}), \quad \alpha = 1 \end{cases}$$

$$\begin{cases} \dfrac{1}{\gamma-1} \log\left[\int \bar{p}\left(\dfrac{p+q}{2p}\right)^\gamma d\mu(\boldsymbol{x}) \int \bar{q}\left(\dfrac{p+q}{2q}\right)^\gamma d\mu(\boldsymbol{x})\right]^{\frac{1}{\gamma}} \\[2mm] \int \left(\bar{p}\log\left(\dfrac{2p}{p+q}\right) + \bar{q}\log\left(\dfrac{2q}{p+q}\right)\right) d\mu(\boldsymbol{x}), \quad \gamma = 0 \\[2mm] \int (\bar{p}+\bar{q})\log\left(\dfrac{p+q}{2\sqrt{pq}}\right) d\mu(\boldsymbol{x}), \quad \gamma = 1. \end{cases}$$

Table 4. Basic Alpha- and Beta-divergences and directly generated corresponding Gamma-divergences (see also Table 3 how the Gamma-divergences can be expressed by power means).

Divergence $D_A^{(\alpha)}(\mathbf{P}\|\mathbf{Q})$ or $D_B^{(\beta)}(\mathbf{P}\|\mathbf{Q})$	Gamma-divergence $D_{AG}^{(\gamma)}(c\mathbf{P}\|c\mathbf{Q})$ and $D_{BG}^{(\gamma)}(c\mathbf{P}\|c\mathbf{Q})$
$\dfrac{1}{\alpha(1-\alpha)}\int\left(\alpha p+(1-\alpha)q-p^\alpha q^{1-\alpha}\right)d\mu(\boldsymbol{x})$	$-\log\dfrac{\left(\int p^\gamma q^{1-\gamma}d\mu(\boldsymbol{x})\right)^{1/(\gamma(1-\gamma))}}{\left(\int pd\mu(\boldsymbol{x})\right)^{1/(1-\gamma)}\left(\int qd\mu(\boldsymbol{x})\right)^{1/\gamma}}$
$\dfrac{1}{\beta(\beta-1)}\int\left(p^\beta+(\beta-1)q^\beta-\beta p q^{\beta-1}\right)d\mu(\boldsymbol{x})$	$-\log\dfrac{\left(\int p q^{\gamma-1}d\mu(\boldsymbol{x})\right)^{1/(\gamma-1)}}{\left(\int p^\gamma d\mu(\boldsymbol{x})\right)^{1/(\gamma(\gamma-1))}\left(\int q^\gamma d\mu(\boldsymbol{x})\right)^{1/\gamma}}$
$\dfrac{1}{\alpha(1-\alpha)}\int\left(p+q-p^\alpha q^{1-\alpha}-p^{1-\alpha}q^\alpha\right)d\mu(\boldsymbol{x})$	$\dfrac{-1}{\gamma(1-\gamma)}\log\dfrac{\left(\int p^\gamma q^{1-\gamma}d\mu(\boldsymbol{x})\right)\left(\int p^{1-\gamma}q^\gamma d\mu(\boldsymbol{x})\right)}{\int pd\mu(\boldsymbol{x})\int qd\mu(\boldsymbol{x})}$
$\dfrac{1}{\beta-1}\int\left(p^\beta+q^\beta-p q^{\beta-1}-p^{\beta-1}q\right)d\mu(\boldsymbol{x})$	$\dfrac{-1}{\gamma-1}\log\dfrac{\left(\int p q^{\gamma-1}d\mu(\boldsymbol{x})\right)\left(\int p^{\gamma-1}q d\mu(\boldsymbol{x})\right)}{\left(\int p^\gamma d\mu(\boldsymbol{x})\right)\left(\int q^\gamma d\mu(\boldsymbol{x})\right)}$
$\dfrac{\int\left(\frac{\alpha}{2}p+(1-\frac{\alpha}{2})q-\left(\frac{p+q}{2}\right)^\alpha q^{\alpha-1}\right)d\mu(\boldsymbol{x})}{\alpha(1-\alpha)}$	$\dfrac{-1}{\gamma(1-\gamma)}\log\dfrac{\int\left(\frac{p+q}{2}\right)^{\gamma-1}q^{\gamma-1}d\mu(\boldsymbol{x})}{\left(\int pd\mu(\boldsymbol{x})\right)^{\gamma/2}\left(\int qd\mu(\boldsymbol{x})\right)^{1-\gamma/2}}$
$\dfrac{1}{1-\alpha}\int(p-q)\left(1-\frac{p+q}{2q}\right)^{\alpha-1}d\mu(\boldsymbol{x})$	$\dfrac{1}{1-\gamma}\log\dfrac{\int pd\mu(\boldsymbol{x})\int q\left(\frac{p+q}{2q}\right)^{\gamma-1}d\mu(\boldsymbol{x})}{\int qd\mu(\boldsymbol{x})\int p\left(\frac{p+q}{2q}\right)^{\gamma-1}d\mu(\boldsymbol{x})}$
$\dfrac{\int\left(p+q-(p^{1-\alpha}+q^{1-\alpha})\left(\frac{p+q}{2}\right)^\alpha\right)d\mu(\boldsymbol{x})}{\alpha(1-\alpha)}$	$\log\left(\dfrac{\int p^{1-\gamma}\left(\frac{p+q}{2}\right)^\gamma d\mu(\boldsymbol{x})\int q^{1-\gamma}\left(\frac{p+q}{2}\right)^\gamma d\mu(\boldsymbol{x})}{\int pd\mu(\boldsymbol{x})\int qd\mu(\boldsymbol{x})}\right)^{\frac{1}{\gamma(\gamma-1)}}$

The asymmetric Beta-Gamma-divergence has the following properties [30,35]:

1. $D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) \geq 0$. The equality holds if and only if $\mathbf{P} = c\mathbf{Q}$ for a positive constant c.

2. It is scale invariant, that is, $D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{BG}^{(\gamma)}(c_1\mathbf{P} \,||\, c_2\mathbf{Q})$, for arbitrary positive scaling constants c_1, c_2.

3. As $\gamma \to 1$, the Gamma-divergence becomes the Kullback–Leibler divergence:

$$\lim_{\gamma \to 1} D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{KL}(\bar{\mathbf{P}} \,||\, \bar{\mathbf{Q}}) = \int \bar{p} \, \log \frac{\bar{p}}{\bar{q}} d\mu(\boldsymbol{x}) \tag{82}$$

where $\bar{p} = p / \int p\,d\mu(\boldsymbol{x})$ and $\bar{q} = q / \int q\,d\mu(\boldsymbol{x})$.

4. For $\gamma \to 0$, the Gamma-divergence can be expressed as follows

$$\lim_{\gamma \to 0} D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = \int \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \, d\mu(\boldsymbol{x}) + \log \left(\int \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \, d\mu(\boldsymbol{x}) \right) \tag{83}$$

For the discrete Gamma divergence we have the corresponding formula

$$D_{BG}^{(0)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{q_i}{p_i} \right) + \log \left(\sum_{i=1}^{n} \frac{p_i}{q_i} \right) - \log(n) = \log \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{q_i}}{\left(\prod_{i=1}^{n} \frac{p_i}{q_i} \right)^{1/n}} \tag{84}$$

Similarly to the Alpha and Beta-divergences, we can also define the symmetric Beta-Gamma-divergence as

$$D_{BGS}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) + D_{BG}^{(\gamma)}(\mathbf{Q} \,||\, \mathbf{P}) = \frac{1}{1-\gamma} \, \log \left[\frac{\left(\int p^{\gamma-1} q \, d\mu(\boldsymbol{x}) \right) \left(\int p \, q^{\gamma-1} d\mu(\boldsymbol{x}) \right)}{\left(\int p^\gamma d\mu(\boldsymbol{x}) \right) \left(\int q^\gamma d\mu(\boldsymbol{x}) \right)} \right]$$

The symmetric Gamma-divergence has similar properties to the asymmetric Gamma-divergence:

1. $D_{BGS}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) \geq 0$. The equality holds if and only if $\mathbf{P} = c\mathbf{Q}$ for a positive constant c, in particular, $p = q, \ \forall i$.

2. It is scale invariant, that is,

$$D_{BGS}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{BGS}^{(\gamma)}(c_1\mathbf{P} \,||\, c_2\mathbf{Q}) \tag{85}$$

for arbitrary positive scaling constants c_1, c_2.

3. For $\gamma \to 1$, it is reduced to a special form of the symmetric Kullback–Leibler divergence (also called the J-divergence)

$$\lim_{\gamma \to 1} D_{BGS}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = \int (\bar{p} - \bar{q}) \log \frac{\bar{p}}{\bar{q}} d\mu(\boldsymbol{x}) \tag{86}$$

where $\bar{p} = p / \int p\,d\mu(\boldsymbol{x})$ and $\bar{q} = q / \int q\,d\mu(\boldsymbol{x})$.

4. For $\gamma = 0$, we obtain a simple divergence expressed by weighted arithmetic means

$$D_{BGS}^{(0)}(\mathbf{P} \,||\, \mathbf{Q}) = \log \left(\int w \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \, d\mu(\boldsymbol{x}) \int w \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \, d\mu(\boldsymbol{x}) \right) \tag{87}$$

where weight function $w > 0$ is such that $\int w \, d\mu(\boldsymbol{x}) = 1$.

For the discrete Beta-Gamma divergence (or simply the Gamma divergence), we obtain divergence

$$
\begin{aligned}
D_{BGS}^{(0)}(\mathbf{P} \,||\, \mathbf{Q}) = \quad & \log \left(\sum_{i=1}^{n} \frac{p_i}{q_i} \sum_{i=1}^{n} \frac{q_i}{p_i} \right) - \log(n)^2 \\
= \quad & \log \left(\left(\frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{q_i} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{q_i}{p_i} \right) \right) \\
= \quad & \log \left(M_1 \left\{ \frac{p_i}{q_i} \right\} M_1 \left\{ \frac{q_i}{p_i} \right\} \right)
\end{aligned}
$$

It is interesting to note that for $n \rightarrow \infty$ the discrete symmetric Gamma-divergence can be expressed by expectation functions $D_{BGS}^{(1)}(\mathbf{P} \,||\, \mathbf{Q}) = \log(E\{\boldsymbol{u}\} \, E\{\boldsymbol{u}^{-1}\})$, where $u_i = \{p_i/q_i\}$ and $u_i^{-1} = \{q_i/p_i\}$.

5. For $\gamma = 2$, the asymmetric Gamma-divergences (equal to a symmetric Gamma-divergence) is reduced to Cauchy–Schwarz divergence, introduced by Principe [83]

$$
\begin{aligned}
D_{BG}^{(2)}(\mathbf{P} \,||\, \mathbf{Q}) \; &= D_{BG}^{(2)}(\mathbf{Q} \,||\, \mathbf{P}) = \frac{1}{2} D_{BGS}^{(2)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{2} D_{CS}(\mathbf{P} \,||\, \mathbf{Q}) \\
&= -\log \frac{\displaystyle\int p(\boldsymbol{x}) \, q(\boldsymbol{x}) \, d\mu(\boldsymbol{x})}{\left(\displaystyle\int p^2(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{1/2} \left(\displaystyle\int q^2(\boldsymbol{x}) d\mu(\boldsymbol{x}) \right)^{1/2}}
\end{aligned}
$$

It should be noted that the basic asymmetric Beta-Gamma divergence (derived from the Beta-divergence) is exactly equivalent to the Gamma divergence defined in [35], while Alpha-Gamma divergences (derived from the family of Alpha-divergences) have different expressions but they are similar in terms of properties.

5. Relationships for Asymmetric Divergences and their Unified Representation

The fundamental relationships and analogies for the generalized divergences discussed in this paper are summarized in Table 5.

The basic difference among Alpha- Beta and Gamma divergences is their scaling properties

$$D_A^{(\alpha)}(c \, \mathbf{P} \,||\, c \, \mathbf{Q}) = c \, D_A^{(\alpha)}(\mathbf{P} \,||\, \mathbf{Q}) \tag{88}$$

$$D_B^{(\beta)}(c \, \mathbf{P} \,||\, c \, \mathbf{Q}) = c^{\beta} \, D_B^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) \tag{89}$$

Table 5. Fundamental asymmetric divergences and their relationships. We can obtain the Beta-divergence from the Alpha-divergence via transformations $p \to p^\beta$, $q \to q^\beta$ and $\alpha = \beta^{-1}$ and the Alpha-divergence from the Beta-divergence via dual transformations: $p \to p^\alpha$, $q \to q^\alpha$ and $\beta = \alpha^{-1}$. Furthermore, we can generate the Gamma-divergences from the Alpha- and Beta- divergences via transformations $c_0 \int p^{c_1}(\boldsymbol{x}) q^{c_2}(\boldsymbol{x}) d\mu(\boldsymbol{x}) \to c_0 \log \int p^{c_1}(\boldsymbol{x}) q^{c_2}(\boldsymbol{x}) d\mu(\boldsymbol{x})$. Moreover, we can generate Beta-Gamma divergences from the Alpha-Gamma divergence via transformations: $p \to p^{\gamma_{new}}$, $q \to q^{\gamma_{new}}$ and $\gamma = \gamma_{new}^{-1}$.

Divergence name	Formula
Alpha	$D_A^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) = \dfrac{\int (p^\alpha q^{1-\alpha} - \alpha\, p + (\alpha - 1)\, q) d\mu(\boldsymbol{x})}{\alpha(\alpha - 1)}$
Beta	$D_B^{(\beta)}(\mathbf{P}\|\mathbf{Q}) = \dfrac{\int (p^\beta + (\beta - 1)\, q^\beta - \beta\, p\, q^{\beta-1}) d\mu(\boldsymbol{x})}{\beta(\beta - 1)}$
Gamma	$D_{AG}^{(\gamma)}(\mathbf{P}\|\mathbf{Q}) = \dfrac{\log\left(\int p^\gamma q^{1-\gamma} d\mu(\boldsymbol{x})\right) - \gamma \log\left(\int p\, d\mu(\boldsymbol{x})\right) + (\gamma - 1) \log\left(\int q\, d\mu(\boldsymbol{x})\right)}{\gamma(\gamma - 1)}$
	$\qquad\qquad = \dfrac{1}{\gamma - 1} \log\left(\int \bar{q}\left(\dfrac{\bar{p}}{\bar{q}}\right)^\gamma d\mu(\boldsymbol{x})\right)^{\frac{1}{\gamma}}$
	$D_{BG}^{(\gamma)}(\mathbf{P}\|\mathbf{Q}) = \dfrac{\log\left(\int p^\gamma d\mu(\boldsymbol{x})\right) + (\gamma - 1)\log\left(\int q^\gamma d\mu(\boldsymbol{x})\right) - \gamma\log\left(\int p\, q^{\gamma-1} d\mu(\boldsymbol{x})\right)}{\gamma(\gamma - 1)}$
	$\qquad\qquad = \dfrac{1}{1 - \gamma} \log\left(\int \widetilde{q}^{\,\gamma}\left(\dfrac{\widetilde{p}}{\widetilde{q}}\right) d\mu(\boldsymbol{x})\right) \quad))$
Alpha-Rényi	$D_{AR}^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) = \dfrac{\log(\int (p^\alpha q^{1-\alpha} - \alpha p + (\alpha - 1)\, q) d\mu(\boldsymbol{x}) + 1)}{\alpha(\alpha - 1)}$
Bregman	$D_\Phi(\mathbf{P}\|\mathbf{Q}) = \int \left(\Phi(p) - \Phi(q) - \dfrac{\delta\Phi}{\delta q}(p - q)\right) d\mu(\boldsymbol{x}), \quad$ (see Eqation(57))
Csiszár-Morimoto	$D_f(\mathbf{P}\|\mathbf{Q}) = \int q f\left(\dfrac{p}{q}\right) d\mu(\boldsymbol{x}), \quad$ (see Equation (57)) for $f(u) = \Phi(t)_{|t=u}$)

$$D_G^{(\gamma)}(c\,\mathbf{P}\,||\,c\,\mathbf{Q}) = D_G^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) \tag{90}$$

for any $c > 0$.

Note that the values of all Alpha-divergences are proportional to a scaling factor and this scaling is independent of the parameter α, while for all Beta-divergences this scaling factor is dependent on the value of the parameter β exponentially. Only in a special case for $\beta = 0$ (Itakura–Saito distance), the Beta-divergences are invariant to a scaling factor, while the Gamma-divergences is invariant to a scaling factor for any value of the parameter γ.

On the basis of our analysis and discussion in the previous sections, we come to conclusion that the wide class of asymmetric divergences can be described by the following generalized function:

$$D_{AC}^{(\alpha,\beta,r)}(\mathbf{P}||\mathbf{Q}) = \frac{1}{\alpha(\beta-1)\lambda\,r}\left[\left(\widetilde{D}_{AC}^{(\alpha,\beta)}+1\right)^r - 1\right] \tag{91}$$

where $r > 0$ and

$$\widetilde{D}_{AC}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \int\left(\alpha p^\lambda + (\beta-1)q^\lambda - \lambda p^\alpha q^{\beta-1}\right)d\mu(\boldsymbol{x}) \tag{92}$$

with $\alpha \neq 0$, $\beta \neq 1$ and $\lambda = \alpha + \beta - 1 \neq 0$.

Note that such defined function has similar structure to generalized logarithm (3). Moreover, it should be noted that the above function is a divergence only for such set of parameters for which $\widetilde{D}_{AC}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \geq 0$ and equals zero if and only if $\mathbf{P} = \mathbf{Q}$.

In the special cases, we obtain the following divergences:

- For $r = \lambda = 1$ and $\alpha + \beta = 2$, we obtain the Alpha-divergence (4)

- For $r = 1$ and $\alpha = 1$ $(\beta = \lambda)$, we obtain the Beta-divergence (46)

- For $r \to 0$ and $\alpha + \beta = 2$ $(\lambda = 1)$, we obtain the Alpha-Rényi divergence (25), which for normalized probabilities densities: $\bar{p} = p/(\int p d\mu(\boldsymbol{x}))$ and $\bar{q} = q/(\int q d\mu(\boldsymbol{x}))$ reduces to the Alpha-Gamma divergence.

- And finally, for $r \to 0$ and $\alpha = 1$ $(\beta = \lambda = \gamma)$ and the normalized densities: $\widetilde{p} = p/(\int p^\gamma d\mu(\boldsymbol{x}))^{1/\gamma}$ and $\widetilde{q} = q/(\int q^\gamma d\mu(\boldsymbol{x}))^{1/\gamma}$, we obtain the Gamma divergence (referred in this paper as the Beta-Gamma divergence) (72), proposed recently by Fujisawa and Eguchi [35].

The function (91) has many interesting general properties. However, this is out of the scope of this paper.

Duality

Note that we can easily establish the following dualities for the Alpha and Alpha-Gamma divergences:

$$D_A^{(\alpha)}(\mathbf{P}\,||\,\mathbf{Q}) = c^{-1}D_A^{(1-\alpha)}(c\,\mathbf{Q}\,||\,c\,\mathbf{P}), \qquad D_{AG}^{(\gamma)}(\mathbf{P}\,||\,\mathbf{Q}) = D_{AG}^{(1-\gamma)}(c_1\mathbf{Q}\,||\,c_2\mathbf{P}) \tag{93}$$

However, in order to establish duality for the Beta-divergence, we need represent it in slightly more general form as two set parameters divergence:

$$D_{AB}^{(\beta_1, \beta_2)}(\mathbf{P} \,||\, \mathbf{Q}) = \frac{1}{\beta_1 \beta_2 (\beta_1 + \beta_2)} \int \left(\beta_1 \, p^{\beta_1 + \beta_2} + \beta_2 \, q^{\beta_1 + \beta_2} - (\beta_1 + \beta_2) p^{\beta_1} q^{\beta_2} \right) d\mu(\boldsymbol{x}) \qquad (94)$$

In this case, we can express a duality of the Beta-divergence as

$$D_B^{(\beta)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{AB}^{(1, \beta - 1)}(\mathbf{P} \,||\, \mathbf{Q}) = c^{-\beta} D_{AB}^{(\beta - 1, 1)}(c \, \mathbf{Q} \,||\, c \, \mathbf{P}) \qquad (95)$$

or more generally

$$D_{AB}^{(\beta_1, \beta_2)}(\mathbf{P} \,||\, \mathbf{Q}) = c^{-(\beta_1 + \beta_2)} D_{AB}^{(\beta_2, \beta_1)}(c \, \mathbf{Q} \,||\, c \, \mathbf{P}) \qquad (96)$$

Analogously, we define family of the asymmetric Gamma divergences in more general form as a two-set parameter divergence:

$$D_{ABG}^{(\gamma_1, \gamma_2)}(\mathbf{P} \| \mathbf{Q}) = \log \frac{\left(\int p^{\gamma_1 + \gamma_2} d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma_2 (\gamma_1 + \gamma_2)}} \left(\int q^{\gamma_1 + \gamma_2} d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma_1 (\gamma_1 + \gamma_2)}}}{\left(\int p^{\gamma_1} q^{\gamma_2} d\mu(\boldsymbol{x}) \right)^{\frac{1}{\gamma_1 \gamma_2}}} \qquad (97)$$

Hence, we express duality property for the Beta-Gamma (basic Gamma) -divergence as follows

$$D_{BG}^{(\gamma)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{ABG}^{(1, \gamma - 1)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{ABG}^{(\gamma - 1, 1)}(c_1 \, \mathbf{Q} \,||\, c_2 \, \mathbf{P}) \qquad (98)$$

or more generally

$$D_{ABG}^{(\gamma_1, \gamma_2)}(\mathbf{P} \,||\, \mathbf{Q}) = D_{ABG}^{(\gamma_2, \gamma_1)}(c_1 \mathbf{Q} \,||\, c_2 \mathbf{P}) \qquad (99)$$

5.1. *Conclusions and Discussion*

The main goal of this paper was to establish and widen bridges between several classes of divergences. The main results are summarized in Tables 2–5. We have extended and unified several important classes of divergence and provided links and correspondences among them. The parametric families of the generalized divergences presented in this paper, especially the Alpha-, Alpha-Rényi and Beta-divergences, enable us to consider smooth connections of various pairs of well-known and frequently used fundamental divergences under one "umbrella". For example, the family of Beta-divergences smoothly connects the squared Euclidean distance (L_2-norm) with a generalized extended version of Itakura–Saito like distances and through the the generalized KL I-divergences. In fact, we have shown how to generate the whole family of the extended robust Itakura–Saito like distances from Beta-divergences for $\beta = 0$, which are scale invariant and can be used in many potential applications, especially in speech processing.

Furthermore, we have shown that two different families of Gamma-divergences can be generated one from Beta-divergences and the other one from Alpha-divergences. The Gamma-divergences characterized by the global scale invariance (independent of value of parameter γ) and high robustness. We have reviewed and extended fundamental properties of the Gamma-divergences. Moreover, our

approach allows us to generate a wide class of new divergences, especially Gamma-divergences and family of extended Itakura–Saito like distances. Special emphasis is given to divergences for singular values of tuning parameters $0, 1$.

These families have many desirable properties such as flexibility, robustness to outliers and they involve only one single tuning parameter (α, β, or γ). An insight using information geometry may further elucidates the fundamental structures of such divergences and geometry [1,3,7].

In comparison to previous closely related works, we explicitly investigated three wide class of divergences (Alpha, Beta- Gama), discussed their properties and links to Csiszár–Morimoto and Bregman divergences. Some provided properties, especially for Gamma-divergences are new.

The properties of divergences discussed in this paper allowed us to develop a wide family of robust algorithms for Independent Component Analysis (ICA) and Nonnegative Matrix Factorization (NMF) presented in separated recent works [30–34,68–71]. They have great potential in many applications, especially in blind separation of statistically dependent sparse and/or smooth sources and in feature extractions and classification problems [30].

One of the important and still open problem (not discussed in this paper) is how to set tuning parameters α, β and γ depending on distributions of available data sets and noise or outliers. Some recent works address this problem [69,71,84]. Usually, there is some trade off between robustness and efficiency [35,37,71]. For example, if the tuning parameter γ is large, then the robustness of the Gamma divergence will be strong but the efficiency could be lower, and vice versa, if the tuning parameter has small positive value, then the robustness of the method would be not strong but the efficiency will be higher [35]. The relation between the efficiency and the tuning parameters was discussed by several authors (see Jones *et al.* [85], Basu *et al.* [67], and Fujisawa and Eguchi [35]).

References

1. Amari, S. *Differential-Geometrical Methods in Statistics*; Springer Verlag: Berlin, Germany, 1985.
2. Amari, S. Dualistic geometry of the manifold of higher-order neurons. *Neural Network.* **1991**, *4*, 443–451.
3. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: New York, NY, USA, 2000.
4. Amari, S. Integration of stochastic models by minimizing alpha-divergence. *Neural Comput.* **2007**, *19*, 2780–2796.
5. Amari, S. Information geometry and its applications: Convex function and dually flat manifold. In *Emerging Trends in Visual Computing*; Nielsen, F., Ed.; Springer: New York, NY, USA; pp. 75–102.
6. Amari, S. Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
7. Amari, S.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci.* **2010**; (in print).
8. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.

9. Fujimoto, Y.; Murata, N. A modified EM Algorithm for mixture models based on Bregman divergence. *Ann. Inst. Stat. Math.* **2007**, *59*, 57–75.

10. Zhu, H.; Rohwer, R. Bayesian Invariant measurements of generalization. *Neural Process. Lett.* **1995**, *2*, 28–31.

11. Zhu, H.; Rohwer, R. Measurements of generalisation based on information geometry. In *Mathematics of Neural Networks: Model Algorithms and Applications*; Ellacott, S.W., Mason, J.C., Anderson, I.J., Eds.; Kluwer: Norwell, MA, USA, 1997; pp. 394–398.

12. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *56*, 2882–2903.

13. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi diagrams. *Discrete and Computational Geometry (Springer)* **2010**; (in print).

14. Yamano, T. A generalization of the Kullback-Leibler divergence and its properties. *J. Math. Phys.* **2009**, *50*, 85–95.

15. Minami, M.; Eguchi, S. Robust blind source separation by Beta-divergence. *Neural Comput.* **2002**, *14*, 1859–1886.

16. Bregman, L. The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Comp. Math. Phys., USSR* **1967**, *7*, 200–217.

17. Csiszár, I. Eine Informations Theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitt von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutat Int. Kzl* **1963**, *8*, 85–108.

18. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273.

19. Csiszár, I. Information measures: A critial survey. In Transactions of the 7th Prague Conference, Prague, Czech Republic, 18–23 August 1974; Reidel: Dordrecht, Netherlands, 1977; pp. 83–86.

20. Ali, M.; Silvey, S. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc.* **1966**, *Ser B*, 131–142.

21. Hein, M.; Bousquet, O. Hilbertian metrics and positive definite kernels on probability measures. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Barbados, 6–8 January 2005; Ghahramani, Z., Cowell, R., Eds.; *AISTATS* **2005**, *10*, 136–143.

22. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.

23. Zhang, J. Referential duality and representational duality on statistical manifolds. In Proceedings of the Second International Symposium on Information Geometry and its Applications, University of Tokyo, Tokyo, Japan, 12–16 December 2005; University of Tokyo: Tokyo, Japan, 2006; pp. 58–67.

24. Zhang, J. A note on curvature of a-connections of a statistical manifold. *Ann. Inst. Stat. Math.* **2007**, *59*, 161–170.

25. Zhang, J.; Matsuzoe, H. Dualistic differential geometry associated with a convex function. In *Springer Series of Advances in Mechanics and Mathematics*, 2008; Springer: New York, NY, USA; pp. 58–67.

26. Lafferty, J. Additive models, boosting, and inference for generalized divergences. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 7–9 July 1999; ACM: New York, NY, USA, 1999.

27. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.

28. Villmann, T.; Haase, S. Divergence based vector quantization using Fréchet derivatives. *Neural Comput.* **2010**, (submitted for publication).

29. Villmann, T.; Haase, S.; Schleif, F.M.; Hammer, B. Divergence based online learning in vector quantization. In Proceedings of the International Conference on Artifial Intelligence and Soft Computing (ICAISC2010), LNAI, Zakopane, Poland, 13–17 June 2010.

30. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S. *Nonnegative Matrix and Tensor Factorizations*; John Wiley & Sons Ltd: Chichester, UK, 2009.

31. Cichocki, A.; Zdunek, R.; Amari, S. Csiszár's divergences for nonnegative matrix factorization: Family of new algorithms. *Springer, LNCS-3889* **2006**, *3889*, 32–39.

32. Cichocki, A.; Amari, S.; Zdunek, R.; Kompass, R.; Hori, G.; He, Z. Extended SMART algorithms for Nonnegative Matrix Factorization. *Springer, LNAI-4029* **2006**, *4029*, 548–562.

33. Cichocki, A.; Zdunek, R.; Choi, S.; Plemmons, R.; Amari, S. Nonnegative tensor factorization using Alpha and Beta divergences. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulose, France, May 2007; Volume III, pp. 1393–1396.

34. Cichocki, A.; Zdunek, R.; Choi, S.; Plemmons, R.; Amari, S.I. Novel multi-layer nonnegative tensor factorization with sparsity constraints. *Springer, LNCS-4432* **2007**, *4432*, 271–280.

35. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.

36. Liese, F.; Vajda, I. Convex Statistical Distances. *Teubner-Texte zur Mathematik Teubner Texts in Mathematics* **1987**, *95*, 1–85.

37. Eguchi, S.; Kato, S. Entropy and divergence associated with power function and the statistical application. *Entropy* **2010**, *12*, 262–274.

38. Taneja, I. On generalized entropies with applications. In *Lectures in Applied Mathematics and Informatics*; Ricciardi, L., Ed.; Manchester University Press: Manchester, UK, 1990; pp. 107–169.

39. Taneja, I. New developments in generalized information measures. In *Advances in Imaging and Electron Physics*; Hawkes, P., Ed.; Elsevier: Amsterdam, Netherlands, 1995; Volume 91, pp. 37–135.

40. Gorban, A.N.; Gorban, P.A.; Judge, G. Entropy: The Markov ordering approach. *Entropy* **2010**, *12*, 1145–1193.

41. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Ann. Math. Statist.* **1952**, *23*, 493–507.

42. Minka, T. Divergence measures and message passing. *Microsoft Research Technical Report (MSR-TR-2005)*, 2005.

43. Taneja, I. On measures of information and inaccuarcy. *J. Statist. Phys.* **1976**, *14*, 203–270.

44. Cressie, N.; Read, T. *Goodness-of-Fit Statistics for Discrete Multivariate Data*; Springer: New York, NY, USA, 1988.

45. Cichocki, A.; Lee, H.; Kim, Y.D.; Choi, S. Nonnegative matrix factorization with Alpha-divergence. *Pattern. Recognit. Lett.* **2008**, *29*, 1433–1440.

46. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Phys.* **1988**, *52*, 479–487.

47. Havrda, J.; Charvát, F. Quantification method of classification processes: Concept of structrual a-entropy. *Kybernetika* **1967**, *3*, 30–35.

48. Cressie, N.; Read, T. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. Ser. B* **1984**, *46*, 440–464.

49. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer Academic Press: Amsterdam, Netherland, 1989.

50. Hellinger, E. Neue Begründung der Theorie Quadratischen Formen von unendlichen vielen Veränderlichen. *J. Reine Ang. Math.* **1909**, *136*, 210–271.

51. Morimoto, T. Markov processes and the *H*-theorem. *J. Phys. Soc. Jap.* **1963**, *12*, 328–331.

52. Österreicher, F. Csiszár's f-divergences-basic properties. Technical report, In *Research Report Collection*; Victoria University: Melbourne, Australia, 2002.

53. Harremoës, P.; Vajda, I. Joint range of *f*-divergences. Accepted for presentation at ISIT 2010, Austin, TX, USA, 13–18 June 2010.

54. Dragomir, S. *Inequalities for Csiszár f-Divergence in Information Theory*; Victoria University: Melbourne, Australia, 2000; (edited monograph).

55. Rényi, A. On the foundation of information theory. *Rev. Inst. Int. Stat.* **1965**, *33*, 1–4.

56. Rényi, A. On measures of entropy and information. In *Proceddings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA; Volome 1, pp. 547–561.

57. Rényi, A. *Probability Theory*; North-Holland: Amsterdam, The Netherlands, 1970.

58. Harremoës, P. Interpretaions of Rényi entropies and divergences. *Physica A* **2006**, *365*, 57–62.

59. Harremoës, P. Joint range of Rényi entropies. *Kybernetika* **2009**, *45*, 901–911.

60. Hero, A.; Ma, B.; Michel, O.; Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Process. Mag.* **2002**, *19*, 85–95.

61. Topsoe, F. Some inequalities for information divergence and related measuresof discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609.

62. Burbea, J.; Rao, C. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multi. Analysis* **1982**, *12*, 575–596.

63. Burbea, J.; Rao, C. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory* **1982**, *IT-28*, 489–495.

64. Sibson, R. Information radius. *Probability Theory and Related Fields* **1969**, *14*, 149–160.

65. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. Lon., Ser. A* **1946**, *186*, 453–461.

66. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86.

67. Basu, A.; Harris, I.R.; Hjort, N.; Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.

68. Mollah, M.; Minami, M.; Eguchi, S. Exploring latent structure of mixture ICA models by the minimum Beta-divergence method. *Neural Comput.* **2006**, *16*, 166–190.

69. Mollah, M.; Eguchi, S.; Minami, M. Robust prewhitening for ICA by minimizing Beta-divergence and its application to FastICA. *Neural Process. Lett.* **2007**, *25*, 91–110.

70. Kompass, R. A Generalized divergence measure for Nonnegative Matrix Factorization. *Neural Comput.* **2006**, *19*, 780–791.

71. Mollah, M.; Sultana, N.; Minami, M.; Eguchi, S. Robust extraction of local structures by the minimum of Beta-divergence method. *Neural Netw.* **2010**, *23*, 226–238.

72. Nielsen, F.; Nock, R. The dual Voronoi diagrams with respect to representational Bregman divergences. In *Proceedings of the International Symposium on Voronoi Diagrams (ISVD)*, Copenhagen, Denmark, 23–26 June 2009.

73. Cichocki, A.; Phan, A. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE (invited paper)* **2009**, *E92-A (3)*, 708–721.

74. Cichocki, A.; Phan, A.; Caiafa, C. Flexible HALS algorithms for sparse non-negative matrix/tensor factorization. In Proceedings of the 18th IEEE workshops on Machine Learning for Signal Processing, Cancun, Mexico, 16–19 October 2008.

75. Dhillon, I.; Sra, S. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In *Neural Information Processing Systems*; MIT Press: Vancouver, Canada, 2005; pp. 283–290.

76. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.

77. Itakura, F.; Saito, F. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the of the 6th International Congress on Acoustics, Tokyo, Japan, 1968; pp. 17–20.

78. Eggermont, P.; LaRiccia, V. On EM-like algorithms for minimum distance estimation. Technical report, Mathematical Sciences; University of Delaware: Newark, DE, USA, 1998.

79. Févotte, C.; Cemgil, A.T. Nonnegative matrix factorizations as probabilistic inference in composite models. In Proceedings of the 17th European Signal Processing Conference (EUSIPCO-09), Glasgow, Scotland, UK, 24–28 August 2009.

80. Banerjee, A.; Dhillon, I.; Ghosh, J.; Merugu, S.; Modha, D. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; ACM Press: New York, NY, USA, 2004; pp. 509–514.

81. Lafferty, J. Additive models, boosting, and inference for generalized divergences. In Proceedings of the 12th Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 7–9 July 1999; ACM Press: New York, USA, 1999, pp. 125–133.

82. B.A.Frigyik.; Srivastan, S.; Gupta, M. Functional Bregman divergence and Bayesion estimation of distributions. *IEEE Trans. Inf. Theory* **2008**, *54*, 5130–5139.

83. Principe, J. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: Berlin, Germany, 2010.

84. Choi, H.; Choi, S.; Katake, A.; Choe, Y. Learning alpha-integration with partially-labeled data. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2010), Dallas, TX, USA, 14–19 March 2010.

85. Jones, M.; Hjort, N.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **1998**, *85*, 865–873.

Appendix A. Divergences Derived from Tsallis α-entropy

In this appendix, we presents links between the Tsallis and the Rényi entropies and Alpha- Beta-divergences and Gamma-divergences. For simplicity of our considerations, we assume first that $\int \bar{p}(\boldsymbol{x})d\mu(\boldsymbol{x}) = 1$ and $\int \bar{q}(\boldsymbol{x})d\mu(\boldsymbol{x}) = 1$, but these constraints can be relaxed.

The Tsallis entropy is defined by using the power function $h_\alpha(\bar{\boldsymbol{p}}) = \int \bar{p}^\alpha d\mu(\boldsymbol{x})$,

$$H_T^{(\alpha)}(\bar{\boldsymbol{p}}) = \frac{1}{1-\alpha}\left\{h_\alpha(\bar{\boldsymbol{p}}) - 1\right\} \tag{100}$$

The Rényi entropy is also written as

$$H_R^{(\alpha)}(\bar{\boldsymbol{p}}) = \frac{1}{1-\alpha}\log\left\{h_\alpha(\bar{\boldsymbol{p}})\right\} = \frac{1}{1-\alpha}\log\left(\int \bar{p}^\alpha(\boldsymbol{x})\,d\mu(\boldsymbol{x})\right) \tag{101}$$

It is shown that the Tsallis entropy is related (corresponds) to the Alpha- and Beta-divergences, while the Rényi entropy is related to the Gamma-divergences.

The Rényi and Tsallis entropy measures are two different generalization of the standard Boltzman-Gibbs entropy (or Shannon's information). In the limit as $\alpha \to 1$ we have $p^{\alpha-1} = \exp((\alpha-1)\log p) \approx 1 + (\alpha-1)\log p$, hence they reduce to the Shannon entropy:

$$H_S(\bar{\boldsymbol{p}}) = -\int \bar{p}(\boldsymbol{x})\log(\bar{p}(\boldsymbol{x}))\,d\mu(\boldsymbol{x}) \tag{102}$$

The Tsallis entropy can also be considered as an α-deformation of the Shannon's entropy by noting that

$$H_T^{(\alpha)}(\bar{\boldsymbol{p}}) = -\int \bar{p}^\alpha(\boldsymbol{x})\log_\alpha(\bar{p}(\boldsymbol{x}))\,d\mu(\boldsymbol{x}) = \frac{1}{1-\alpha}\left(\int \bar{p}^\alpha(\boldsymbol{x})d\mu(\boldsymbol{x}) - 1\right) \tag{103}$$

and by using the generalized logarithm function defined by

$$\log_\alpha x = \frac{1}{1-\alpha}\left(x^{1-\alpha} - 1\right) \tag{104}$$

The Tsallis relative entropy or the Tsallis divergence is defined similarly by

$$D_T^{(\alpha)}(\bar{\boldsymbol{p}}\|\bar{\boldsymbol{q}}) = -\int \bar{p}\log_\alpha\left(\frac{\bar{q}}{\bar{p}}\right)d\mu(\boldsymbol{x}) = \frac{1}{1-\alpha}\left(1 - \int \bar{p}^\alpha \bar{q}^{1-\alpha}d\mu(\boldsymbol{x})\right) \tag{105}$$

This is a rescaled version of the Alpha-divergence,

$$D_T^{(\alpha)}(\bar{\boldsymbol{p}}\|\bar{\boldsymbol{q}}) = \alpha D_A^{(\alpha)}(\bar{\boldsymbol{p}}\|\bar{\boldsymbol{q}}) \tag{106}$$

It should be noted that the Alpha-divergence is more general since it is defined for any positive arrays $\mathbf{P}$ and $\mathbf{Q}$ and also for any values of α including singular values of $\alpha = 0$ and $\alpha = 1$.

Since the Tsallis Alpha-divergence is an f-divergence, it is possible to extend it to the applicable general cases. This is given in Appendix B, where $\log_\alpha(x)$ is replaced by another power function.

Another way of deriving a divergence is to use the convex generating functions. This has been deeply investigated by Zhang [22].

Since

$$\Phi_\alpha(\bar{\boldsymbol{p}}) = -h_\alpha(\bar{\boldsymbol{p}}) \tag{107}$$

is a convex function, we derive a Bregman divergence associated with the Tsallis entropy,

$$\widetilde{D}_T^{(\alpha)}(\bar{\boldsymbol{p}} \,\|\, \bar{\boldsymbol{q}}) = \int \left\{ -\bar{p}^\alpha + \bar{q}^\alpha + \alpha \bar{p}^{\alpha-1}(\bar{p} - \bar{q}) \right\} d\mu(\boldsymbol{x}) \tag{108}$$

This is a rescaled and simplified version of the Beta-divergence with $\alpha = \beta$,

$$\widetilde{D}_T^{(\alpha)}(\bar{\boldsymbol{p}} \,\|\, \bar{\boldsymbol{q}}) = \alpha(1 - \alpha) D_B^{(\beta)}(\bar{\boldsymbol{p}} \| \bar{\boldsymbol{q}}) \tag{109}$$

Again, note that the Beta-divergence is more general than the Tsallis divergence, since it is defined for any positive arrays and any β including singular values of $\beta = 0$ and $\beta = 1$. We need to extend the definition of $h_\alpha(\boldsymbol{p})$ to obtain the whole range of the Beta-divergences.

Appendix B. Entropies and Divergences

An entropy $H(\bar{\boldsymbol{p}})$ is maximized at the uniform distribution. Therefore, when a divergence function is given, we can define an entropy $H(\bar{\boldsymbol{p}})$ by

$$H(\bar{\boldsymbol{p}}) = D(\bar{\boldsymbol{p}} \,\|\, \mathbf{1}) + c \tag{110}$$

for some constant c and normalized probability measure $\bar{p}$.

In many cases, $\Psi(\bar{\boldsymbol{p}}) = -D(\bar{\boldsymbol{p}} \,\|\, \mathbf{1})$ is a convex function of $\bar{p}$. For example, from the KL-divergence, we have

$$H_{KL}(\bar{\boldsymbol{p}}) = -\int \bar{p}(\boldsymbol{x}) \log \bar{p}(\boldsymbol{x}) d\mu(\boldsymbol{x}) \tag{111}$$

which is the Shannon entropy (neglecting a constant).

We define a power function of probabilities by

$$h_\alpha(\bar{\boldsymbol{p}}) = \int \bar{p}^\alpha(\boldsymbol{x}) d\mu(\boldsymbol{x}) \tag{112}$$

where $0 < \alpha < 1$.

From the Alpha-divergence $D_A^{(\alpha)}$ and Beta-divergence $D_B^{(\beta)}$ we have the entropies

$$H(\bar{\boldsymbol{p}}) = c\, h_\alpha(\bar{\boldsymbol{p}}) + d \tag{113}$$

where c and d are constants and $\alpha = \beta$.

Similarly, from the Gamma-divergence $D_G^{(\gamma)}$, we have

$$H(\bar{\boldsymbol{p}}) = c \log(h_\alpha(\bar{\boldsymbol{p}})) + d \tag{114}$$

with $\alpha = \gamma$.

This implies that various divergence functions can generate a similar family of entropy functions. On the other hand, there are a number of ways to generate divergences from entropy functions. A typical way is to use the Bregman divergence derived from a convex generating function $\Phi(\bar{p}) = -H(\bar{p})$,

$$D_\Phi(\bar{\boldsymbol{p}} \,\|\, \bar{\boldsymbol{q}}) = \int \left(\boldsymbol{\Phi}(\bar{p}) - \boldsymbol{\Phi}(\bar{q}) - \frac{\delta\Phi(\bar{q})}{\delta\bar{q}}(\bar{p} - \bar{q}) \right) d\mu(\boldsymbol{x}) \tag{115}$$

Appendix C. Tsallis and Rényi Entropies

It is worth mentioning that the Rényi and Tsallis α-entropies are monotonically related through

$$H_R^{(\alpha)}(\bar{\boldsymbol{p}}) = \frac{1}{1-\alpha} \log\left(1 + (1-\alpha)H_T^{(\alpha)}(\bar{\boldsymbol{p}})\right) \tag{116}$$

or using the α-logarithm

$$H_T^{(\alpha)}(\bar{\boldsymbol{p}}) = \log_\alpha e^{H_R^{(\alpha)}(\bar{\boldsymbol{p}})} \tag{117}$$

$\Psi_R^{(\alpha)} = -H_R^{(\alpha)}$ is convex function of $\bar{\boldsymbol{p}}$, and hence we have a corresponding divergence.

The Bregman divergence can be defined in a more general form as shown below.

If $\Phi(\bar{\boldsymbol{p}})$: $\mathbb{R}^n \to \mathbb{R}$ is a strictly convex and C^1 (*i.e.*, first-order differentiable) function, the corresponding Bregman distance (divergence) for discrete positive measures is defined by

$$D_\Phi(\bar{\boldsymbol{p}}\,||\,\bar{\boldsymbol{q}}) = \Phi(\bar{\boldsymbol{p}}) - \Phi(\bar{\boldsymbol{q}}) - (\bar{\boldsymbol{p}} - \bar{\boldsymbol{q}})^T \nabla\Phi(\bar{\boldsymbol{q}}) \tag{118}$$

where $\nabla\Phi(\bar{\boldsymbol{q}})$ represents the gradient of Φ evaluated at $\bar{\boldsymbol{q}}$.

The Bregman divergence derived from the Rényi entropy for discrete probability densities is given by

$$D_R^{(\alpha)}(\bar{\boldsymbol{p}}\,||\,\bar{\boldsymbol{q}}) = \frac{1}{1-\alpha} \log \frac{h_\alpha(\bar{\boldsymbol{q}})}{h_\alpha(\boldsymbol{p})} - \frac{\alpha}{1-\alpha} \frac{1}{h_\alpha(\bar{\boldsymbol{q}})}\left(1 - \sum_i \bar{p}_i\,\bar{q}_i^{\alpha-1}\right) \tag{119}$$

The Tsallis entropy is closely related to the α-exponential family of probability distributions. A family $S = \{\bar{p}(\boldsymbol{x}, \boldsymbol{\theta})\}$, parameterized by a vector parameter $\boldsymbol{\theta}$, of probability distributions is called a α-exponential family, when it is written as

$$\log_\alpha \bar{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_i \theta_i x_i - \Psi(\boldsymbol{\theta}) \tag{120}$$

where $\Psi(\boldsymbol{\theta})$ is a convex function. This is the same as the α-family defined in [6]. The set $S = \{\boldsymbol{p}\}$ of discrete probability distributions is a α-exponential family for any α. To show this, we introduce random variables $\boldsymbol{x} = (x)$,

$$x_i = \delta_i \tag{121}$$

which is 1 when the outcome is i of which probability is $\bar{p}$.

Then, the probability is written as

$$\log_\alpha \bar{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_i \theta_i \delta_i - \Psi(\boldsymbol{\theta}) \tag{122}$$

where

$$\theta_i = \log_\alpha \bar{p}_i \tag{123}$$

Since $\sum \bar{p}_i = 1$ holds, or

$$\sum \delta_i = 1 \tag{124}$$

and hence free variables of $\boldsymbol{\theta}$ are $n - 1$, say $(\theta_1, \cdots, \theta_{n-1})$. In this case θ_n is a function of the other variables, and we get

$$\Psi(\boldsymbol{\theta}) = -\log_\alpha \theta_n \tag{125}$$

We can derive a divergence from this convex function $\Psi(\boldsymbol{\theta})$. However, it is more intuitive to use the dual of $\Psi(\boldsymbol{\theta})$, which is given by

$$\varphi(\boldsymbol{\eta}) = \frac{1}{1 - \alpha} \left(\frac{1}{h_\alpha(\bar{\boldsymbol{p}})} - 1 \right) \tag{126}$$

where $\boldsymbol{\eta}$ is the dual variable of $\boldsymbol{\theta}$ by the Legendre transformation

$$\boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}) \tag{127}$$

The derived divergence is

$$D_{\exp}^{(\alpha)}(\bar{\boldsymbol{p}} \,\|\, \bar{\boldsymbol{q}}) = \frac{1}{h_\alpha(\bar{\boldsymbol{p}})} D_T^{(\alpha)}(\bar{\boldsymbol{p}} \,\|\, \bar{\boldsymbol{q}}) \tag{128}$$

This is a conformal transformation of the Alpha-divergence [7]. A conformal transformation, also called a conformal map or angle-preserving transformation is a transformation that preserves local angles of intersection of any two lines or curves.