entropy

MDPI

Article

A Generalized Relative (α, β)-Entropy: Geometric Properties and Applications to Robust Statistical Inference

Abhik Ghosh and Ayanendranath Basu *

Indian Statistical Institute, Kolkata 700108, India; abhianik@gmail.com
* Correspondence: ayanbasu@isical.ac.in; Tel.: +91-33-2575-2806

Received: 30 March 2018; Accepted: 1 May 2018; Published: 6 May 2018

Abstract: Entropy and relative entropy measures play a crucial role in mathematical information theory. The relative entropies are also widely used in statistics under the name of divergence measures which link these two fields of science through the minimum divergence principle. Divergence measures are popular among statisticians as many of the corresponding minimum divergence methods lead to robust inference in the presence of outliers in the observed data; examples include the ϕ-divergence, the density power divergence, the logarithmic density power divergence and the recently developed family of logarithmic super divergence (LSD). In this paper, we will present an alternative information theoretic formulation of the LSD measures as a two-parameter generalization of the relative α-entropy, which we refer to as the general (α, β)-entropy. We explore its relation with various other entropies and divergences, which also generates a two-parameter extension of Renyi entropy measure as a by-product. This paper is primarily focused on the geometric properties of the relative (α, β)-entropy or the LSD measures; we prove their continuity and convexity in both the arguments along with an extended Pythagorean relation under a power-transformation of the domain space. We also derive a set of sufficient conditions under which the forward and the reverse projections of the relative (α, β)-entropy exist and are unique. Finally, we briefly discuss the potential applications of the relative (α, β)-entropy or the LSD measures in statistical inference, in particular, for robust parameter estimation and hypothesis testing. Our results on the reverse projection of the relative (α, β)-entropy establish, for the first time, the existence and uniqueness of the minimum LSD estimators. Numerical illustrations are also provided for the problem of estimating the binomial parameter.

Keywords: relative entropy; logarithmic super divergence; robustness; minimum divergence inference; generalized renyi entropy

1. Introduction

Decision making under uncertainty is the backbone of modern information science. The works of C. E. Shannon and the development of his famous entropy measure [1–3] represent the early mathematical foundations of information theory. The Shannon entropy and the corresponding relative entropy, commonly known as the Kullback-Leibler divergence (KLD), has helped to link information theory simultaneously with probability [4–8] and statistics [9–13]. If P and Q are two probability measures on a measurable space $(\Omega, \mathcal{A})$ and have absolutely continuous densities p and q, respectively, with respect to a common dominating σ-finite measure μ, then the Shannon entropy of P is defined as

$$\mathcal{E}(P) = -\int p \log(p) d\mu, \tag{1}$$

and the KLD measure between P and Q is given by

$$\mathcal{RE}(P,Q) = \int p \log\left(\frac{p}{q}\right) d\mu. \tag{2}$$

In statistics, the minimization of the KLD measure produces the most likely approximation as given by the maximum likelihood principle; the latter, in turn, has a direct equivalence to the (Shannon) entropy maximization criterion in information theory. For example, if Ω is finite and μ is the counting measure, it is easy to see that $\mathcal{RE}(P,U) = \log|\Omega| - \mathcal{E}(P)$, where U is the uniform measure on Ω. Minimization of this relative entropy, or equivalently maximization of the Shannon entropy, with respect to P within a suitable convex set $\mathbb{E}$, generates the most probable distribution for an independent identically distributed finite source having true marginal probability in $\mathbb{E}$ with non-informative (uniform) prior probability of guessing [14,15]. In general, with a finite source, $\mathcal{RE}(P,Q)$ denotes the penalty in expected compressed length if the compressor assumes a mismatched probability Q [16,17]. The corresponding general minimizer of $\mathcal{RE}(P,Q)$ given Q, namely its forward projection, and other geometric properties of $\mathcal{RE}(P,Q)$ are well studied in the literature; see [18–29] among others.

Although the maximum entropy or the minimum divergence criterion based on the classical Shannon entropy $\mathcal{E}(P)$ and the KLD measure $\mathcal{RE}(P,Q)$ is still widely used in major (probabilistic) decision making problems in information science and statistics [30–43], there also exist many different useful generalizations of these quantities to address eminent issues in quantum statistical physics, complex codings, statistical robustness and many other topics of interest. For example, if we consider the standardized cumulant of compression length in place of the expected compression length in Shannon's theory, the optimum distribution turns out to be the maximizer of a generalization of the Shannon entropy [44,45] which is given by

$$\mathcal{E}_\alpha(P) = \frac{1}{1-\alpha}\log\left(\int p^\alpha d\mu\right), \quad \alpha > 0,\ \alpha \neq 1 \tag{3}$$

provided $p \in L_\alpha(\mu)$, the complete vector space of functions for which the α-th power of their absolute values are μ-integrable. This general entropy functional is popular by the name Renyi entropy of order α [46] and covers many important entropy measures like Hartley entropy at $\alpha \to 0$ (for finite source), Shannon entropy at $\alpha \to 1$, collision entropy at $\alpha = 2$ and the min-entropy at $\alpha \to \infty$. The corresponding Renyi divergence measure is given by

$$\mathcal{D}_\alpha(P,Q) = \frac{1}{\alpha-1}\log\left(\int p^\alpha q^{1-\alpha} d\mu\right), \ \alpha > 0,\ \alpha \neq 1, \tag{4}$$

whenever $p, q \in L_\alpha(\mu)$ and coincides with the classical KLD measure at $\alpha \to 1$. The Renyi entropy and the Renyi divergence are widely used in recent complex physical and statistical problems; see, for example, [47–56]. Other non-logarithmic extensions of Shannon entropy include the classical f-entropies [57], the Tsallis entropy [58] as well as the more recent generalized (α,β,γ)-entropy [59,60] among many others; the corresponding divergences and the minimum divergence criteria are widely used in critical information theoretic and statistical problems; see [57,59–70] for details.

We have noted that there is a direct information theoretic connection of KLD to the Shannon entropy under mismatched guessing by minimizing the expected compressed length. However, such a connection does not exist between the Renyi entropy $\mathcal{E}_\alpha(P)$ and the Renyi divergence $\mathcal{D}_\alpha(P,Q)$ as recently noted by [17,71]. Herein, it has been shown that, for a finite source with marginal distribution P and a (prior) mismatched compressor distribution Q, the penalty in the normalized cumulant of compression length is not $\mathcal{D}_\alpha(P,Q)$; rather it is given by $\mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha)$ where P_α and Q_α are defined by

$$\frac{dP_\alpha}{d\mu} = p_\alpha = \frac{p^\alpha}{\int p^\alpha d\mu}, \quad \frac{dQ_\alpha}{d\mu} = q_\alpha = \frac{q^\alpha}{\int q^\alpha d\mu}. \tag{5}$$

The new quantity $\mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha)$ also gives a measure of discrimination (i.e., is a divergence) between the probability distributions P and Q and coincides with the KLD at $\alpha \to 1$. This functional is referred to as the relative α-entropy in the terminology of [72] and has the simpler form

$$\begin{aligned}\mathcal{RE}_\alpha(P,Q) &:= \mathcal{D}_{1/\alpha}(P_\alpha, Q_\alpha)\\ &= \frac{\alpha}{1-\alpha}\log\int pq^{\alpha-1}d\mu - \frac{1}{1-\alpha}\log\int p^\alpha d\mu + \log\int q^\alpha d\mu, \quad \alpha > 0,\ \alpha \neq 1.\end{aligned} \tag{6}$$

The geometric properties of this relative α-entropy along with its forward and reverse projections have been studied recently [16,73]; see Section 2.1 for some details. This quantity had, however, already been proposed earlier as a statistical divergence, although for $\alpha \geq 1$ only, by [74] while developing a robust estimation procedure following the generalized method-of-moments approach of [75]. Later authors referred to the divergence proposed in [74] as the logarithmic density power divergence (LDPD) measure. The advantages of the minimum LDPD estimator in terms of robustness against outliers in data have been studied by, among other, [66,74]. Fujisawa [76], Fujisawa and Eguchi [77] have also used the same divergence measure with $\gamma = (\alpha - 1) \geq 0$ in different statistical problems and have referred to it as the γ-divergence. Note that, the formulation in (6) extends the definition of the divergence over the $0 < \alpha < 1$ region as well.

Motivated by the substantial advantages of the minimum LDPD inference in terms of statistical robustness against outlying observations, Maji et al. [78,79] have recently developed a two-parameter generalization of the LDPD family, namely the logarithmic super divergence (LSD) family, given by

$$\begin{aligned}\mathcal{LSD}_{\tau,\gamma}(P,Q) = \frac{1}{B}\log\int p^{1+\tau}d\mu - \frac{1+\tau}{AB}\log\int p^A q^B d\mu + \frac{1}{A}\log\int q^{1+\tau}d\mu,\\ \text{with} \quad A = 1+\gamma(1-\tau), B = 1+\tau-A, \ \ \tau \geq 0, \gamma \in \mathbb{R}.\end{aligned} \tag{7}$$

This rich superfamily of divergences contain many important divergence measures including the LDPD at $\gamma = 0$ and the Kullback-Leibler divergence at $\tau = \gamma \to 0$; this family also contains a transformation of Renyi divergence at $\tau = 0$ which has been referred to as the logarithmic power-divergence family by [80]. As shown in [78,79], the statistical inference based on some of the new members of this LSD family, outside the existing ones including the LDPD, provide much better trade-off between the robustness and efficiency of the corresponding minimum divergence estimators.

The statistical benefits of the LSD family over the LDPD family raise a natural question: is it possible to translate this robustness advantage of the LSD family of divergences to the information theoretic context, through the development of a corresponding generalization of the relative α-entropy in (6)? In this paper, we partly answer this question by defining an independent information theoretic generalization of the relative α-entropy measure coinciding with the LSD measure. We will refer to this new generalized relative entropy measure as the "*Relative* (α, β)*-entropy*" and study its properties for different values of $\alpha > 0$ and $\beta \in \mathbb{R}$. In particular, this new formulation will extend the scope of the LSD measure for $-1 < \tau < 0$ as well and generate several interesting new divergence and entropy measures. We also study the geometric properties of all members of the relative (α, β)-entropy family, or equivalently the LSD measures, including their continuity in both the arguments and a Pythagorean-type relation. The related forward projection problem, i.e., the minimization of the relative (α, β)-entropy in its first argument, is also studied extensively.

In summary, the main objective of the present paper is to study the geometric properties of the LSD measure through the new information theoretic or entropic formulation (or the relative (α, β)-entropy). Our results indeed generalize the properties of the relative α-entropy from [16,73]. The specific and significant contributions of the paper can be summarized as follows.

1. We present a two parameter extension of the relative α-entropy measure in (6) motivated by the logarithmic S-divergence measures. These divergence measures are known to generate more robust statistical inference compared to the LDPD measures related to the relative α-entropy.

2. In the new formulation of the relative (α, β)-entropy, the LSD measures are linked with several important information theoretic divergences and entropy measures like the ones named after Renyi. A new divergence family is discovered corresponding to $\alpha \rightarrow 0$ case (properly standardized) for the finite measure cases.
3. As a by-product of our new formulation, we get a new two-parameter generalization of the Renyi entropy measure, which we refer to as the Generalized Renyi entropy (GRE). This opens up a new area of research to examine the detailed properties of GRE and its use in complex problems in statistical physics and information theory. In this paper, we show that this new GRE satisfies the basic entropic characteristics, i.e., it is zero when the argument probability is degenerate and is maximum when the probability is uniform.
4. Here we provide a detailed geometric analysis of the robust LSD measure, or equivalently the relative (α, β)-entropy in our new formulation. In particular, we show their continuity or lower semi-continuity with respect to the first argument depending on the values of the tuning parameters α and β. Also, its lower semi-continuity with respect to the second argument is proved.
5. We also study the convexity of the LSD measures (or the relative (α, β)-entropies) with respect to its argument densities. The relative α-entropy (i.e, the relative (α, β)-entropy at $\beta = 1$) is known to be quasi-convex [16] only in its first argument. Here, we will show that, for general $\alpha > 0$ and $\beta \neq 1$, the relative (α, β)-entropies are not quasi-convex on the space of densities, but they are always quasi-convex with respect to both the arguments on a suitably (power) transformed space of densities. Such convexity results in the second argument were unavailable in the literature even for the relative α-entropy, which we will introduce in this paper through a transformation of space.
6. Like the relative α-entropy, but unlike the relative entropy in (2), our new relative (α, β)-entropy also does not satisfy the data processing inequalities. However, we prove an extended Pythagorean relation for the relative (α, β)-entropy which makes it reasonable to treat them as "squared distances" and talk about their projections.
7. The forward projection of a relative entropy or a suitable divergence, i.e., their minimization with respect to the first argument, is very important for both statistical physics and information theory. This is indeed equivalent to the maximum entropy principle and is also related to the Gibbs conditioning principle. In this paper, we will examine the conditions under which such a forward projection of the relative (α, β)-entropy (or, LSD) exists and is unique.
8. Finally, for completeness, we briefly present the application of the LSD measure or the relative (α, β)-entropy measure in robust statistical inference in the spirit of [78,79] but now with extended range of tuning parameters. It uses the reverse projection principle; a result on the existence of the minimum LSD functional is first presented with the new formulation of this paper. Numerical illustrations are provided for the binomial model, where we additionally study their properties for the extended tuning parameter range $\alpha \in (0, 1)$ as well as for some new divergence families (related to $\alpha = 0$). Brief indications of the potential use of these divergences in testing of statistical hypotheses are also provided.

Although we are primarily discussing the logarithmic entropies like the Renyi entropy and its generalizations in this paper, it is important to point out that non-logarithmic entropies including the f-entropy and the Tsallis entropy are also very useful in several applications with real systems. Recently, several complex physical and social systems have been observed to follow the theory developed from such non-logarithmic, non-additive entropies instead of the classical additive Shannon entropy. In particular, the Tsallis entropy has led to the development of the nonextensive statistical mechanics [61,64] to solve several critical issues in modern physics. Important areas of application include, but certainly are not limited to, the motion of cold atoms in dissipative optical lattices [81,82], the magnetic field fluctuations in the solar wind and related q-triplet [83], the distribution of velocity in driven dissipative dusty plasma [84], spin glass relaxation [85], the interaction of trapped ion with a classical buffer gas [86], different high energy collisional experiments [87–89], derivation of the black hole entropy [90], along with water engineering [63], text mining [65] and many others. Therefore, it is also important to investigate the possible generalizations and manipulations of such non-logarithmic entropies both from mathematical and application point of view. However, as our primary interest here is in logarithmic entropies, we have, to keep the focus clear, otherwise avoided the description and development of non-logarithmic entropies in this paper.

Although there are many applications of extended and general non-additive entropy and divergence measures, there are also some criticisms of these non-additive measures that should be kept in mind. It is of course possible to employ such quantities simply as new descriptors of the complexity of systems, but at the same time, it is known that the minimization of a generalized divergence (or maximization of the corresponding entropy) under constraints in order to determine an optimal probability assignment leads to inconsistencies for information measures other than the Kullback-Leibler divergence. See, for instance [91–96], among others. So, one needs to be very careful in discriminating the application of the newly introduced entropies and divergence measures for the purposes of inference under given information, from the ones where it is used as a measure of complexity. In this respect, we would like to emphasize that, the main advantage of our two-parameter extended family of LSD or relative (α, β)-entropy measures in parametric statistical inference is in their strong robustness property against possible contamination (generally manifested through outliers) in the sample data. The classical additive Shannon entropy and Kullback-Leibler divergence produce non-robust inference even under a small proportion of data contamination, but the extremely high robustness of the LSD has been investigated in detail, with both theoretical and empirical justifications, by [78,79]; in this respect, we will present some numerical illustrations in Section 5.2. Another important issue could be to decide whether to stop at the two-parameter level for information measures or to extend it to three-parameters, four-parameters, etc. It is not an easy question to answer. However, we have seen that many members of the two-parameter family of LSD measures generate highly robust inference along with a desirable trade-off between efficiency under pure data and robustness under contaminated data. Therefore a two-parameter system appears to work well in practice. Since it is a known principle that one "should not multiply entities beyond necessity", we will, for the sake of parsimony, restrict ourselves to the second level of generalization for robust statistical inference, at least until there is further convincing evidence that the next higher level of generalization can produce a significant improvement.

2. The Relative (α, β)-Entropy Measure

2.1. Definition: An Extension of the Relative α-Entropy

In order to motivate the development of our generalized relative (α, β)-entropy measure, let us first briefly describe an alternative formulation of the relative α-entropy following [16]. Consider the mathematical set-up of Section 1 with $\alpha > 0$ and assume that the space $L_\alpha(\mu)$ is equipped with the norm

$$||f||_\alpha = \begin{cases} \left(\int |f|^\alpha d\mu\right)^{1/\alpha} & \text{if} \quad \alpha \geq 1, \ f \in L_\alpha(\mu), \\ \int |f|^\alpha d\mu & \text{if} \quad 0 < \alpha < 1, \ f \in L_\alpha(\mu), \end{cases} \tag{8}$$

and the corresponding metric $d_\alpha(g, f) = ||g - f||_\alpha$ for $g, f \in L_\alpha(\mu)$. Then, the relative α-entropy between two distributions P and Q is obtained as a function of the Cressie-Read power divergence measure [97], defined below in (11), between the escort measures P_α and Q_α defined in (5). Note that the disparity family or the ϕ-divergence family [18,98–103] between P and Q is defined as

$$D_\phi(P, Q) = \int q\phi\left(\frac{p}{q}\right) d\mu, \tag{9}$$

for a continuous convex function ϕ on $[0, \infty)$ satisfying $\phi(0) = 0$ and with the usual convention $0\phi(0/0) = 0$. We consider the ϕ-function given by

$$\phi(u) = \phi_\lambda(u) = sign(\lambda(\lambda+1))\left(u^{\lambda+1} - 1\right), \quad \lambda \in \mathbb{R}, u \geq 0, \tag{10}$$

with the convention that, for any $u > 0$, $0\phi_\lambda(u/0) = 0$ if $\lambda < 0$ and $0\phi_\lambda(u/0) = \infty$ if $\lambda > 0$. The corresponding ϕ-divergence has the form

$$D_\lambda(P, Q) = D_{\phi_\lambda}(P, Q) = sign(\lambda(\lambda+1)) \int q \left[\left(\frac{p}{q}\right)^{\lambda+1} - 1\right] d\mu, \tag{11}$$

which is just a positive multiple of the Cressie-Read power divergence with the multiplicative constant being $|\lambda(1+\lambda)|$; when this constant is present, the case $\lambda = 0$ leads to the KLD measure in a limiting sense. Note that, our ϕ-function in (10) differs slightly from the one used by [16] in that we use $sign(\lambda(\lambda+1))$ in place of $sign(\lambda)$ there; this is to make the divergence in (11) non-negative for all $\lambda \in \mathbb{R}$ ([16] considered only $\lambda > -1$) which will be needed to define our generalized relative entropy. Then, given an $\alpha > 0$, [16,17] set $\lambda = \alpha^{-1} - 1 (> -1)$ and show that the relative α-entropy of P with respect to Q can be obtained as

$$\mathcal{RE}_\alpha(P, Q) = \mathcal{RE}^\mu_\alpha(P, Q) = \frac{1}{\lambda}\log\left[sign(\lambda)D_\lambda(P_\alpha, Q_\alpha) + 1\right]. \tag{12}$$

It is straightforward to see that the above formulation (12) coincides with the definition given in (6). We often suppress the superscript μ whenever the underlying measure is clear from the context; in most applications in information theory and statistics it is either counting measure or the Lebesgue measure depending on whether the distribution is discrete or continuous.

We can now change the tuning parameters in the formulation given by (12) suitably as to arrive at the more general form of the LSD family in (7). For this purpose, let us fix $\alpha > 0$, $\beta \in \mathbb{R}$ and assume that $p, q \in L_\alpha(\mu)$ are the μ-densities of P and Q, respectively. Instead of considering the re-parametrization $\lambda = \alpha^{-1} - 1$ as above, we now consider the two-parameter re-parametrization $\lambda = \beta\alpha^{-1} - 1 \in \mathbb{R}$. Note that, the feasible range of λ, in order to make $\alpha > 0$, now clearly depends on β through $\alpha = \frac{\beta}{1+\lambda} > 0$; whenever $\beta > 0$ we have $-1 < \lambda < \infty$ and if $\beta < 0$ we need $-\infty < \lambda < -1$. We have already taken care of this dependence through the modified ϕ function defined in (10) which ensures that $D_\lambda(\cdot, \cdot)$ is non-negative for all $\lambda \in \mathbb{R}$. So we can again use the relation as in (12), after suitable standardization due to the additional parameter β, to define a new generalized relative entropy measure as given in the following definition.

Definition 1 (Relative (α, β)-entropy)**.** *Given any $\alpha > 0$ and $\beta \in \mathbb{R}$, put $\lambda = \frac{\beta}{\alpha} - 1$ (i.e., $\alpha = \frac{\beta}{1+\lambda}$). Then, the relative (α, β)-entropy of P with respect to Q is defined as*

$$\mathcal{RE}_{\alpha,\beta}(P, Q) = \mathcal{RE}^\mu_{\alpha,\beta}(P, Q) = \frac{1}{\beta\lambda}\log\left[sign(\beta\lambda)D_\lambda(P_\alpha, Q_\alpha) + 1\right]. \tag{13}$$

The cases $\beta = 0$ and $\lambda = 0$ (i.e, $\beta = \alpha$) are defined in limiting sense; see Equations (15) and (16) below.

A straightforward simplification gives a simpler form of this new relative (α, β)-entropy which coincides with the LSD measure as follows.

$$\begin{aligned}\mathcal{RE}_{\alpha,\beta}(P, Q) &= \frac{1}{\alpha-\beta}\log\int p^\alpha d\mu - \frac{\alpha}{\beta(\alpha-\beta)}\log\int p^\beta q^{\alpha-\beta} d\mu + \frac{1}{\beta}\log\int q^\alpha d\mu, \\ &= \mathcal{LSD}_{\alpha-1, \frac{\beta-1}{2-\alpha}}(P, Q).\end{aligned} \tag{14}$$

Note that, it coincides with the relative α-entropy $\mathcal{RE}_\alpha(P,Q)$ at the choice $\beta = 1$. For the limiting cases, it leads to the forms

$$\mathcal{RE}_{\alpha,0}(P,Q) = \frac{\int \log(q/p) q^\alpha d\mu}{\int q^\alpha d\mu} + \frac{1}{\alpha}\log\left(\frac{\int p^\alpha d\mu}{\int q^\alpha d\mu}\right), \tag{15}$$

$$\mathcal{RE}_{\alpha,\alpha}(P,Q) = \frac{\int \log(p/q) p^\alpha d\mu}{\int p^\alpha d\mu} + \frac{1}{\alpha}\log\left(\frac{\int q^\alpha d\mu}{\int p^\alpha d\mu}\right). \tag{16}$$

By the divergence property of $D_\lambda(\cdot,\cdot)$, all the relative (α,β)-entropies are non-negative and valid statistical divergences. Note that, in view of (14), the formulation (13) extends the scope of LSD measure, defined in (7), for $\tau \in (-1,0)$.

Proposition 1. *For any $\alpha > 0$ and $\beta \in \mathbb{R}$, $\mathcal{RE}_{\alpha,\beta}(P,Q) \geq 0$ for all probability measures P and Q, whenever it is defined. Further, $\mathcal{RE}_{\alpha,\beta}(P,Q) = 0$ if and only in $P = Q[\mu]$.*

Also, it is important to identify the cases where the relative (α,β)-entropy is not finitely defined, which can be obtained from the definition and convention related to D_λ divergence; these are summarized in the following proposition.

Proposition 2. *For any $\alpha > 0$, $\beta \in \mathbb{R}$ and distributions P, Q having μ-densities in $L_\alpha(\mu)$, the relative (α,β)-entropy $\mathcal{RE}_{\alpha,\beta}(P,Q)$ is a finite positive number except for the following three cases:*

1. *P is not absolutely continuous with respect to Q and $\alpha < \beta$, in which case $\mathcal{RE}_{\alpha,\beta}(P,Q) = +\infty$.*
2. *P is mutually singular to Q and $\alpha > \beta$, in which case also $\mathcal{RE}_{\alpha,\beta}(P,Q) = +\infty$.*
3. *$0 < \beta < \alpha$ and $D_\lambda(P_\alpha, Q_\alpha) \geq 1$, in which case also $\mathcal{RE}_{\alpha,\beta}(P,Q)$ is undefined.*

The above two propositions completely characterize the values and existence of our new relative (α,β)-entropy measure. In the next subsection, we will now explore its relation with other existing entropies and divergence measures; along the way we will get some new ones as by-products of our generalized relative entropy formulation.

2.2. Relations with Different Existing or New Entropies and Divergences

The relative (α,β)-entropy measures form a large family containing several existing relative entropies and divergences. Its relation with some popular ones are summarized in the following proposition; the proof is straightforward from definitions and hence omitted.

Proposition 3. *For $\alpha > 0$, $\beta \in \mathbb{R}$ and distributions P, Q, the following results hold (whenever the relevant integrals and divergences are defined finitely, even in limiting sense).*

1. *$\mathcal{RE}_{1,1}(P,Q) = \mathcal{RE}(P,Q)$, the KLD measure.*
2. *$\mathcal{RE}_{\alpha,1}(P,Q) = \mathcal{RE}_\alpha(P,Q)$, the relative α-entropy.*
3. *$\mathcal{RE}_{1,\beta}(P,Q) = \frac{1}{\beta}\mathcal{D}_\beta(P,Q)$, a scaled Renyi divergence, which also coincides with the logarithmic power divergence measure of [80].*
4. *$\mathcal{RE}_{\alpha,\beta}(P,Q) = \frac{1}{\beta}\mathcal{D}_{\beta/\alpha}(P_\alpha,Q_\alpha)$, where P_α and Q_α are as defined in (5).*

Remark 1. *Note that, items 3 and 4 in Proposition 3 indicate a possible extension of the Renyi divergence measure over negative values of the tuning parameter β as follows:*

$$\mathcal{D}^*_\beta(P,Q) = \frac{1}{\beta}\mathcal{D}_\beta(P,Q), \ \ \beta \in \mathbb{R}\backslash\{0\}, \quad \mathcal{D}^*_0(P,Q) = \int q \log\left(\frac{q}{p}\right) d\mu.$$

Note that this modified Renyi divergence also coincides with the KLD measure at $\beta = 1$. Statistical applications of this divergence family have been studied by [80].

However, not all the members of the family of relative (α, β)-entropies are distinct or symmetric. For example, $\mathcal{RE}_{\alpha,0}(P, Q) = \mathcal{RE}_{\alpha,\alpha}(Q, P)$ for any $\alpha > 0$. The following proposition characterizes all such identities.

Proposition 4. *For $\alpha > 0$, $\beta \in \mathbb{R}$ and distributions P, Q, the relative (α, β)-entropy $\mathcal{RE}_{\alpha,\beta}(P, Q)$ is symmetric if and only if $\beta = \frac{\alpha}{2}$. In general, we have $\mathcal{RE}_{\alpha,\frac{\alpha}{2}-\gamma}(P, Q) = \mathcal{RE}_{\alpha,\frac{\alpha}{2}+\gamma}(Q, P)$ for any $\alpha > 0, \gamma \in \mathbb{R}$.*

Recall that the KLD measure is linked to the Shannon entropy and the relative α-entropy is linked with the Renyi entropy when the prior mismatched probability is uniform over the finite space. To derive such a relation for our general relative (α, β)-entropy, let us assume $\mu(\Omega) < \infty$ and let U denote the uniform probability measure on Ω. Then, we get

$$\mathcal{RE}_{\alpha,\beta}(P, U) = \frac{1}{\beta}\left[\log \mu(\Omega) - \mathcal{E}_{\alpha,\beta}(P)\right], \quad \beta \neq 0 \tag{17}$$

where the functional $\mathcal{E}_{\alpha,\beta}(P)$ is given in Definition 2 below and coincides with the Renyi entropy at $\beta = 1$. Thus, it can be used to define a two-parameter generalization of the Renyi entropy as follows.

Definition 2 (Generalized Renyi Entropy). *For any probability measure P over a measurable space Ω, we define the generalized Renyi entropy (GRE) of order (α, β) as*

$$\mathcal{E}_{\alpha,\beta}(P) = \frac{1}{\beta - \alpha}\log\left[\frac{\left(\int p^{\alpha}d\mu\right)^{\beta}}{\left(\int p^{\beta}d\mu\right)^{\alpha}}\right], \quad \alpha > 0, \beta \in \mathbb{R},\ \beta \neq 0, \alpha; \tag{18}$$

$$\mathcal{E}_{\alpha,\alpha}(P) = -\frac{\int \log(p)p^{\alpha}d\mu}{\int p^{\alpha}d\mu} + \frac{1}{\alpha}\log\left(\int p^{\alpha}d\mu\right), \quad \alpha > 0. \tag{19}$$

Note that, at $\beta = 1$, we have $\mathcal{E}_{\alpha,1}(P) = \mathcal{E}_{\alpha}(P)$, the usual Renyi entropy measure of order α.

The GRE is a new entropy to the best of our knowledge, and does not belong to the general class of entropy functionals as given in [104] which covers many existing entropies (including most, if not all, classical entropies). The following property of the functional $\mathcal{E}_{\alpha,\beta}(P)$ is easy to verify and justifies its use as a new entropy functional. To keep the focus of the present paper clear on the relative (α, β)-entropy, further properties of the GRE will be explored in our future work.

Theorem 1 (Entropic characteristics of GRE). *For any probability measure P over a finite measure space Ω, we have $0 \leq \mathcal{E}_{\alpha,\beta}(P) \leq \log \mu(\Omega)$ for all $\alpha > 0$ and $\beta \in \mathbb{R}\backslash\{0\}$. The two extremes are attained as follows.*

1. *$\mathcal{E}_{\alpha,\beta}(P) = 0$ if P is degenerate at a point in Ω (no uncertainty).*
2. *$\mathcal{E}_{\alpha,\beta}(P) = \log \mu(\Omega)$ if P is uniform over Ω (maximum uncertainty).*

Example 1 (Normal Distribution). *Consider distributions P_i from the most common class of multivariate (s-dimensional) normal distributions having mean $\boldsymbol{\mu}_i \in \mathbb{R}^s$ and variance matrix $\boldsymbol{\Sigma}_i$ for $i = 1, 2$. It is known that the Shannon and the Renyi entropies of P_1 are, respectively, given by*

$$\mathcal{E}(P_1) = \frac{s}{2} + \frac{s}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_1|,$$

$$\mathcal{E}_{\alpha}(P_1) = \frac{s}{2}\frac{\log\alpha}{\alpha - 1} + \frac{s}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_1|, \quad \alpha > 0, \alpha \neq 1.$$

With the new entropy measure, GRE, the entropy of the normal distribution P_1 can be seen to have the form

$$\begin{aligned}\mathcal{E}_{\alpha,\beta}(P_1) &= \frac{s}{2}\frac{(\alpha\log\beta-\beta\log\alpha)}{(\beta-\alpha)}+\frac{s}{2}\log(2\pi)+\frac{1}{2}\log|\boldsymbol{\Sigma}_1|,\ \ \alpha>0,\beta\in\mathbb{R}\backslash\{0,\alpha\},\\ \mathcal{E}_{\alpha,\alpha}(P_1) &= \frac{s}{2}(1-\log\alpha)+\frac{s}{2}\log(2\pi)+\frac{1}{2}\log|\boldsymbol{\Sigma}_1|,\ \ \alpha>0.\end{aligned}$$

Interestingly, the GRE of a normal distribution is effectively the same as its Shannon entropy or Renyi entropy up to an additive constant. However, similar characteristic does not hold between the relative entropy (KLD) and relative (α,β)-entropy. The KLD measure between two normal distributions P_1 and P_2 is given by

$$\mathcal{RE}(P_1,P_2) = \frac{1}{2}Trace(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)+\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right)-\frac{s}{2},$$

whereas the general relative (α,β)-entropy, with $\alpha>0$ and $\beta\in\mathbb{R}\backslash\{0,\alpha\}$, has the form

$$\begin{aligned}\mathcal{RE}_{\alpha,\beta}(P_1,P_2) &= \frac{\alpha}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\left[\beta\boldsymbol{\Sigma}_2+(\alpha-\beta)\boldsymbol{\Sigma}_1\right]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\\ &\quad+\frac{1}{2\beta(\beta-\alpha)}\log\left(\frac{|\boldsymbol{\Sigma}_2|^\beta|\boldsymbol{\Sigma}_1|^{\alpha-\beta}}{|\beta\boldsymbol{\Sigma}_2+(\alpha-\beta)\boldsymbol{\Sigma}_1|^\alpha}\right)-\frac{s\alpha\log\alpha}{2\beta(\alpha-\beta)}.\end{aligned}$$

Note that the relative (α,β)-entropy gives a more general divergence measure which utilizes different weights for the variance (or precision) matrix of the two normal distributions.

Example 2 (Exponential Distribution)**.** *Consider the exponential distribution P having density $p_\theta(x)=\theta e^{-\theta x}I(x\geq 0)$ with $\theta>0$. This distribution is very useful in lifetime modeling and reliability engineering; it is also the maximum entropy distribution of a non-negative random variable with fixed mean. The Shannon and the Renyi entropies of P are, respectively, given by*

$$\mathcal{E}(P)=1-\log\theta,\ \ \textit{and}\ \ \mathcal{E}_\alpha(P)=\frac{\log\alpha}{\alpha-1}-\log\theta,\ \ \alpha>0,\alpha\neq 1.$$

A simple calculation leads to the following form of the our new GRE measure of the exponential distribution P.

$$\begin{aligned}\mathcal{E}_{\alpha,\beta}(P) &= \frac{(\alpha\log\beta-\beta\log\alpha)}{(\beta-\alpha)}-\log\theta,\ \ \alpha>0,\beta\in\mathbb{R}\backslash\{0,\alpha\},\\ \mathcal{E}_{\alpha,\alpha}(P) &= (1-\log\alpha)-\log\theta,\ \ \alpha>0.\end{aligned}$$

Once again, the new GRE is effectively the same as the Shannon entropy or the Renyi entropy, up to an additive constant, for the exponential distribution as well.

Further, if P_1 and P_2 are two exponential distributions with parameters θ_1 and θ_2, respectively, the relative entropy (KLD) and the relative (α,β)-entropy between them are given by

$$\begin{aligned}\mathcal{RE}(P_1,P_2) &= \frac{\theta_2}{\theta_1}+\log\theta_1-\log\theta_2-1,\\ \mathcal{RE}_{\alpha,\beta}(P_1,P_2) &= \frac{\alpha}{\beta(\alpha-\beta)}\log\left[\beta\theta_1+(\alpha-\beta)\theta_2\right]-\frac{1}{\alpha-\beta}\log\theta_1-\frac{1}{\beta}\log\theta_2-\frac{\alpha\log\alpha}{\beta(\alpha-\beta)},\end{aligned}$$

for $\alpha>0$ and $\beta\in\mathbb{R}\backslash\{0,\alpha\}$. Clearly, the contributions of both the distribution is weighted differently by β and $(\alpha-\beta)$ in their relative (α,β)-entropy measure.

Before concluding this section, we study the nature of our relative (α,β)-entropy as $\alpha\to 0$. For this purpose, we restrict ourselves to the case of finite measure spaces with $\mu(\Omega)<\infty$. It is again straightforward to note that $\lim_{\alpha\to 0}\mathcal{RE}_{\alpha,\beta}(P,Q)=0$ for any $\beta\in\mathbb{R}$ and any distributions P and Q on Ω.

However, if we take the limit after scaling the relative entropy measure by α we get a non-degenerate divergence measure as follows.

$$\mathcal{RE}_{\beta}^{*}(P,Q) = \lim_{\alpha\downarrow 0}\frac{1}{\alpha}\mathcal{RE}_{\alpha,\beta}(P,Q) = \frac{1}{\beta^2}\left[\log\int\left(\frac{p}{q}\right)^{\beta}d\mu - \frac{\beta}{\mu(\Omega)}\int\log\left(\frac{p}{q}\right)d\mu - \log\mu(\Omega)\right],$$

for $\beta \in \mathbb{R}\backslash\{0\}$, and

$$\mathcal{RE}_{0}^{*}(P,Q) = \lim_{\alpha\downarrow 0}\frac{1}{\alpha}\mathcal{RE}_{\alpha,0}(P,Q) = \frac{1}{2\mu(\Omega)}\left[\int\{\log(p/q)\}^2 d\mu - \frac{1}{\mu(\Omega)}\left\{\int\log(p/q)\,d\mu\right\}^2\right].$$

These interesting relative entropy measures again define a subfamily of valid statistical divergences, from its construction. The particular member at $\beta = 1$ is linked to the LDPD (or the γ-divergence) with tuning parameter -1 and can be thought of as a logarithmic extension of the famous Itakura–Saito divergence [105] given by

$$D_{IS}(P,Q) = \int\left(\frac{p}{q}\right)d\mu - \int\log\left(\frac{p}{q}\right)d\mu - \mu(\Omega). \tag{20}$$

This Itakura–Saito-divergence has been successfully applied to non-negative matrix factorization in different applications [106] which can be extended by using the new divergence family $\mathcal{RE}_{\beta}^{*}(P,Q)$ in future works.

3. Geometry of the Relative (α,β)-Entropy

3.1. Continuity

We start the exploration of the geometric properties of the relative (α,β)-entropy with its continuity over the functional space $L_\alpha(\mu)$. In the following, we interchangeably use the notation $\mathcal{RE}_{\alpha,\beta}(p,q)$ and $D_\lambda(p,q)$ to denote $\mathcal{RE}_{\alpha,\beta}(P,Q)$ and $D_\lambda(P,Q)$, respectively. Our results generalize the corresponding properties of the relative α-entropy from [16,73] to our relative (α,β)-entropy or equivalent LSD measure.

Proposition 5. *For a given $q \in L_\alpha(\mu)$, consider the function $p \mapsto \mathcal{RE}_{\alpha,\beta}(p,q)$ from $p \in L_\alpha(\mu)$ to $[0,\infty]$. This function is lower semi-continuous in $L_\alpha(\mu)$ for any $\alpha > 0$, $\beta \in \mathbb{R}$. Additionally, it is continuous in $L_\alpha(\mu)$ when $\alpha > \beta > 0$ and the relative entropy is finitely defined.*

Proof. First let us consider any $\alpha > 0$ and take $p_n \to p$ in $L_\alpha(\mu)$. Then, $||p_n||_\alpha \to ||p||_\alpha$. Also, $|p_n^\alpha - p^\alpha| \leq |p_n|^\alpha + |p|^\alpha$ and hence a general version of the dominated convergence theorem yields $p_n^\alpha \to p^\alpha$ in $L_1(\mu)$. Thus, we get

$$p_{n,\alpha} := \frac{p_n^\alpha}{\int p_n^\alpha d\mu} \to p_\alpha \quad \text{in } L_1(\mu). \tag{21}$$

Further, following ([107], Lemma 1), we know that the function $h \to \int\phi_\lambda(h)d\nu$ is lower semi-continuous in $L_1(\nu)$ for any $\lambda \in \mathbb{R}$ and any probability measure ν on $(\Omega,\mathcal{A})$. Taking $\nu = Q_\alpha$, we get from (21) that $p_{n,\alpha}/q_\alpha \to p_\alpha/q_\alpha$ in $L_1(\nu)$. Therefore, the above lower semi-continuity result along with (9) implies that

$$\liminf_{n\to\infty} D_\lambda(p_{n,\alpha},q_\alpha) \geq D_\lambda(p_\alpha,q_\alpha) \geq 0, \;\; \lambda \in \mathbb{R}. \tag{22}$$

Now, note that the function $\psi(u) = \frac{1}{\rho}\log(sign(\rho)u + 1)$ is continuous and increasing on $[0, \infty)$ for $\rho > 0$ and on $[0, 1)$ for $\rho < 0$. Thus, combining (22) with the definition of the relative (α, β)-entropy in (13), we get that

$$\liminf_{n\to\infty} \mathcal{RE}_{\alpha,\beta}(p_n, q) \geq \mathcal{RE}_{\alpha,\beta}(p, q), \tag{23}$$

i.e., the function $p \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is lower semi-continuous.

Finally, consider the case $\alpha > \beta > 0$. Note that the dual space of $L_{\alpha/\beta}(\mu)$ is $L_{\frac{\alpha}{\alpha-\beta}}(\mu)$ since $\alpha > \beta > 0$. Also, for $q \in L_\alpha(\mu)$, we have $\left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} \in L_{\frac{\alpha}{\alpha-\beta}}(\mu)$, the dual space of the Banach space $L_{\alpha/\beta}(\mu)$. Therefore, the function $T : L_{\alpha/\beta}(\mu) \mapsto \mathbb{R}$ defined by

$$T(h) = \int h \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu, \;\; h \in L_{\alpha/\beta}(\mu),$$

is a bounded linear functional and hence continuous. Now, take $p_n \to p$ in $L_\alpha(\mu)$ so that $||p_n||_\alpha \to ||p||_\alpha$ as $n \to \infty$. Therefore, $\left(\frac{p_n}{||p_n||_\alpha}\right) \to \left(\frac{p}{||p||_\alpha}\right)$ in $L_\alpha(\mu)$ implying $\left(\frac{p_n}{||p_n||_\alpha}\right)^\beta \to \left(\frac{p}{||p||_\alpha}\right)^\beta$ in $L_{\alpha/\beta}(\mu)$. Hence, by the continuity of T on $L_{\alpha/\beta}(\mu)$, we get

$$T\left(\left(\frac{p_n}{||p_n||_\alpha}\right)^\beta\right) \to T\left(\left(\frac{p}{||p||_\alpha}\right)^\beta\right), \quad \text{as } n \to \infty.$$

However, from (14), we get

$$\mathcal{RE}_{\alpha,\beta}(p_n, q) = \frac{\alpha}{\beta(\beta-\alpha)} \log T\left(\left(\frac{p_n}{||p_n||_\alpha}\right)^\beta\right) \to \frac{\alpha}{\beta(\beta-\alpha)} \log T\left(\left(\frac{p}{||p||_\alpha}\right)^\beta\right) = \mathcal{RE}_{\alpha,\beta}(p, q). \tag{24}$$

This proves the continuity of $\mathcal{RE}_{\alpha,\beta}(p, q)$ in its first argument when $\alpha > \beta > 0$. □

Remark 2. *Whenever Ω is finite (discrete) equipped with the counting measure μ, all integrals in the definition of $\mathcal{RE}_{\alpha,\beta}(P, Q)$ become finite sums and any limit can be taken inside these finite sums. Thus, whenever defined finitely, the function $p \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is always continuous in this case.*

Remark 3. *For a general infinite space Ω, the function $p \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is not necessarily continuous for the cases $\alpha < \beta$. This can be seen by using the same counterexample as given in Remark 3 of [16]. However, it is yet to be verified if this function can be continuous for $\beta < 0$ cases.*

Proposition 6. *For a given $p \in L_\alpha(\mu)$, consider the function $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ from $q \in L_\alpha(\mu)$ to $[0, \infty]$. This function is lower semi-continuous in $L_\alpha(\mu)$ for any $\alpha > 0$ and $\beta \in \mathbb{R}$.*

Proof. Fix an $\alpha > 0$ and $\beta \in \mathbb{R}$, which in turn fixes a $\lambda \in \mathbb{R}$. Note that, the relative (α, β)-entropy measure can be re-expressed from (13) as

$$\mathcal{RE}_{\alpha,\beta}(p, q) = \frac{1}{\beta\lambda} \log\left[sign(\beta\lambda) D_{-(\lambda+1)}(q_\alpha, p_\alpha) + 1\right]. \tag{25}$$

Now, consider a sequence $q_n \to q$ in $L_\alpha(\mu)$ and proceed as in the proof of Proposition 5 using ([107], Lemma 1) to obtain

$$\liminf_{n\to\infty} D_{-(\lambda+1)}(q_{n,\alpha}, p_\alpha) \geq D_{-(\lambda+1)}(q_\alpha, p_\alpha) \geq 0, \;\; \lambda \in \mathbb{R}. \tag{26}$$

Now, whenever $D_{-(\lambda+1)}(q_\alpha, p_\alpha) = 1$ with $\beta\lambda < 0$ or $D_{-(\lambda+1)}(q_\alpha, p_\alpha) = \infty$ with $\beta\lambda > 0$, we get from (25) and (26) that

$$\liminf_{n\to\infty} \mathcal{RE}_{\alpha,\beta}(p, q_n) = \mathcal{RE}_{\alpha,\beta}(p, q) = +\infty. \tag{27}$$

In all other cases, we consider the function $\psi(u) = \frac{1}{\rho}\log(sign(\rho)u + 1)$ as in the proof of Proposition 5. This function is continuous and increasing whenever the corresponding relative entropy is finitely defined for all tuning parameter values; on $[0, \infty)$ for $\rho > 0$ and on $[0, 1)$ for $\rho < 0$. Hence, again combining (26) with (25) through the function ψ, we conclude that

$$\liminf_{n\to\infty} \mathcal{RE}_{\alpha,\beta}(p, q_n) \geq \mathcal{RE}_{\alpha,\beta}(p, q). \tag{28}$$

Therefore, the function $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is also lower semi-continuous. □

Remark 4. *As in Remark 2, whenever Ω is finite (discrete) and is equipped with the counting measure μ, the function $q \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is continuous in $L_\alpha(\mu)$ for any fixed $p \in L_\alpha(\mu)$, $\alpha > 0$ and $\beta \in \mathbb{R}$.*

3.2. Convexity

It has been shown in [16] that the relative α-entropy (i.e., $\mathcal{RE}_{\alpha,1}(p, q)$) is neither convex nor bi-convex, but it is quasi-convex in p. For general $\beta \neq 1$, however, the relative (α, β)-entropy $\mathcal{RE}_{\alpha,\beta}(p, q)$ is not even quasi-convex in $p \in L_\alpha(\mu)$; rather it is quasi-convex on the β-power transformed space of densities, $L_\alpha(\mu)^\beta = \{p^\beta : p \in L_\alpha(\mu)\}$, as described in the following theorem. Note that, for $\alpha, \beta > 0$, $L_\alpha(\mu)^\beta = L_{\alpha/\beta}(\mu)$. Here we define the lower level set $B_{\alpha,\beta}(q, r) = \{p : \mathcal{RE}_{\alpha,\beta}(p, q) \leq r\}$ and its power-transformed set $B_{\alpha,\beta}(q, r)^\beta = \{p^\beta : p \in B_{\alpha,\beta}(q, r)\}$, for any $q \in L_\alpha(\mu)$ and $r > 0$.

Theorem 2. *For any given $\alpha > 0$, $\beta \in \mathbb{R}$ and $q \in L_\alpha(\mu)$, the sets $B_{\alpha,\beta}(q, r)^\beta$ are convex for all $r > 0$. Therefore, the function $p^\beta \mapsto \mathcal{RE}_{\alpha,\beta}(p, q)$ is quasi-convex on $L_\alpha(\mu)^\beta$.*

Proof. Note that, at $\beta = 1$, our theorem coincides with Proposition 5 of [16]; so we will prove the result for the case $\beta \neq 1$. Fix $\alpha, r > 0$, a real $\beta \notin \{1, \alpha\}$, $q \in L_\alpha(\mu)$, and $p_0, p_1 \in B_{\alpha,\beta}(q, r)$. Then $p_0^\beta, p_1^\beta \in B_{\alpha,\beta}(q, r)^\beta$. For $\tau \in [0, 1]$, we consider $p_\tau^\beta = \tau p_1^\beta + \bar{\tau} p_0^\beta$ with $\bar{\tau} = 1 - \tau$. We need to show that $p_\tau^\beta \in B_{\alpha,\beta}(q, r)^\beta$, i.e., $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \leq r$.

Now, from (14), we have

$$\mathcal{RE}_{\alpha,\beta}(p, q) = \frac{1}{\beta\lambda} \log \int \left(\frac{p}{||p||_\alpha}\right)^\beta \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu = \frac{1}{\beta\lambda} \log \int \left(\frac{p_\alpha}{q_\alpha}\right)^{\beta/\alpha} dQ_\alpha. \tag{29}$$

Since $p_0^\beta, p_1^\beta \in B_{\alpha,\beta}(q, r)^\beta$, we have

$$sign(\beta\lambda) \int \left(\frac{p_\tau}{||p_\tau||_\alpha}\right)^\beta \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu \leq sign(\beta\lambda) e^{r\beta\lambda}, \quad \text{for } \tau = 0, 1. \tag{30}$$

For any $\tau \in (0, 1)$, we get

$$\begin{aligned} sign(\beta\lambda) \int \left(\frac{p_\tau}{||p_\tau||_\alpha}\right)^\beta \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu = {} & sign(\beta\lambda) \int \left(\frac{\tau p_1^\beta + \bar{\tau} p_0^\beta}{||p_\tau||_\alpha^\beta}\right) \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu, && \text{[by definition of } p_\tau\text{]} \\ \leq {} & sign(\beta\lambda) e^{r\beta\lambda} \frac{\tau ||p_1||_\alpha^\beta + \bar{\tau} ||p_0||_\alpha^\beta}{||p_\tau||_\alpha^\beta}, && \text{[by (30)].} \end{aligned} \tag{31}$$

Now, using the extended Minkowski's inequalities from Lemma 1, given below, along with (31) and noting that $\beta\lambda = \beta(\beta-\alpha)/\alpha$, we get that

$$sign(\beta\lambda)\int\left(\frac{p_\tau}{||p_\tau||_\alpha}\right)^\beta\left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta}d\mu \quad\le\quad sign(\beta\lambda)e^{r\beta\lambda}.$$

Therefore, by (29) and the fact that $\frac{1}{\rho}\log(sign(\rho)u)$ is increasing in u, we finally get $\mathcal{RE}_{\alpha,\beta}(p_\tau,q)\le r$. This proves the result for $\alpha\neq\beta$.

The case $\beta=\alpha$ can be proved in a similar manner and is left as an exercise to the readers. □

Lemma 1 (Extended Minkowski's inequality). *Fix $\alpha>0$, a real $\beta\notin\{1,\alpha\}$, $p_0,p_1\in L_\alpha(\mu)$, and $\tau\in[0,1]$. Define $p_\tau^\beta=\tau p_1^\beta+\bar{\tau}p_0^\beta$ with $\bar{\tau}=1-\tau$. Then we have the following inequalities:*

$$||p_\tau||_\alpha^\beta \quad\ge\quad \tau||p_1||_\alpha^\beta+\bar{\tau}||p_0||_\alpha^\beta, \quad \text{if } \beta(\beta-\alpha)>0, \tag{32}$$

$$||p_\tau||_\alpha^\beta \quad\le\quad \tau||p_1||_\alpha^\beta+\bar{\tau}||p_0||_\alpha^\beta, \quad \text{if } \beta(\beta-\alpha)<0. \tag{33}$$

Proof. It follows by using the Jensen's inequality and the convexity of the function $x^{\beta/\alpha}$. □

Next, note in view of Proposition 4 that, for any $p,q\in L_\alpha(\mu)$, $\mathcal{RE}_{\alpha,\beta}(p,q)=\mathcal{RE}_{\alpha,\alpha-\beta}(q,p)$. Using this result along with the above theorem, we also get the quasi-convexity of the relative (α,β)-entropy $\mathcal{RE}_{\alpha,\beta}(p,q)$ in q over a different power transformed space of densities. This leads to the following theorem.

Theorem 3. *For any given $\alpha>0$, $\beta\in\mathbb{R}$ and $p\in L_\alpha(\mu)$, the function $q^{\alpha-\beta}\mapsto\mathcal{RE}_{\alpha,\beta}(p,q)$ is quasi-convex on $L_\alpha(\mu)^{\alpha-\beta}$. In particular, for the choice $\beta=\alpha-1$, the function $q\mapsto\mathcal{RE}_{\alpha,\beta}(p,q)$ is quasi-convex on $L_\alpha(\mu)$.*

Remark 5. *Note that, at $\alpha=\beta=1$, the $\mathcal{RE}_{1,1}(p,q)$ coincides with the KLD measure (or relative entropy) which is quasi-convex in both the arguments p and q on $L_\alpha(\mu)$.*

3.3. Extended Pythagorean Relation

Motivated by the quasi-convexity of $\mathcal{RE}_{\alpha,\beta}(p,q)$ on $L_\alpha(\mu)^\beta$, we now present a Pythagorean-type result for the general relative (α,β)-entropy over the power-transformed space. It generalizes the corresponding result for relative α-entropy [16]; the proof is similar to that in [16] with necessary modifications due to the transformation of the domain space.

Theorem 4 (Pythagorean Property). *Fix an $\alpha>0$, $\beta\in\mathbb{R}$ with $\beta\neq\alpha$ and $p_0,p_1,q\in L_\alpha(\mu)$. Define $p_\tau\in L_\alpha(\mu)$ by $p_\tau^\beta=\tau p_1^\beta+\bar{\tau}p_0^\beta$ for $\tau\in[0,1]$ and $\bar{\tau}=1-\tau$.*

(i) *Suppose $\mathcal{RE}_{\alpha,\beta}(p_0,q)$ and $\mathcal{RE}_{\alpha,\beta}(p_1,q)$ are finite. Then, $\mathcal{RE}_{\alpha,\beta}(p_\tau,q)\ge\mathcal{RE}_{\alpha,\beta}(p_0,q)$ for all $\tau\in[0,1]$, i.e., the back-transformation of line segment joining p_1^β and p_0^β on $L_\alpha(\mu)^\beta$ to $L_\alpha(\mu)$ does not intersect $B_{\alpha,\beta}(q,\mathcal{RE}_{\alpha,\beta}(p_0,q))$, if and only if*

$$\mathcal{RE}_{\alpha,\beta}(p_1,q)\ge\mathcal{RE}_{\alpha,\beta}(p_1,p_0)+\mathcal{RE}_{\alpha,\beta}(p_0,q). \tag{34}$$

(ii) *Suppose $\mathcal{RE}_{\alpha,\beta}(p_\tau,q)$ is finite for some fixed $\tau\in(0,1)$. Then, the back-transformation of line segment joining p_1^β and p_0^β on $L_\alpha(\mu)^\beta$ to $L_\alpha(\mu)$ does not intersect $B_{\alpha,\beta}(q,\mathcal{RE}_{\alpha,\beta}(p_\tau,q))$ if and only if*

$$\mathcal{RE}_{\alpha,\beta}(p_1,q) \quad=\quad \mathcal{RE}_{\alpha,\beta}(p_1,p_\tau)+\mathcal{RE}_{\alpha,\beta}(p_\tau,q), \tag{35}$$

$$\text{and}\quad \mathcal{RE}_{\alpha,\beta}(p_0,q) \quad=\quad \mathcal{RE}_{\alpha,\beta}(p_0,p_\tau)+\mathcal{RE}_{\alpha,\beta}(p_\tau,q). \tag{36}$$

Proof of Part (i). Let $P_{\tau,\alpha}$ to be the probability measure having μ-density $p_{\tau,\alpha} = \frac{p_\tau^\alpha}{\int p_\tau^\alpha d\mu}$ for $\tau \in [0,1]$. Also note that, with $\lambda = \beta/\alpha - 1$, we have

$$D_\lambda(P_\alpha, Q_\alpha) = sign(\beta\lambda)\left[\int \left(\frac{p}{||p||_\alpha}\right)^\beta (q_\alpha)^{-\lambda} d\mu - 1\right], \text{ for } p, q \in L_\alpha(\mu). \tag{37}$$

Thus, (34) is equivalent to the statement

$$sign(\beta\lambda)||p_0||_\alpha^\beta \int p_1^\beta (q_\alpha)^{-\lambda} d\mu \geq sign(\beta\lambda) \int p_1^\beta (p_{0,\alpha})^{-\lambda} d\mu \cdot \int p_0^\beta (q_\alpha)^{-\lambda} d\mu. \tag{38}$$

and we have

$$D_\lambda(P_{\tau,\alpha}, Q_\alpha) = sign(\beta\lambda)\left[\int \left(\frac{p_\tau}{||p_\tau||_\alpha}\right)^\beta (q_\alpha)^{-\lambda} d\mu - 1\right] = sign(\beta\lambda)\frac{s(\tau)}{t(\tau)}, \tag{39}$$

where $s(\tau) = \int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu$ and $t(\tau) = ||p_\tau||_\alpha^\beta$. Now consider the two implications separately.

Only if statement: Now, let us assume that $\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, q)$ for all $\tau \in (0,1)$. Then, we get $\frac{1}{\tau}\left[D_\lambda(P_{\tau,\alpha}, Q_\alpha) - D_\lambda(P_{0,\alpha}, Q_\alpha)\right] \geq 0$ for all $\tau \in (0,1)$. Letting $\tau \downarrow 0$, we get that

$$\frac{\partial}{\partial\tau} D_\lambda(P_{\tau,\alpha}, Q_\alpha)\Big|_{\tau=0} \geq 0. \tag{40}$$

In order to find the derivative of $D_\lambda(P_{\tau,\alpha}, Q_\alpha)$, we first note that

$$\frac{s(\tau) - s(0)}{\tau} = \frac{1}{\tau}\left[\int p_\tau^\beta (q_\alpha)^{-\lambda} d\mu - \int p_0^\beta (q_\alpha)^{-\lambda} d\mu\right] = \int (p_1^\beta - p_0^\beta)(q_\alpha)^{-\lambda} d\mu,$$

and hence

$$s'(0) = \lim_{\tau\downarrow 0} \frac{s(\tau) - s(0)}{\tau} = \int (p_1^\beta - p_0^\beta)(q_\alpha)^{-\lambda} d\mu. \tag{41}$$

Further, using a simple modification of the techniques in the proof of ([16], Theorem 9), it is easy to verify that the derivative of $t(\tau)$ with respect to τ exists and is given by

$$t'(\tau) = \left(\int p_\tau^\alpha d\mu\right)^{\frac{(\beta-\alpha)}{\alpha}} \int p_\tau^{\alpha-\beta}(p_1^\beta - p_0^\beta) d\mu.$$

Hence we get

$$t'(0) = \left(\int p_0^\alpha d\mu\right)^{\frac{(\beta-\alpha)}{\alpha}} \int p_0^{\alpha-\beta}(p_1^\beta - p_0^\beta) d\mu = \int p_1^\beta (p_{0,\alpha})^{-\lambda} d\mu - ||p_0||_\alpha^\beta. \tag{42}$$

Therefore, the derivative of $D_\lambda(P_{\tau,\alpha}, Q_\alpha) = sign(\beta\lambda)s(\tau)/t(\tau)$ exists and is given by $sign(\beta\lambda)\left[t(0)s'(0) - t'(0)s(0)\right]/t(0)^2$. Therefore, using (40), we get that

$$sign(\beta\lambda)t(0)s'(0) \geq sign(\beta\lambda)t'(0)s(0), \tag{43}$$

which implies (38) after substituting the values from (41) and (42).

If statement: Now, let us assume that (34)—or equivalently (38)—holds true. Further, as in the derivation of (38), we can start from the trivial statement

$$\mathcal{RE}_{\alpha,\beta}(p_0, q) = \mathcal{RE}_{\alpha,\beta}(p_0, p_0) + \mathcal{RE}_{\alpha,\beta}(p_0, q),$$

to deduce

$$sign(\beta\lambda)||p_0||_\alpha^\beta \int p_0^\beta \left(q_\alpha\right)^{-\lambda} d\mu = sign(\beta\lambda) \int p_0^\beta \left(p_{0,\alpha}\right)^{-\lambda} d\mu \cdot \int p_0^\beta \left(q_\alpha\right)^{-\lambda} d\mu. \tag{44}$$

Now, multiply (38) by τ and (44) by $\bar{\tau}$, and add to get

$$sign(\beta\lambda)||p_0||_\alpha^\beta \int p_\tau^\beta \left(q_\alpha\right)^{-\lambda} d\mu \geq sign(\beta\lambda) \int p_\tau^\beta \left(p_{0,\alpha}\right)^{-\lambda} d\mu \cdot \int p_0^\beta \left(q_\alpha\right)^{-\lambda} d\mu.$$

In view of (37), this implies that

$$\mathcal{RE}_{\alpha,\beta}(p_\tau, q) \geq \mathcal{RE}_{\alpha,\beta}(p_\tau, p_0) + \mathcal{RE}_{\alpha,\beta}(p_0, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, q).$$

This proves the if statement of Part (i) completing the proof. □

Proof of Part (ii). Note that the *if statement* follows directly from Part (i).
To prove the *only if statement*, we first show that $\mathcal{RE}_{\alpha,\beta}(p_1, q)$ and $\mathcal{RE}_{\alpha,\beta}(p_0, q)$ are finite since $\mathcal{RE}_{\alpha,\beta}(p_\tau, q)$ is finite. For this purpose, we note that $p_1^\beta \leq \tau^{-1} p_\tau^\beta$ by the definition of p_τ and hence $(p_1/q)^\beta \leq \tau^{-1}(p_\tau/q)^\beta$. Therefore, we get

$$\left(\frac{p_{1,\alpha}}{q_\alpha}\right)^{\beta/\alpha} = \left(\frac{p_1}{q}\right)^\beta \left(\frac{||q||}{||p_1||}\right)^\beta \leq \frac{1}{\tau}\left(\frac{p_\tau}{q}\right)^\beta \left(\frac{||q||}{||p_1||}\right)^\beta = \frac{1}{\tau}\left(\frac{p_{\tau,\alpha}}{q_\alpha}\right)^\beta \left(\frac{||p_\tau||}{||p_1||}\right)^\beta. \tag{45}$$

Integration with respect to Q_α and using (29), we get $\mathcal{RE}_{\alpha,\beta}(p_1, q) \leq \mathcal{RE}_{\alpha,\beta}(p_\tau, q) + c < \infty$, where c is a constant. Similarly one can also show that $\mathcal{RE}_{\alpha,\beta}(p_0, q) < \infty$.

Therefore, we can apply Part (i) to conclude that

$$\mathcal{RE}_{\alpha,\beta}(p_1, q) \geq \mathcal{RE}_{\alpha,\beta}(p_1, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q), \text{ and } \mathcal{RE}_{\alpha,\beta}(p_0, q) \geq \mathcal{RE}_{\alpha,\beta}(p_0, p_\tau) + \mathcal{RE}_{\alpha,\beta}(p_\tau, q). \tag{46}$$

These relations imply that

$$sign(\beta\lambda)||p_\tau||_\alpha^\beta \int p_1^\beta \left(q_\alpha\right)^{-\lambda} d\mu \geq sign(\beta\lambda) \int p_1^\beta \left(p_{\tau,\alpha}\right)^{-\lambda} d\mu \cdot \int p_\tau^\beta \left(q_\alpha\right)^{-\lambda} d\mu, \tag{47}$$

$$\text{and } sign(\beta\lambda)||p_\tau||_\alpha^\beta \int p_0^\beta \left(q_\alpha\right)^{-\lambda} d\mu \geq sign(\beta\lambda) \int p_0^\beta \left(p_{\tau,\alpha}\right)^{-\lambda} d\mu \cdot \int p_\tau^\beta \left(q_\alpha\right)^{-\lambda} d\mu. \tag{48}$$

The proof of the above results proceed in a manner analogous to the proof of (38). Now, if either of the inequalities in (46) is strict, the corresponding inequality in (47) or (48) will also be strict. Then, multiplying (47) and (48) by τ and $\bar{\tau}$, respectively, and adding them we get (44) with a strict inequality (in place of an equality), which is a contradiction. Hence, both inequalities in (46) must be equalities implying (35) and (36). This completes the proof. □

Note that, at $\beta = 1$, the above theorem coincides with Theorem 9 of [16]. However, for general α, β as well, the above extended Pythagorean relation for the relative (α, β)-entropy suggests that it behaves “like" a squared distance (although with a non-linear space transformation). So, one can meaningfully define its projection on to a suitable set which we will explore in the following sections.

4. The Forward Projection of Relative (α, β)-Entropy

The forward projection, i.e., minimization with respect to the first argument given a fixed second argument, leads to the important maximum entropy principle of information theory; it also relates to the Gibbs conditioning principle from statistical physics [16]. Let us now formally define and study the forward projection of the relative (α, β)-entropy. Let $\mathbb{S}^*$ denote the set of probability measure on $(\Omega, \mathcal{A})$ and let the set of corresponding μ-densities be denoted by $\mathbb{S} = \{p = dP/d\mu : P \in \mathbb{S}^*\}$.

Definition 3 (Forward (α, β)-Projection)**.** *Fix $Q \in \mathbb{S}^*$ having μ-density $q \in L_\alpha(\mu)$. Let $\mathbb{E} \subset \mathbb{S}$ with $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$ for some $p \in \mathbb{E}$. Then, $p^* \in \mathbb{E}$ is called the forward projection of the relative (α, β)-entropy or simply the forward (α, β)-projection (or forward LSD projection) of q on $\mathbb{E}$ if it satisfies the relation*

$$\mathcal{RE}_{\alpha,\beta}(p^*, q) = \inf_{p\in\mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q). \tag{49}$$

Note that we must assume that, $\mathbb{E} \subset L_\alpha(\mu)$ so that the above relative (α, β)-entropy is finitely defined for $p \in \mathbb{E}$.

We first prove the uniqueness of the forward (α, β)-projection from the Pythagorean property, whenever it exists. The following theorem describe the connection of the forward (α, β)-projection with Pythagorean relation; the proof is same as that of ([16], Theorem 10) using Theorem 4 and hence omitted for brevity.

Theorem 5. *Consider the set $\mathbb{E} \subset \mathbb{S}$ such that $\mathbb{E}^\beta$ is convex and fix $q \in L_\alpha(\mu)$. Then, $p^* \in \mathbb{E} \cap B_{\alpha,\beta}(q, \infty)$ is a forward (α, β)-projection of q on $\mathbb{E}$ if and only if every $p \in \mathbb{E} \cap B_{\alpha,\beta}(q, \infty)$ satisfies*

$$\mathcal{RE}_{\alpha,\beta}(p, q) \geq \mathcal{RE}_{\alpha,\beta}(p, p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q). \tag{50}$$

Further, if $(p^*)^\beta$ is an algebraic inner point of $\mathbb{E}^\beta$, i.e., for every $p \in \mathbb{E}$ there exists $p' \in \mathbb{E}$ and $\tau \in (0, 1)$ such that $(p^*)^\beta = \tau p^\beta + (1 - \tau)(p')^\beta$, then every $p \in \mathbb{E}$ satisfies $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$ and

$$\mathcal{RE}_{\alpha,\beta}(p, q) = \mathcal{RE}_{\alpha,\beta}(p, p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q), \text{ and } \mathcal{RE}_{\alpha,\beta}(p', q) = \mathcal{RE}_{\alpha,\beta}(p', p^*) + \mathcal{RE}_{\alpha,\beta}(p^*, q).$$

Corollary 1 (Uniqueness of Forward (α, β)-Projection)**.** *Consider the set $\mathbb{E} \subset \mathbb{S}$ such that $\mathbb{E}^\beta$ is convex and fix $q \in L_\alpha(\mu)$. If a forward (α, β)-projection of q on $\mathbb{E}$ exists, it must be unique a.s.$[\mu]$.*

Proof. Suppose p_1^* and p_2^* are two forward (α, β)-projection of q on $\mathbb{E}$. Then, by definition, $\mathcal{RE}_{\alpha,\beta}(p_1^*, q) = \mathcal{RE}_{\alpha,\beta}(p_2^*, q) < \infty$. Applying Theorem 5 with $p^* = p_1^*$ and $p = p_2^*$, we get

$$\mathcal{RE}_{\alpha,\beta}(p_2^*, q) \geq \mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) + \mathcal{RE}_{\alpha,\beta}(p_1^*, q).$$

Hence $\mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) \leq 0$ or $\mathcal{RE}_{\alpha,\beta}(p_2^*, p_1^*) = 0$ by non-negativity of relative entropy, which further implies that $p_1^* = p_2^*$ a.s.$[\mu]$ by Proposition 1. □

Next we will show the existence of the forward (α, β)-projection under suitable conditions. We need to use an extended Apollonius Theorem for the ϕ-divergence measure D_λ used in the definition (13) of the relative (α, β)-entropy. Such a result is proved in [16] for the special case $\alpha(1 + \lambda) = 1$; the following lemma extends it for the general case $\alpha(1 + \lambda) = \beta \in \mathbb{R}$.

Lemma 2. *Fix $p_0, p_1, q \in L_\alpha(\mu)$, $\tau \in [0, 1]$ and $\alpha(1 + \lambda) = \beta \in \mathbb{R}$ with $\alpha > 0$ and define r satisfying*

$$r^\beta = \frac{\frac{\tau}{||p_1||_\alpha^\beta} p_1^\beta + \frac{1-\tau}{||p_0||_\alpha^\beta} p_0^\beta}{\frac{\tau}{||p_1||_\alpha^\beta} + \frac{1-\tau}{||p_0||_\alpha^\beta}}. \tag{51}$$

Let $p_{j,\alpha} = p_j^\alpha / \int p_j^\alpha d\mu$ for $j = 0, 1$, and similarly q_α and r_α. Then, if $\beta(\beta - \alpha) > 0$ we have

$$\tau D_\lambda(p_{1,\alpha}, q_\alpha) + (1 - \tau) D_\lambda(p_{0,\alpha}, q_\alpha) \geq \tau D_\lambda(p_{1,\alpha}, r_\alpha) + (1 - \tau) D_\lambda(p_{0,\alpha}, r_\alpha) + D_\lambda(r_\alpha, q_\alpha), \tag{52}$$

but the inequality gets reversed if $\beta(\beta - \alpha) < 0$.

Proof. By (37), we get

$$\begin{aligned}
&\tau D_\lambda(p_{1,\alpha}, q_\alpha) + (1-\tau)D_\lambda(p_{0,\alpha}, q_\alpha) - \tau D_\lambda(p_{1,\alpha}, r_\alpha) - (1-\tau)D_\lambda(p_{0,\alpha}, r_\alpha)\\
&= sign(\beta\lambda)\tau \int \left(\frac{p_1}{||p_1||_\alpha}\right)^\beta \left[(q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda}\right] d\mu + sign(\beta\lambda)(1-\tau)\int \left(\frac{p_0}{||p_0||_\alpha}\right)^\beta \left[(q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda}\right] d\mu\\
&= sign(\beta\lambda)||r||_\alpha^\beta \left[\frac{\tau}{||p_1||_\alpha^\beta} + \frac{1-\tau}{||p_0||_\alpha^\beta}\right] \int \left(\frac{r}{||r||_\alpha}\right)^\beta \left[(q_\alpha)^{-\lambda} - (r_\alpha)^{-\lambda}\right] d\mu\\
&= sign(\beta\lambda)||r||_\alpha^\beta \left[\frac{\tau}{||p_1||_\alpha^\beta} + \frac{1-\tau}{||p_0||_\alpha^\beta}\right] D_\lambda(R_\alpha, Q_\alpha).
\end{aligned}$$

Then the Lemma follows by an application of the extended Minkowski's inequalities (32) and (33) from Lemma 1. □

We now present the sufficient conditions for the existence of the forward (α, β)-projection in the following theorem.

Theorem 6 (Existence of Forward (α, β)-Projection). *Fix $\alpha > 0$ and $\beta \in \mathbb{R}$ with $\beta \neq \alpha$ and $q \in L_\alpha(\mu)$. Given any set $\mathbb{E} \subset \mathbb{S}$ for which $\mathbb{E}^\beta$ is convex and closed and $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$ for some $p \in \mathbb{E}$, a forward (α, β)-projection of q on $\mathbb{E}$ always exists (and it is unique by Corollary 1).*

Proof. We prove it separately for the cases $\beta\lambda > 0$ and $\beta\lambda < 0$, extending the arguments from [16]. The case $\beta\lambda = 0$ can be obtained from these two cases by standard limiting arguments and hence omitted for brevity.

The Case $\beta\lambda > 0$:

Consider a sequence $\{p_n\} \subset \mathbb{E}$ such that $D_\lambda(p_{n,\alpha}, q_\alpha) < \infty$ for each n and $D_\lambda(p_{n,\alpha}, q_\alpha) \to \inf_{p\in\mathbb{E}} D_\lambda(p_\alpha, q_\alpha)$ as $n \to \infty$. Then, by Lemma 2 applied to p_m and p_n with $\tau = 1/2$, we get

$$\frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) \geq \frac{1}{2}D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + \frac{1}{2}D_\lambda(p_{n,\alpha}, r_{m,n,\alpha}) + D_\lambda(r_{m,n,\alpha}, q_\alpha), \tag{53}$$

where $r_{m,n}$ is defined by

$$r_{m,n}^\beta = \frac{\frac{\tau}{||p_m||_\alpha^\beta} p_m^\beta + \frac{1-\tau}{||p_n||_\alpha^\beta} p_n^\beta}{\frac{\tau}{||p_m||_\alpha^\beta} + \frac{1-\tau}{||p_n||_\alpha^\beta}}. \tag{54}$$

Note that, since $\mathbb{E}^\beta$ is convex, $r_{m,n} \in \mathbb{E}^\beta$ and so $r_{m,n} \in \mathbb{E}$. Also, using the non-negativity of divergence, (53) leads to

$$0 \leq \frac{1}{2}D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + \frac{1}{2}D_\lambda(p_{n,\alpha}, r_{m,n,\alpha}) \leq \frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) - D_\lambda(r_{m,n,\alpha}, q_\alpha). \tag{55}$$

Taking limit as $m, n \to \infty$, one can see that $\left[\frac{1}{2}D_\lambda(p_{m,\alpha}, q_\alpha) + \frac{1}{2}D_\lambda(p_{n,\alpha}, q_\alpha) - D_\lambda(r_{m,n,\alpha}, q_\alpha)\right] \to 0$ and hence $[D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) + D_\lambda(p_{n,\alpha}, r_{m,n,\alpha})] \to 0$. Thus, $D_\lambda(p_{m,\alpha}, r_{m,n,\alpha}) \to 0$ as $m, n \to \infty$ by non-negativity. This along with a generalization of Pinker's inequality for ϕ-divergence ([100], Theorem 1) gives

$$\lim_{m,n\to\infty} ||p_{m,\alpha} - r_{m,n,\alpha}||_T = 0, \tag{56}$$

whenever $\lambda(1+\lambda) > 0$ (which is true since $\beta\lambda > 0$); here $||\cdot||_T$ denotes the total variation norm. Now, by triangle inequality

$$||p_{m,\alpha} - p_{n,\alpha}||_T \leq ||p_{m,\alpha} - r_{m,n,\alpha}||_T + ||p_{n,\alpha} - r_{m,n,\alpha}||_T \to 0, \quad \text{as} \quad m,n \to \infty.$$

Thus, $\{p_{n,\alpha}\}$ is Cauchy in $L_1(\mu)$ and hence converges to some $g \in L_1(\mu)$, i.e.,

$$\lim_{n\to\infty} \int |p_{n,\alpha} - g| d\mu = 0, \tag{57}$$

and g is a probability density with respect to μ since each p_n is so. Also, (57) implies that $p_{n,\alpha} \to g$ in $[\mu]$-measure and hence $p_{n,\alpha}^{1/\alpha} \to g^{1/\alpha}$ in $L_\alpha(\mu)$ by an application of generalized dominated convergence theorem.

Next, as in the proof of ([16], Theorem 8), we can show that $||p_n||_\alpha$ is bounded and hence $||p_n||_\alpha \to c$ for some $c > 0$, possibly working with a subsequence if needed. Thus we have $p_n = ||p_n||_\alpha p_{n,\alpha}^{1/\alpha} \to cg^{1/\alpha}$ in $L_\alpha(\mu)$. However, since $\mathbb{E}^\beta$ is closed, we have $\mathbb{E}$ is closed and hence $cg^{1/\alpha} = p^*$ for some $p^* \in \mathbb{E}$. Further, since $\int gd\mu = 1$, we must have $c = ||p^*||_\alpha$ and hence $g = p^*_\alpha$. Since $p_n \to p^*$ and $p^* \in \mathbb{E}$, Proposition 5 implies that

$$\mathcal{RE}_{\alpha,\beta}(p^*, q) \leq \liminf_{n\to\infty} \mathcal{RE}_{\alpha,\beta}(p_n, q) = \inf_{p\in\mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q) \leq \mathcal{RE}_{\alpha,\beta}(p^*, q),$$

where the second equality follows by continuity of the function $f(u) = (\beta\lambda)^{-1}\log(sign(\beta\lambda)u + 1)$, definitions of p_n sequence and (13). Hence, we must have $\mathcal{RE}_{\alpha,\beta}(p^*, q) = \inf\limits_{p\in\mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q)$, i.e., p^* is a forward (α, β)-projection of q on $\mathbb{E}$.

The Case $\beta\lambda < 0$:

Note that, in this case, we must have $0 < \beta < \alpha$, since $\alpha > 0$. Then, using (29), we can see that

$$\begin{aligned}
\inf_{p\in\mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q) &= \frac{1}{\beta\lambda} \log\left[\sup_{p\in\mathbb{E}} \int \left(\frac{p}{||p||_\alpha}\right)^\beta \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} d\mu\right] \\
&= \frac{1}{\beta\lambda} \log\left[\sup_{h\in\widetilde{\mathbb{E}}} \int hg d\mu\right],
\end{aligned} \tag{58}$$

where $g = \left(\frac{q}{||q||_\alpha}\right)^{\alpha-\beta} \in L_{\frac{\alpha}{\alpha-\beta}}(\mu)$ and

$$\widetilde{\mathbb{E}} = \left\{ s\left(\frac{p}{||p||_\alpha}\right)^\beta : p \in \mathbb{E}, s \in [0,1] \right\} \subset L_{\alpha/\beta}(\mu).$$

Now, since $\mathbb{E}^\beta$ and hence $\mathbb{E}$ is closed, one can show that $\widetilde{E}$ is also closed; see, e.g., the proof of ([16], Theorem 8). Next, we will show that $\widetilde{E}$ is also convex. For take $s_1\left(\frac{p_1}{||p_1||_\alpha}\right)^\beta \in \widetilde{\mathbb{E}}$ and $s_0\left(\frac{p_0}{||p_0||_\alpha}\right)^\beta \in \widetilde{\mathbb{E}}$ for some $s_0, s_1 \in [0,1]$ and $p_0, p_1 \in \mathbb{E}$, and take any $\tau \in [0,1]$. Note that

$$\tau s_1 \left(\frac{p_1}{||p_1||_\alpha}\right)^\beta + (1-\tau)s_0\left(\frac{p_0}{||p_0||_\alpha}\right)^\beta = s_\tau\left(\frac{p_\tau}{||p_\tau||_\alpha}\right)^\beta,$$

where

$$p_\tau^\beta = \frac{\tau s_1 \left(\frac{p_1}{||p_1||_\alpha}\right)^\beta + (1-\tau)s_0\left(\frac{p_0}{||p_0||_\alpha}\right)^\beta}{\frac{\tau s_1}{||p_1||_\alpha^\beta} + \frac{(1-\tau)s_0}{||p_0||_\alpha^\beta}}, \quad \text{and} \quad s_\tau = \left[\frac{\tau s_1}{||p_1||_\alpha^\beta} + \frac{(1-\tau)s_0}{||p_0||_\alpha^\beta}\right] ||p_\tau||_\alpha^\beta.$$

However, by convexity of $\mathbb{E}^{\beta}$, $p_{\tau} \in \mathbb{E}$ and also $0 \le s_{\tau} \le 1$ by the extended Minkowski inequality (33). Therefore, $s_{\tau}\left(\frac{p_{\tau}}{||p_{\tau}||_{\alpha}}\right)^{\beta} \in \widetilde{\mathbb{E}}$ and hence $\widetilde{\mathbb{E}}$ is convex.

Finally, since $0 < \beta < \alpha$, $L_{\alpha/\beta}(\mu)$ is a reflexive Banach space and hence the closed and convex $\widetilde{\mathbb{E}} \subset L_{\alpha/\beta}(\mu)$ is also closed in the weak topology. So, the unit ball is compact in the weak topology by the Banach-Alaoglu theorem and hence its closed subset $\widetilde{\mathbb{E}}$ is also weakly compact. However, since g belongs to the dual space of $L_{\alpha/\beta}(\mu)$, the linear functional $h \mapsto \int hgd\mu$ is continuous in weak topology and also increasing in s. Hence its supremum over $\widetilde{\mathbb{E}}$ is attained at $s = 1$ and some $p^* \in \mathbb{E}$, which is the required forward (α, β)-projection. □

Before concluding this section, we will present one example of the forward (α, β)-projection onto a transformed-linear family of distributions.

Example 3 (An example of the forward (α, β)-projection). *Fix $\alpha > 0$, $\beta \in \mathbb{R}\backslash\{0, \alpha\}$ and $q \in L_{\alpha}(\mu)$ related to the measure Q. Consider measurable functions $f_i : \Omega \mapsto \mathbb{R}$ for $i \in I$, an index set, and the family of distributions*

$$\mathbb{L}_{\beta}^* = \left\{P \in \mathbb{S}^* : \int f_{\gamma} dP_{\beta} = 0\right\} \subset \mathbb{S}^*.$$

Let us denote the corresponding μ-density set by $\mathbb{L}_{\beta} = \left\{p = \frac{dP}{d\mu} : P \in \mathbb{L}_{\beta}^*\right\}$. We assume that, $\mathbb{L}_{\beta}^*$ is non-empty, every $P \in \mathbb{L}_{\beta}^*$ is absolute continuous with respect to μ and $\mathbb{L}_{\beta} \subset L_{\alpha}(\mu)$.

Then, p^* is the forward (α, β)-projection of q on $\mathbb{L}_{\beta}$ if and only if there exists a function g in the $L_1(Q_{\beta})$-closure of the linear space spanned by $\{f_i : i \in I\}$ and a subset $N \subset \Omega$ such that, for every $P \in \mathbb{L}_{\beta}^*$

$$\begin{cases} P(N) = 0 & \text{if } \alpha < \beta, \\ c\int_N q^{\alpha-\beta} dP_{\beta} \le \int_{\Omega\backslash N} g dP_{\beta} & \text{if } \alpha > \beta, \end{cases}$$

with $c = \frac{\int (p^*)^{\alpha} d\mu}{\int (p^*)^{\beta} q^{\alpha-\beta} d\mu}$ and p^* satisfies

$$\begin{aligned} & p^*(x)^{\alpha-\beta} = cq(x)^{\alpha-\beta} + g(x), \quad \text{if } x \notin N, \\ & p^*(x) = 0, \qquad\qquad \text{if } x \in N. \end{aligned}$$

The proof follows by extending the arguments of the proof of ([16], Theorem 11) and hence it is left as an exercise to the readers.

Remark 6. *Note that, at the special case $\beta = 1$, $\mathbb{L}_1^*$ is a linear family of distributions and the above example coincides with ([16], Theorem 11) on the forward projection of relative α-entropy on $\mathbb{L}_1^*$. However, it is still an open question to derive the forward (α, β)-projection on $\mathbb{L}_1^*$.*

5. Statistical Applications: The Minimum Relative Entropy Inference

5.1. The Reverse Projection and Parametric Estimation

As in the case of the forward projection of a relative entropy measure, we can also define the reverse projection by minimizing it with respect to the second argument over a convex set $\mathbb{E}$ keeping the first argument fixed. More formally, we use the following definition.

Definition 4 (Reverse (α, β)-Projection). *Fix $p \in L_{\alpha}(\mu)$ and let $\mathbb{E} \subset \mathbb{S}$ with $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$ for some $q \in \mathbb{E}$. Then, $q^* \in \mathbb{E}$ is called the reverse projection of the relative (α, β)-entropy or simply the reverse (α, β)-projection (or reverse LSD projection) of p on $\mathbb{E}$ if it satisfies the relation*

$$\mathcal{RE}_{\alpha,\beta}(p, q^*) = \inf_{q \in \mathbb{E}} \mathcal{RE}_{\alpha,\beta}(p, q). \tag{59}$$

We can get sufficient conditions for the existence and uniqueness of the reverse (α, β)-projection directly from Theorem 6 and the fact that $\mathcal{RE}_{\alpha,\beta}(p, q) = \mathcal{RE}_{\alpha,\alpha-\beta}(q, p)$; this is presented in the following theorem.

Theorem 7 (Existence and Uniqueness of Reverse (α, β)-Projection)**.** *Fix $\alpha > 0$ and $\beta \in \mathbb{R}$ with $\beta \neq \alpha$ and $p \in L_\alpha(\mu)$. Given any set $\mathbb{E} \subset \mathbb{S}$ for which $\mathbb{E}^{\alpha-\beta}$ is convex and closed and $\mathcal{RE}_{\alpha,\beta}(p, q) < \infty$ for some $q \in \mathbb{E}$, a reverse (α, β)-projection of p on $\mathbb{E}$ exists and is unique.*

The reverse projection is mostly used in statistical inference where we fix the first argument of a relative entropy measure (or divergence measure) at the empirical data distribution and minimize the relative entropy with respect to the model family of distributions in its second argument. The resulting estimator, commonly known as the minimum distance or minimum divergence estimator, yields the reverse projection of the observed data distribution on the family of model distributions with respect to the relative entropy or divergence under consideration. This approach was initially studied by [9–13] to obtain the popular maximum likelihood estimator as the reverse projection with respect to the relative entropy in (2). More recently, this approach has become widely popular, but with more general relative entropies or divergence measures, to obtain robust estimators against possible contamination in the observed data. Let us describe it more rigorously in the following for our relative (α, β)-entropy.

Suppose we have independent and identically distributed data $X_1, \ldots, X_n$ from a true distribution G having density g with respect to some common dominating measure μ. We model g by a parametric model family of μ-densities $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$, where it is assumed that both g and $f_{\boldsymbol{\theta}}$ have the same support independent of $\boldsymbol{\theta}$. Our objective is to infer about the unknown parameter $\boldsymbol{\theta}$. In minimum divergence inference, an estimator of $\boldsymbol{\theta}$ is obtained by minimizing the divergence measure between (an estimate of) g and $f_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta} \in \Theta$. Maji et al. [78] have considered the LSD (or equivalently the relative (α, β)-entropy) as the divergence under consideration and defined the corresponding minimum divergence functional at G, say $\boldsymbol{T}_{\alpha,\beta}(G)$, through the relation

$$\mathcal{RE}_{\alpha,\beta}\left(g, f_{\boldsymbol{T}_{\alpha,\beta}(G)}\right) = \min_{\boldsymbol{\theta} \in \Theta} \mathcal{RE}_{\alpha,\beta}(g, f_{\boldsymbol{\theta}}), \tag{60}$$

whenever the minimum exists. We will refer to $\boldsymbol{T}_{\alpha,\beta}(G)$ as the minimum relative (α, β)-entropy (MRE) functional, or the minimum LSD functional in the language of [78,79]. Note that, if $g \in \mathcal{F}$, i.e., $g = f_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \Theta$, then we must have $\boldsymbol{T}_{\alpha,\beta}(G) = \boldsymbol{\theta}_0$. If $g \notin \mathcal{F}$, we call $\boldsymbol{T}_{\alpha,\beta}(G)$ as the “best fitting parameter" value, since $f_{\boldsymbol{T}_{\alpha,\beta}(G)}$ is the closest model element to g in the LSD sense. In fact, for $g \notin \mathcal{F}$, $\boldsymbol{T}_{\alpha,\beta}(G)$ is nothing but the reverse (α, β)-projection of the true density g on the model family $\mathcal{F}$, which exists and is unique under the sufficient conditions of Theorem 7. Therefore, under identifiability of the model family $\mathcal{F}$ we get the existence and uniqueness of the MRE functional, which is presented in the following corollary. Although this estimator was first introduced by [78] in terms of the LSD, the results concerning the existence of the estimate were not provided.

Corollary 2 (Existence and Uniqueness of the MRE Functional)**.** *Consider the above parametric estimation problem with $g \in L_\alpha(\mu)$ and $\mathcal{F} \subset L_\alpha(\mu)$. Fix $\alpha > 0$ and $\beta \in \mathbb{R}$ with $\beta \neq \alpha$ and assume that the model family $\mathcal{F}$ is identifiable in $\boldsymbol{\theta}$.*

1. *Suppose $g = f_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \Theta$. Then the unique MRE functional is given by $\boldsymbol{T}_{\alpha,\beta}(G) = \boldsymbol{\theta}_0$.*
2. *Suppose $g \notin \mathcal{F}$. If $\mathcal{F}^{\alpha-\beta}$ is convex and closed and $\mathcal{RE}_{\alpha,\beta}(g, f_{\boldsymbol{\theta}}) < \infty$ for some $\boldsymbol{\theta} \in \Theta$, the MRE functional $\boldsymbol{T}_{\alpha,\beta}(G)$ exists and is unique.*

Further, under standard differentiability assumptions, we can obtain the estimating equation of the MRE functional $\boldsymbol{T}_{\alpha,\beta}(G)$ as given by

$$\left[\int f_{\boldsymbol{\theta}}^{\alpha}\boldsymbol{u}_{\boldsymbol{\theta}}d\mu\right]\left[\int f_{\boldsymbol{\theta}}^{\alpha-\beta}g^{\beta}d\mu\right]=\left[\int f_{\boldsymbol{\theta}}^{\alpha-\beta}g^{\beta}\boldsymbol{u}_{\boldsymbol{\theta}}d\mu\right]\left[\int f_{\boldsymbol{\theta}}^{\alpha}d\mu\right], \tag{61}$$

where $\boldsymbol{u}_{\boldsymbol{\theta}}(x)=\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(x)$. It is important to note that, at $\beta=\alpha=1$, the MRE functional $\boldsymbol{T}_{1,1}(G)$ coincides with the maximum likelihood functional since $\mathcal{RE}_{1,1}=\mathcal{RE}$, the KLD measure. Based on the estimating Equation (61), Maji et al. [78] extensively studied the theoretical robustness properties of the MRE functional against gross-error contamination in data through the higher order influence function analysis. The classical first order influence function was seen to be inadequate for this purpose; it becomes independent of β at the model but the real-life performance of the MRE functional critically depends on both α and β [78,79] as we will also see in Section 5.2.

In practice, however, the true data generating density is not known and so we need to use some empirical estimate in place of g and the resulting value of the MRE functional is called the minimum relative (α,β)-entropy estimator (MREE) or the minimum LSD estimator in the terminology of [78,79]. Note that, when the data are discrete and μ is the counting measure, one can use a simple estimate of g given by the relative frequencies $r_n(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i=x)$, where $I(A)$ is the indicator function of the event A; the corresponding MREE is then obtained by solving (61) with $g(x)$ replaced by $r_n(x)$ and integrals replaced by sums over the discrete support. Asymptotic properties of this MREE under discrete models are well-studied by [78,79] for the tuning parameters $\alpha\geq 1$ and $\beta\in\mathbb{R}$; the same line of argument can be used to extend them also for the cases $\alpha\in(0,1)$ in a straightforward manner.

However, in case of continuous data, there is no such simple estimator available to use in place of g unless $\beta=1$. When $\beta=1$, the estimating Equation (61) depends on g through the terms $\int f_{\boldsymbol{\theta}}^{\alpha-1}gd\mu=\int f_{\boldsymbol{\theta}}^{\alpha-1}dG$ and $\int f_{\boldsymbol{\theta}}^{\alpha-1}\boldsymbol{u}_{\boldsymbol{\theta}}gd\mu=\int f_{\boldsymbol{\theta}}^{\alpha-1}\boldsymbol{u}_{\boldsymbol{\theta}}dG$; so we can simply use the empirical distribution function G_n in place of G and solve the resulting equation to obtain the corresponding MREE. However, for $\beta\neq 1$, we must use a non-parametric kernel estimator g_n of g in (61) to obtain the MREE under continuous models; this leads to complications including bandwidth selection while deriving the asymptotics of the resulting MREE. One possible approach to avoid such complications is to use the smoothed model technique, which has been applied in [108] for the case of minimum ϕ-divergence estimators. Another alternative approach has been discussed in [109,110]. However, the detailed analyses of the MREE under the continuous model, in either of the above approaches, are yet to be studied so far.

5.2. Numerical Illustration: Binomial Model

Let us now present numerical illustrations under the common binomial model to study the finite sample performance of the MREEs. Along with the known properties of the MREE at $\alpha\geq 1$ (i.e., the minimum LSD estimators with $\tau\geq 0$ from [78,79]), here we will additionally explore their properties in case of $\alpha\in(0,1)$ and for the new divergences $\mathcal{RE}_{\beta}^{*}(P,Q)$ related to $\alpha=0$.

Suppose $X_1,\ldots,X_n$ are random observations from a true density g having support $\chi=\{0,1,2,\ldots,m\}$ for some positive integer m. We model g by the Binomial(m,θ) densities $f_\theta(x)=\binom{n}{x}\theta^x(1-\theta)^{m-x}$ for $x\in\chi$ and $\theta\in[0,1]$. Here an estimate $\widehat{g}$ of g is given by the relative frequency $\widehat{g}(x)=r_n(x)$. For any $\alpha>0$ and $\beta\in\mathbb{R}$, the relative (α,β)-entropy between $\widehat{g}$ and f_θ is given by

$$\begin{aligned}\mathcal{RE}_{\alpha,\beta}(\widehat{g},f_\theta) &= \frac{1}{\beta}\log\left[\sum_{x=0}^{m}\binom{n}{x}^{\alpha}\left(\frac{\theta}{1-\theta}\right)^{\alpha x}(1-\theta)^{m\alpha}\right]+\frac{1}{\alpha-\beta}\log\left[\sum_{x=0}^{m}r_n(x)^{\alpha}\right]\\ &\quad -\frac{\alpha}{\beta(\alpha-\beta)}\log\left[\sum_{x=0}^{m}\binom{n}{x}^{\alpha-\beta}\left(\frac{\theta}{1-\theta}\right)^{(\alpha-\beta)x}(1-\theta)^{m(\alpha-\beta)}r_n(x)^{\beta}\right],\end{aligned}$$

which can be minimized with respect to $\theta \in [0, 1]$ to obtain the corresponding MREE of θ. Note that, it is also the solution of the estimating Equation (61) with $g(x)$ replaced by the relative frequency $r_n(x)$. However, in this example, $u_\theta(x) = \frac{x - m\theta}{\theta(1-\theta)}$ and hence the MREE estimating equation simplifies to

$$\frac{\sum_{x=0}^{m} \binom{n}{x}^{\alpha} (x - m\theta) \left(\frac{\theta}{1-\theta}\right)^{\alpha x}}{\sum_{x=0}^{m} \binom{n}{x}^{\alpha} \left(\frac{\theta}{1-\theta}\right)^{\alpha x}} = \frac{\sum_{x=0}^{m} (x - m\theta) \binom{n}{x}^{\alpha-\beta} \left(\frac{\theta}{1-\theta}\right)^{(\alpha-\beta)x} r_n(x)^{\beta}}{\sum_{x=0}^{m} \binom{n}{x}^{\alpha-\beta} \left(\frac{\theta}{1-\theta}\right)^{(\alpha-\beta)x} r_n(x)^{\beta}}. \tag{62}$$

We can numerically solve the above estimating equation over $\theta \in [0, 1]$, or equivalently over the transformed parameter $p := \frac{\theta}{1-\theta} \in [0, \infty]$, to obtain the corresponding MREE (i.e., the minimum LSD estimator).

We simulate random sample of size n from a binomial population with true parameter $\theta_0 = 0.1$ with $m = 10$ and numerically compute the MREE. Repeating this exercise 1000 times, we can obtain an empirical estimate of the bias and the mean squared error (MSE) of the MREE of 10θ (since θ is very small in magnitude). Tables 1 and 2 present these values for sample sizes $n = 20, 50, 100$ and different values of tuning parameters $\alpha > 0$ and $\beta > 0$; their existences are guaranteed by Corollary 2. Note that the choice $\alpha = 1 = \beta$ gives the maximum likelihood estimator whereas $\beta = 1$ only yields the minimum LDPD estimator with parameter α. Next, in order to study the robustness, we contaminate 10% of each sample by random observations from a distant binomial distribution with parameters $\theta = 0.9$ and $m = 10$ and repeat the above simulation exercise; the resulting bias and MSE for the contaminated samples are given in Tables 3 and 4. Our observations from these tables can be summarized as follows.

- Under pure data with no contamination, the maximum likelihood estimator (the MREE at $\alpha = 1 = \beta$) has the least bias and MSE as expected, which further decrease as sample size increases.
- As we move away from $\alpha = 1$ and $\beta = 1$ in either direction, the MSEs of the corresponding MREEs under pure data increase slightly; but as long as the tuning parameters remain within a reasonable window of the $(1, 1)$ point and neither component is very close to zero, this loss in efficiency is not very significant.
- When α or β approaches zero, the MREEs become somewhat unstable generating comparatively larger MSE values. This is probably due to the presence of inliers under the discrete binomial model. Note that, the relative (α, β)-entropy measures with $\beta \leq 0$ are not finitely defined for the binomial model if there is just only one empty cell present in the data.
- Under contamination, the bias and MSE of the maximum likelihood estimator increase significantly but many MREEs remains stable. In particular, the MREEs with $\beta \geq \alpha$ and the MREEs with β close to zero are non-robust against data contamination. Many of the remaining members of the MREE family provide significantly improved robust estimators.
- In the entire simulation, the combination $(\alpha = 1, \beta = 0.7)$ appears to provide the most stable results. In Table 4, the best results are available along a tubular region which moves from the top left-hand to the bottom right-hand of the table subject to the conditions that $\alpha > \beta$ and none of them are very close to zero.
- Based on our numerical experiments, the optimum range of values of α, β providing the most robust minimum relative (α, β)-estimators are $\alpha = 0.9, 1$, $0.5 \leq \beta \leq 0.7$ and $1 < \alpha \leq 1.5$, $0.5 \leq \beta < 1$. Note that this range includes the estimators based on the logarithmic power divergence measure as well as the new LSD measures with $\alpha < 1$.
- Many of the MREEs, which belong to the optimum range mentioned in the last item and are close to the combination $\alpha = 1 = \beta$, generally also provide the best trade-off between efficiency under pure data and robustness under contaminated data.

In summary, many MREEs provide highly robust estimators under data contamination along with only a very small loss in efficiency under pure data. These numerical findings about the finite sample behavior of the MREEs under the binomial model and the corresponding optimum range of tuning parameters, for the subclass with $\alpha \geq 1$, are consistent with the findings of [78,79] who used a Poisson model. Additionally, our illustrations shed lights on the properties of the MREEs at $\alpha < 1$ as well and show that some MREEs in this range, e.g., at $\alpha = 0.9$ and $\beta = 0.5$, also yield optimum estimators in terms of the dual goal of high robustness and high efficiency.

Table 1. Bias of the MREE for different α, β and sample sizes n under pure data.

β	α									
	0.3	**0.5**	**0.7**	**0.9**	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 20$					
0.1	−0.210	−0.416	−0.397	−0.311	−0.277	−0.227	−0.130	0.021	0.024	0.122
0.3	2.218	−0.273	−0.229	−0.160	−0.141	−0.115	−0.096	−0.068	−0.036	0.034
0.5	−0.127	0.001	−0.125	−0.088	−0.082	−0.069	−0.058	−0.042	−0.032	−0.019
0.7	−0.093	−0.110	−0.010	−0.046	−0.044	−0.029	−0.023	−0.031	−0.023	−0.020
0.9	−0.066	−0.056	−0.028	−0.001	−0.015	−0.002	0.008	0.000	−0.006	−0.013
1	−0.041	−0.045	−0.017	0.005	−0.002	0.011	0.014	0.012	0.008	−0.003
1.3	−0.035	−0.013	0.023	0.036	0.030	0.039	0.088	0.039	0.035	0.021
1.5	−0.003	0.012	0.048	0.053	0.047	0.058	0.053	0.170	0.048	0.035
1.7	0.012	0.028	0.058	0.067	0.061	0.070	0.070	0.058	0.269	0.045
2	0.008	0.049	0.078	0.084	0.078	0.086	0.087	0.078	0.069	0.444
					$n = 50$					
0.1	−0.085	−0.301	−0.254	−0.183	−0.156	−0.106	−0.002	0.114	0.292	0.245
0.3	1.829	−0.176	−0.150	−0.078	−0.066	−0.042	−0.045	−0.014	0.005	0.030
0.5	−0.056	0.099	−0.054	−0.037	−0.033	−0.026	−0.019	−0.009	−0.007	−0.005
0.7	−0.009	−0.059	0.035	−0.012	−0.013	−0.005	−0.002	−0.009	−0.002	0.006
0.9	−0.031	−0.031	−0.009	0.012	0.002	0.013	0.021	0.015	0.008	0.004
1	0.014	−0.023	0.000	0.011	0.009	0.019	0.022	0.020	0.018	0.004
1.3	0.002	−0.004	0.022	0.034	0.027	0.030	0.084	0.034	0.035	0.028
1.5	0.009	0.023	0.038	0.044	0.037	0.042	0.034	0.174	0.040	0.032
1.7	0.028	0.029	0.049	0.054	0.047	0.050	0.047	0.036	0.277	0.039
2	0.040	0.051	0.065	0.068	0.059	0.063	0.060	0.051	0.041	0.464
					$n = 100$					
0.1	−0.028	−0.216	−0.175	−0.113	−0.103	−0.063	0.036	0.169	0.452	0.349
0.3	1.874	−0.135	−0.125	−0.052	−0.044	−0.022	−0.038	−0.023	0.009	0.024
0.5	−0.002	0.146	−0.034	−0.026	−0.025	−0.021	−0.019	-0.001	−0.008	−0.009
0.7	0.000	−0.042	0.045	−0.009	−0.013	−0.009	0.000	−0.009	−0.008	−0.001
0.9	0.007	−0.025	−0.015	0.001	−0.004	0.005	0.009	0.013	−0.001	−0.003
1	0.014	−0.010	−0.007	−0.001	−0.001	0.005	0.009	0.014	0.010	0.009
1.3	0.036	0.010	0.006	0.015	0.010	0.010	0.065	0.012	0.019	0.014
1.5	0.041	0.023	0.018	0.022	0.017	0.018	0.006	0.158	0.016	0.015
1.7	0.052	0.027	0.028	0.032	0.024	0.025	0.016	0.009	0.267	0.019
2	0.056	0.043	0.042	0.043	0.033	0.034	0.023	0.020	0.013	0.454

Table 2. MSE of the MREE for different α, β and sample sizes n under pure data.

β	α									
	0.3	**0.5**	**0.7**	**0.9**	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 20$					
0.1	0.347	0.251	0.222	0.145	0.122	0.106	0.098	0.242	0.206	0.240
0.3	7.506	0.147	0.100	0.069	0.063	0.059	0.059	0.062	0.098	0.169
0.5	0.238	0.076	0.067	0.051	0.049	0.047	0.050	0.055	0.064	0.101
0.7	0.177	0.091	0.056	0.045	0.044	0.043	0.045	0.055	0.056	0.071
0.9	0.163	0.085	0.061	0.045	0.042	0.043	0.047	0.053	0.058	0.064
1	0.171	0.085	0.064	0.045	0.042	0.045	0.048	0.053	0.058	0.063
1.3	0.148	0.082	0.065	0.052	0.046	0.046	0.061	0.055	0.058	0.065
1.5	0.146	0.085	0.069	0.056	0.050	0.050	0.051	0.087	0.061	0.065
1.7	0.150	0.085	0.070	0.060	0.053	0.055	0.055	0.056	0.134	0.066
2	0.132	0.091	0.076	0.065	0.059	0.060	0.060	0.060	0.061	0.265
					$n = 50$					
0.1	0.334	0.170	0.118	0.066	0.044	0.037	0.067	0.195	0.401	0.275
0.3	5.050	0.093	0.051	0.026	0.021	0.020	0.024	0.027	0.035	0.050
0.5	0.196	0.059	0.030	0.018	0.017	0.018	0.021	0.026	0.030	0.037
0.7	0.191	0.053	0.031	0.018	0.016	0.017	0.023	0.025	0.028	0.035
0.9	0.131	0.050	0.029	0.019	0.016	0.018	0.022	0.025	0.028	0.029
1	0.154	0.044	0.031	0.018	0.017	0.020	0.022	0.024	0.027	0.031
1.3	0.112	0.046	0.029	0.023	0.018	0.018	0.033	0.028	0.029	0.031
1.5	0.108	0.049	0.033	0.024	0.020	0.022	0.022	0.059	0.031	0.031
1.7	0.119	0.049	0.036	0.026	0.022	0.023	0.025	0.025	0.108	0.033
2	0.108	0.053	0.040	0.030	0.025	0.026	0.028	0.029	0.028	0.249
					$n = 100$					
0.1	0.295	0.139	0.085	0.038	0.022	0.022	0.068	0.201	0.583	0.403
0.3	4.770	0.075	0.039	0.016	0.011	0.011	0.017	0.019	0.023	0.035
0.5	0.189	0.061	0.022	0.011	0.009	0.012	0.016	0.017	0.022	0.023
0.7	0.141	0.038	0.024	0.010	0.009	0.010	0.014	0.017	0.018	0.021
0.9	0.123	0.035	0.021	0.011	0.009	0.011	0.012	0.015	0.019	0.021
1	0.122	0.036	0.019	0.010	0.009	0.011	0.013	0.016	0.017	0.020
1.3	0.114	0.035	0.019	0.012	0.009	0.010	0.021	0.016	0.017	0.019
1.5	0.105	0.037	0.019	0.012	0.010	0.011	0.012	0.045	0.017	0.020
1.7	0.097	0.034	0.021	0.014	0.011	0.012	0.014	0.014	0.092	0.020
2	0.088	0.039	0.023	0.016	0.012	0.013	0.013	0.016	0.016	0.227

Table 3. Bias of the MREE for different α, β and sample sizes n under contaminated data.

β	α									
	0.3	**0.5**	**0.7**	0.9	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 20$					
0.1	−0.104	−0.382	−0.340	−0.243	−0.131	−0.071	0.090	0.188	0.295	0.379
0.3	3.287	−0.157	−0.187	−0.135	−0.113	−0.091	−0.045	0.013	0.107	0.237
0.5	2.691	1.483	−0.024	−0.067	−0.069	−0.043	−0.031	−0.010	−0.003	0.051
0.7	3.004	2.546	1.168	0.036	−0.017	−0.008	0.003	0.006	0.005	0.010
0.9	3.133	2.889	2.319	0.917	0.222	0.058	0.019	0.023	0.017	0.022
1	3.183	2.986	2.558	1.619	0.805	0.214	0.039	0.030	0.031	0.019
1.3	3.239	3.121	2.902	2.550	2.262	1.872	0.613	0.077	0.049	0.040
1.5	3.255	3.170	3.012	2.775	2.606	2.396	1.676	0.571	0.069	0.051
1.7	3.271	3.194	3.071	2.903	2.790	2.661	2.256	1.489	0.578	0.057
2	3.289	3.216	3.122	3.012	2.942	2.865	2.649	2.305	1.690	0.682

Table 3. *Cont.*

β	α									
	0.3	**0.5**	**0.7**	**0.9**	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 50$					
0.1	0.384	−0.170	−0.189	−0.132	−0.054	0.024	0.104	0.171	0.261	0.382
0.3	3.549	0.000	−0.122	−0.086	−0.077	−0.053	−0.023	0.029	0.054	0.118
0.5	2.875	1.771	0.040	−0.048	−0.048	−0.029	−0.013	−0.015	−0.017	0.003
0.7	3.091	2.698	1.294	0.048	−0.010	−0.014	−0.001	0.004	0.001	−0.005
0.9	3.205	2.945	2.379	0.939	0.226	0.045	0.009	0.013	0.012	0.013
1	3.240	3.011	2.612	1.609	0.793	0.196	0.018	0.014	0.021	0.012
1.3	3.316	3.171	2.925	2.548	2.239	1.819	0.554	0.034	0.020	0.020
1.5	3.346	3.223	3.034	2.780	2.596	2.363	1.589	0.502	0.035	0.022
1.7	3.362	3.254	3.100	2.916	2.791	2.643	2.199	1.383	0.518	0.025
2	3.373	3.281	3.162	3.035	2.955	2.865	2.622	2.236	1.575	0.650
					$n = 100$					
0.1	0.610	−0.138	−0.105	−0.031	0.002	0.040	0.117	0.184	0.270	0.381
0.3	3.906	0.136	−0.071	−0.050	−0.052	−0.028	−0.028	−0.008	0.023	0.066
0.5	2.927	1.934	0.101	−0.034	−0.027	−0.016	0.006	0.000	−0.003	−0.008
0.7	3.122	2.761	1.348	0.066	0.004	−0.007	0.007	0.011	0.012	0.000
0.9	3.241	2.955	2.406	0.958	0.238	0.047	0.004	0.014	0.022	0.017
1	3.289	3.045	2.651	1.622	0.798	0.202	0.010	0.011	0.016	0.023
1.3	3.362	3.204	2.944	2.567	2.245	1.812	0.533	0.028	0.015	0.022
1.5	3.384	3.269	3.058	2.802	2.610	2.369	1.567	0.485	0.027	0.018
1.7	3.405	3.305	3.133	2.940	2.811	2.658	2.196	1.357	0.504	0.018
2	3.421	3.327	3.204	3.065	2.980	2.886	2.633	2.234	1.541	0.637

Table 4. MSE of the MREE for different α, β and sample sizes n under contaminated data.

β	α									
	0.3	**0.5**	**0.7**	**0.9**	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 20$					
0.1	0.403	0.248	0.465	0.576	1.025	1.093	1.613	1.565	1.626	1.591
0.3	12.595	0.142	0.103	0.075	0.192	0.188	0.362	0.590	1.016	1.537
0.5	7.443	2.268	0.088	0.062	0.058	0.059	0.065	0.189	0.241	0.527
0.7	9.209	6.645	1.410	0.069	0.056	0.058	0.063	0.068	0.119	0.208
0.9	9.982	8.493	5.512	0.882	0.119	0.068	0.065	0.069	0.075	0.090
1	10.292	9.072	6.672	2.692	0.693	0.117	0.068	0.070	0.076	0.087
1.3	10.664	9.916	8.574	6.641	5.240	3.610	0.430	0.079	0.079	0.089
1.5	10.778	10.229	9.238	7.850	6.940	5.883	2.917	0.389	0.079	0.087
1.7	10.884	10.379	9.599	8.582	7.942	7.234	5.235	2.326	0.403	0.087
2	11.004	10.515	9.915	9.233	8.814	8.369	7.177	5.472	2.998	0.547
					$n = 50$					
0.1	1.552	0.815	0.741	0.703	0.966	1.190	1.129	1.224	1.165	1.210
0.3	14.969	0.105	0.047	0.030	0.078	0.075	0.280	0.559	0.566	0.881
0.5	8.345	3.190	0.049	0.025	0.021	0.022	0.025	0.029	0.035	0.184
0.7	9.634	7.335	1.694	0.031	0.020	0.022	0.027	0.029	0.033	0.039
0.9	10.353	8.723	5.712	0.898	0.077	0.027	0.028	0.030	0.032	0.039
1	10.578	9.126	6.871	2.619	0.645	0.067	0.027	0.030	0.033	0.039
1.3	11.069	10.129	8.608	6.548	5.064	3.359	0.329	0.033	0.034	0.038
1.5	11.263	10.457	9.268	7.787	6.801	5.648	2.576	0.279	0.032	0.038
1.7	11.371	10.655	9.676	8.567	7.854	7.051	4.908	1.968	0.298	0.037
2	11.449	10.833	10.060	9.275	8.793	8.276	6.947	5.079	2.560	0.461

Table 4. *Cont.*

β	α									
	0.3	**0.5**	**0.7**	**0.9**	**1**	**1.1**	**1.3**	**1.5**	**1.7**	**2**
					$n = 100$					
0.1	2.102	0.399	0.808	0.945	0.924	0.929	0.891	1.012	1.233	1.120
0.3	17.185	0.141	0.033	0.018	0.013	0.014	0.018	0.142	0.258	0.453
0.5	8.624	3.768	0.056	0.015	0.011	0.015	0.017	0.018	0.022	0.028
0.7	9.809	7.646	1.828	0.024	0.011	0.013	0.018	0.019	0.020	0.023
0.9	10.559	8.764	5.812	0.927	0.070	0.018	0.017	0.020	0.021	0.023
1	10.870	9.312	7.058	2.648	0.645	0.057	0.017	0.019	0.021	0.023
1.3	11.342	10.306	8.691	6.619	5.068	3.312	0.297	0.020	0.020	0.023
1.5	11.494	10.727	9.379	7.880	6.845	5.646	2.484	0.251	0.021	0.021
1.7	11.632	10.960	9.848	8.675	7.932	7.101	4.866	1.873	0.272	0.022
2	11.739	11.102	10.297	9.422	8.910	8.363	6.973	5.040	2.420	0.430

5.3. Application to Testing Statistical Hypothesis

We end the paper with a very brief indication on the potential of the relative (α, β)-entropy or the LSD measure in statistical hypothesis testing problems. The minimum possible value of the relative entropy or divergence measure between the data and the null distribution indicates the amount of departure from null and hence can be used to develop a statistical testing procedure.

Consider the parametric estimation set-up as in Section 5.1 with $g \in \mathcal{F}$ and fix a parameter value $\boldsymbol{\theta}_0 \in \Theta$. Suppose we want to test the simple null hypothesis in the one sample case given by

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{against} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Maji et al. [78] have developed the LSD-based test statistics for the above testing problem as given by

$$T^{(1)}_{n,\alpha,\beta} = 2n\mathcal{RE}_{\alpha,\beta}(f_{\widehat{\boldsymbol{\theta}}_{\alpha,\beta}}, f_{\boldsymbol{\theta}_0}), \tag{63}$$

where $\widehat{\boldsymbol{\theta}}_{\alpha,\beta}$ is the MREE with parameters α and β. [78,79] have also developed the LSD-based test for a simple two-sample problem where two independent samples of sizes n_1 and n_2 are given from true densities $f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2} \in \mathcal{F}$, respectively and we want to test for the homogeneity of the two samples trough the hypothesis

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \quad \text{against} \quad H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2.$$

The proposed test statistics for this two-sample problem has the form

$$T^{(2)}_{n,\alpha,\beta} = \frac{2n_1 n_2}{n_1 + n_2}\mathcal{RE}_{\alpha,\beta}(f_{{}^{(1)}\widehat{\boldsymbol{\theta}}_{\alpha,\beta}}, f_{{}^{(2)}\widehat{\boldsymbol{\theta}}_{\alpha,\beta}}), \tag{64}$$

where ${}^{(1)}\widehat{\theta}_{\alpha,\beta}$ and ${}^{(2)}\widehat{\theta}_{\alpha,\beta}$ are the MREEs of θ_1 and θ_2, respectively, obtained from the two samples separately Note that, at $\alpha = \beta = 1$, both the test statistics in (63) and (64) become asymptotically equivalent to the corresponding likelihood ratio tests under the respective null hypothesis. Maji et al. [78,79] have studied the asymptotic properties of these two tests, which have asymptotic null distributions as linear combinations of chi-square distributions. They have also numerically illustrated the benefits of these LSD or relative (α, β)-entropy-based tests, although with tuning parameters $\alpha \geq 1$ only, to achieve robust inference against possible contamination in the sample data.

The same approach can also be used to develop robust tests for more complex hypothesis testing problems based on the relative (α, β)-entropy or the LSD measures, now with parameters $\alpha > 0$, and also using the new divergences $\mathcal{RE}^*_\beta(\cdot, \cdot)$. For example, consider the above one sample set-up and a subset $\Theta_0 \subset \Theta$ and let we are interested in testing the composite hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{against} \quad H_1 : \boldsymbol{\theta} \notin \Theta_0.$$

with similar motivation from (63) and (64), we can construct relative entropy or LSD-based test statistics for testing the above composite hypothesis as given by

$$\widetilde{T_{n,\alpha,\beta}}^{(1)} = 2n\mathcal{RE}_{\alpha,\beta}(f_{\widehat{\boldsymbol{\theta}}_{\alpha,\beta}}, f_{\widetilde{\boldsymbol{\theta}}_{\alpha,\beta}}), \tag{65}$$

where $\widetilde{\boldsymbol{\theta}}_{\alpha,\beta}$ is the restricted MREE with parameters α and β obtained by minimizing the relative entropy over $\boldsymbol{\theta} \in \Theta_0$ and $\widehat{\boldsymbol{\theta}}_{\alpha,\beta}$ is the corresponding unrestricted MREE obtained by minimizing over $\boldsymbol{\theta} \in \Theta$. It will surely be of significant interest to study the asymptotic and robustness properties of this relative entropy-based test for the above composite hypothesis under one sample or even more general hypotheses with two or more samples. However, considering the length of the present paper, which is primarily focused on the geometric properties of entropies and relative entropies, we have deferred the detailed analyses of such MREE-based hypothesis testing procedures in a future report.

6. Conclusions

We have explored the geometric properties of the LSD measures through a new information theoretic formulation when we develop this divergence measure as a natural extension of the relative α-entropy; we refer to it as the two-parameter relative (α, β)-entropy. It is shown to be always lower semicontinuous in both the arguments, but is continuous in its first argument only if $\alpha > \beta > 0$. We also proved that the relative (α, β)-entropy is quasi-convex in both its arguments after a suitable (different) transformation of the domain space and derive an extended Pythagorean relation under these transformations. Along with the study of its forward and reverse projections, statistical applications are also discussed.

It is worthwhile to note that the information theoretic divergences can also be used to define new measures of robustness and efficiency of a parameter estimate; one can then obtain the optimum robust estimator, along Hampel's infinitesimal principle, to achieve the best trade-off between these divergence-based summary measures [111–113]. In particular, the LDPD measure, a prominent member of our LSD or relative (α, β)-entropy family, has been used by [113] who have illustrated important theoretical properties including different types of equivariance of the resulting optimum estimators besides their strong robustness properties. A similar approach can also be used with our general relative (α, β)-entropies to develop estimators with enhanced optimality properties, establishing a better robustness-efficiency trade-off.

The present work opens up several interesting problems to be solved in future research as already noted throughout the paper. In particular, we recall that the relative α-entropy has an interpretation from the problem of guessing under source uncertainty [17,71]. As an extension of relative α-entropy, a similar information theoretic interpretation of the relative (α, β)-entropy (i.e., the LSD) is expected and its proper interpretation will be a useful development. Additionally, we have obtained a new extension of the Renyi entropy as a by-product and detailed study of this new entropy measure and its potential applications may lead to a new aspect of the mathematical information theory. Also, statistical applications of these measures need to be studied thoroughly specially for the continuous models, where the complications of a kernel density estimator is unavoidable, and for testing complex composite hypotheses from one or more samples. We hope to pursue some of these interesting extensions in future.

Author Contributions: Conceptualization, A.B. and A.G.; Methodology, A.B. and A.G.; Coding and Numerical Work, A.G.; Validation, A.G.; Formal Analysis, A.G. and A.B.; Investigation, A.G. and A.B.

Funding: The research of the first author is funded by the INSPIRE Faculty Research Grant from the Department of Science and Technology, Government of India.

Acknowledgments: The authors wish to thank four anonymous referees whose comments have led to a significantly improved version of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KLD	Kullback-Leibler Divergence
LDPD	Logarithmic Density Power Divergence
LSD	Logarithmic Super Divergence
GRE	Generalized Renyi Entropy
MRE	Minimum Relative (α, β)-entropy
MREE	Minimum Relative (α, β)-entropy Estimator
MSE	Mean Squared Error

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Shannon, C.E. Communication in the presence of noise. *Proc. IRE* **1949**, *37*, 10–21.
3. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, USA, 1949.
4. Khinchin, A.I. The entropy concept in probability theory. *Uspekhi Matematicheskikh Nauk* **1953**, *8*, 3–20.
5. Khinchin, A.I. On the fundamental theorems of information theory. *Uspekhi Matematicheskikh Nauk* **1956**, *11*, 17–75.
6. Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover Publications: New York, NY, USA, 1957.
7. Kolmogorov, A.N. *Foundations of the Theory of Probability*; Chelsea Publishing Co.: New York, NY, USA, 1950.
8. Kolmogorov, A.N. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Inf. Theory* **1956**, *IT-2*, 102–108.
9. Kullback, S. An application of information theory to multivariate analysis. *Ann. Math. Stat.* **1952**, *23*, 88–102.
10. Kullback, S. A note on information theory. *J. Appl. Phys.* **1953**, *24*, 106–107.
11. Kullback, S. Certain inequalities in information theory and the Cramer-Rao inequality. *Ann. Math. Stat.* **1954**, *25*, 745–751.
12. Kullback, S. An application of information theory to multivariate analysis II. *Ann. Math. Stat.* **1956**, *27*, 122–145.
13. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
14. Rosenkrantz, R.D. *E T Jaynes: Papers on Probability, Statistics and Statistical Physics*; Springer Science and Business Media: New York, NY, USA, 1983.
15. Van Campenhout, J.M.; Cover, T.M. Maximum entropy and conditional probability. *IEEE Trans. Inf. Theory* **1981**, *27*, 483–489.
16. Kumar, M.A.; Sundaresan, R. Minimization Problems Based on Relative α-Entropy I: Forward Projection. *IEEE Trans. Inf. Theory* **2015**, *61*, 5063–5080.
17. Sundaresan, R. Guessing under source uncertainty. *Proc. IEEE Trans. Inf. Theory* **2007**, *53*, 269–287.
18. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.
19. Csiszár, I. Sanov property, generalized I -projection, and a conditional limit theorem. *Ann. Probab.* **1984**, *12*, 768–793.
20. Csiszár, I.; Shields, P. *Information Theory and Statistics: A Tutorial*; NOW Publishers: Hanover, NH, USA, 2004.
21. Csiszár, I.; Tusnady, G. Information geometry and alternating minimization procedures. *Stat. Decis.* **1984**, *1*, 205–237.
22. Amari, S.I.; Karakida, R.; Oizumi, M. Information Geometry Connecting Wasserstein Distance and Kullback-Leibler Divergence via the Entropy-Relaxed Transportation Problem. *arXiv* **2017**, arXiv:1709.10219.
23. Costa, S.I.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. *Discret. Appl. Math.* **2015**, *197*, 59–69.

24. Nielsen, F.; Sun, K. Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures. *IEEE Signal Process. Lett.* **2016**, *23*, 1543–1546.
25. Amari, S.I.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci. Tech. Sci.* **2010**, *58*, 183–195.
26. Contreras-Reyes, J.E.; Arellano-Valle, R.B. Kullback-Leibler divergence measure for multivariate skew-normal distributions. *Entropy* **2012**, *14*, 1606–1626.
27. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya Centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
28. Pinski, F.J.; Simpson, G.; Stuart, A.M.; Weber, H. Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *SIAM J. Math. Anal.* **2015**, *47*, 4091–4122.
29. Attouch, H.; Bolte, J.; Redont, P.; Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.* **2010**, *35*, 438–457.
30. Eliazar, I.; Sokolov, I.M. Maximization of statistical heterogeneity: From Shannon's entropy to Gini's index. *Phys. A Stat. Mech. Appl.* **2010**, *389*, 3023–3038.
31. Monthus, C. Non-equilibrium steady states: Maximization of the Shannon entropy associated with the distribution of dynamical trajectories in the presence of constraints. *J. Stat. Mech. Theory Exp.* **2011**, *2011*, P03008.
32. Bafroui, H.H.; Ohadi, A. Application of wavelet energy and Shannon entropy for feature extraction in gearbox fault detection under varying speed conditions. *Neurocomputing* **2014**, *133*, 437–445.
33. Batty, M. Space, Scale, and Scaling in Entropy Maximizing. *Geogr. Anal.* **2010**, *42*, 395–421.
34. Oikonomou, T.; Bagci, G.B. Entropy Maximization with Linear Constraints: The Uniqueness of the Shannon Entropy. *arXiv* **2018**, arXiv:1803.02556.
35. Hoang, D.T.; Song, J.; Periwal, V.; Jo, J. Maximizing weighted Shannon entropy for network inference with little data. *arXiv* **2017**, arXiv:1705.06384.
36. Sriraman, T.; Chakrabarti, B.; Trombettoni, A.; Muruganandam, P. Characteristic features of the Shannon information entropy of dipolar Bose-Einstein condensates. *J. Chem. Phys.* **2017**, *147*, 044304.
37. Sun, M.; Li, Y.; Gemmeke, J.F.; Zhang, X. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence. *IEEE Trans. Audio Speech Lang. Process.* **2015**, *23*, 1233–1242.
38. Garcia-Fernandez, A.F.; Vo, B.N. Derivation of the PHD and CPHD Filters Based on Direct Kullback-Leibler Divergence Minimization. *IEEE Trans. Signal Process.* **2015**, *63*, 5812–5820.
39. Giantomassi, A.; Ferracuti, F.; Iarlori, S.; Ippoliti, G.; Longhi, S. Electric motor fault detection and diagnosis by kernel density estimation and Kullback-Leibler divergence based on stator current measurements. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1770–1780.
40. Harmouche, J.; Delpha, C.; Diallo, D.; Le Bihan, Y. Statistical approach for nondestructive incipient crack detection and characterization using Kullback-Leibler divergence. *IEEE Trans. Reliab.* **2016**, *65*, 1360–1368.
41. Hua, X.; Cheng, Y.; Wang, H.; Qin, Y.; Li, Y.; Zhang, W. Matrix CFAR detectors based on symmetrized Kullback-Leibler and total Kullback-Leibler divergences. *Digit. Signal Process.* **2017**, *69*, 106–116.
42. Ferracuti, F.; Giantomassi, A.; Iarlori, S.; Ippoliti, G.; Longhi, S. Electric motor defects diagnosis based on kernel density estimation and Kullback-Leibler divergence in quality control scenario. *Eng. Appl. Artif. Intell.* **2015**, *44*, 25–32.
43. Matthews, A.G.D.G.; Hensman, J.; Turner, R.; Ghahramani, Z. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *J. Mach. Learn. Res.* **2016**, *51*, 231–239.
44. Arikan, E. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Inf. Theory* **1996**, *42*, 99–105.
45. Campbell, L.L. A coding theorem and Renyi's entropy. *Inf. Control* **1965**, *8*, 423–429.
46. Renyi, A. On measures of entropy and information. In *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability I*; University of California: Berkeley, CA, USA, 1961; pp. 547–561.
47. Wei, B.B. Relations between heat exchange and Rényi divergences. *Phys. Rev. E* **2018**, *97*, 042107.
48. Kumar, M.A.; Sason, I. On projections of the Rényi divergence on generalized convex sets. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016.

49. Sadeghpour, M.; Baratpour, S.; Habibirad, A. Exponentiality test based on Renyi distance between equilibrium distributions. *Commun. Stat.-Simul. Comput.* **2017**, doi:10.1080/03610918.2017.1366514.
50. Markel, D.; El Naqa, I.I. PD-0351: Development of a novel regmentation framework using the Jensen Renyi divergence for adaptive radiotherapy. *Radiother. Oncol.* **2014**, *111*, S134.
51. Bai, S.; Lepoint, T.; Roux-Langlois, A.; Sakzad, A.; Stehlé, D.; Steinfeld, R. Improved security proofs in lattice-based cryptography: Using the Rényi divergence rather than the statistical distance. *J. Cryptol.* **2018**, *31*, 610–640.
52. Dong, X. The gravity dual of Rényi entropy. *Nat. Commun.* **2016**, *7*, 12472.
53. Kusuki, Y.; Takayanagi, T. Renyi entropy for local quenches in 2D CFT from numerical conformal blocks. *J. High Energy Phys.* **2018**, *2018*, 115.
54. Kumbhakar, M.; Ghoshal, K. One-Dimensional velocity distribution in open channels using Renyi entropy. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 949–959.
55. Xing, H.J.; Wang, X.Z. Selective ensemble of SVDDs with Renyi entropy based diversity measure. *Pattern Recog.* **2017**, *61*, 185–196.
56. Nie, F.; Zhang, P.; Li, J.; Tu, T. An Image Segmentation Method Based on Renyi Relative Entropy and Gaussian Distribution. *Recent Patents Comput. Sci.* **2017**, *10*, 122–130.
57. Ben Bassat, M. f-entropies, probability of error, and feature selection. *Inf. Control* **1978**, *39*, 277–292.
58. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys* **1988**, *52*, 479–487.
59. Kumar, S.; Ram, G.; Gupta, V. Axioms for (α, β, γ)-entropy of a generalized probability scheme. *J. Appl. Math. Stat. Inf.* **2013**, *9*, 95–106.
60. Kumar, S.; Ram, G. A generalization of the Havrda-Charvat and Tsallis entropy and its axiomatic characterization. *Abstr. Appl. Anal.* **2014**, *2014*, 505184.
61. Tsallis, C.; Brigatti, E. Nonextensive statistical mechanics: A brief introduction. *Contin. Mech. Thermodyn.* **2004**, *16*, 223–235.
62. Rajesh, G.; Sunoj, S.M. Some properties of cumulative Tsallis entropy of order α. *Stat. Pap.* **2016**, doi:10.1007/s00362-016-0855-7.
63. Singh, V.P. *Introduction to Tsallis Entropy Theory in Water Engineering*; CRC Press: Boca Raton, FL, USA, 2016.
64. Pavlos, G.P.; Karakatsanis, L.P.; Iliopoulos, A.C.; Pavlos, E.G.; Tsonis, A.A. Nonextensive Statistical Mechanics: Overview of Theory and Applications in Seismogenesis, Climate, and Space Plasma. In *Advances in Nonlinear Geosciences*; Tsonis, A., Ed.; Springer: Cham, Switzerland, 2018; pp. 465–495.
65. Jamaati, M.; Mehri, A. Text mining by Tsallis entropy. *Phys. A Stat. Mech. Appl.* **2018**, *490*, 1368–1376.
66. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011.
67. Leise, F.; Vajda, I. On divergence and information in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.
68. Pardo, L. *Statistical Inference Based on Divergences*; CRC/Chapman-Hall: London, UK, 2006.
69. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer: Boston, MA, USA, 1989.
70. Stummer, W.; Vajda, I. On divergences of finite measures and their applicability in statistics and information theory. *Statistics* **2010**, *44*, 169–187.
71. Sundaresan, R. A measure of discrimination and its geometric properties. In Proceedings of the IEEE International Symposium on Information Theory, Lausanne, Switzerland, 30 June–5 July 2002.
72. Lutwak, E.; Yang, D.; Zhang, G. Cramear-Rao and moment-entropy inequalities for Renyi entropy and generalized Fisher information. *IEEE Trans. Inf. Theory* **2005**, *51*, 473–478.
73. Kumar, M.A.; Sundaresan, R. Minimization Problems Based on Relative α-Entropy II: Reverse Projection. *IEEE Trans. Infor. Theory* **2015**, *61*, 5081–5095.
74. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873.
75. Windham, M. Robustifying model fitting. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 599–609.
76. Fujisawa, H. Normalized estimating equation for robust parameter estimation. *Elect. J. Stat.* **2013**, *7*, 1587–1606.
77. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.

78. Maji, A.; Ghosh, A.; Basu, A. The Logarithmic Super Divergence and its use in Statistical Inference. *arXiv* **2014**, arXiv:1407.3961.
79. Maji, A.; Ghosh, A.; Basu, A. The Logarithmic Super Divergence and Asymptotic Inference Properties. *AStA Adv. Stat. Anal.* **2016**, *100*, 99–131.
80. Maji, A.; Chakraborty, S.; Basu, A. Statistical Inference Based on the Logarithmic Power Divergence. *Rashi* **2017**, *2*, 39–51.
81. Lutz, E. Anomalous diffusion and Tsallis statistics in an optical lattice. *Phys. Rev. A* **2003**, *67*, 051402.
82. Douglas, P.; Bergamini, S.; Renzoni, F. Tunable Tsallis Distributions in Dissipative Optical Lattices. *Phys. Rev. Lett.* **2006**, *96*, 110601.
83. Burlaga, L.F.; Viñas, A.F. Triangle for the entropic index q of non-extensive statistical mechanics observed by Voyager 1 in the distant heliosphere. *Phys. A Stat. Mech. Appl.* **2005**, *356*, 375.
84. Liu, B.; Goree, J. Superdiffusion and Non-Gaussian Statistics in a Driven-Dissipative 2D Dusty Plasma. *Phys. Rev. Lett.* **2008**, *100*, 055003.
85. Pickup, R.; Cywinski, R.; Pappas, C.; Farago, B.; Fouquet, P. Generalized Spin-Glass Relaxation. *Phys. Rev. Lett.* **2009**, *102*, 097202.
86. Devoe, R. Power-Law Distributions for a Trapped Ion Interacting with a Classical Buffer Gas. *Phys. Rev. Lett.* **2009**, *102*, 063001.
87. Khachatryan, V.; Sirunyan, A.; Tumasyan, A.; Adam, W.; Bergauer, T.; Dragicevic, M.; Erö, J.; Fabjan, C.; Friedl, M.; Frühwirth, R.; et al. Transverse-Momentum and Pseudorapidity Distributions of Charged Hadrons in pp Collisions at $\sqrt{s} = 7$ TeV. *Phys. Rev. Lett.* **2010**, *105*, 022002.
88. Chatrchyan, S.; Khachatryan, V.; Sirunyan, A.M.; Tumasyan, A.; Adam, W.; Bergauer, T.; Dragicevic, M.; Erö, J.; Fabjan, C.; Friedl, M.; et al. Charged particle transverse momentum spectra in pp collisions at $\sqrt{s} = 0.9$ and 7 TeV. *J. High Energy Phys.* **2011**, *2011*, 86.
89. Adare, A.; Afanasiev, S.; Aidala, C.; Ajitanand, N.; Akiba, Y.; Al-Bataineh, H.; Alexander, J.; Aoki, K.; Aphecetche, L.; Armendariz, R.; et al. Measurement of neutral mesons in $p + p$ collisions at $\sqrt{s} = 200$ GeV and scaling properties of hadron production. *Phys. Rev. D* **2011**, *83*, 052004.
90. Majhi, A. Non-extensive statistical mechanics and black hole entropy from quantum geometry. *Phys. Lett. B* **2017**, *775*, 32–36.
91. Shore, J.E.; Johnson, R.W. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.
92. Caticha, A.; Giffin, A. Updating Probabilities. *AIP Conf. Proc.* **2006**, *872*, 31–42.
93. Presse, S.; Ghosh, K.; Lee, J.; Dill, K.A. Nonadditive Entropies Yield Probability Distributions with Biases not Warranted by the Data. *Phys. Rev. Lett.* **2013**, *111*, 180604.
94. Presse, S. Nonadditive entropy maximization is inconsistent with Bayesian updating. *Phys. Rev. E* **2014**, *90*, 052149.
95. Presse, S.; Ghosh, K.; Lee, J.; Dill, K.A. Reply to C. Tsallis' "Conceptual Inadequacy of the Shore and Johnson Axioms for Wide Classes of Complex Systems". *Entropy* **2015**, *17*, 5043–5046.
96. Vanslette, K. Entropic Updating of Probabilities and Density Matrices. *Entropy* **2017**, *19*, 664.
97. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* **1984**, *46*, 440–464.
98. Csiszár, I. Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Publ. Math. Inst. Hung. Acad. Sci.* **1963**, *3*, 85–107. (In German)
99. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Scientiarum Math. Hung.* **1967**, *2*, 299–318.
100. Csiszár, I. On topological properties of f-divergences. *Stud. Scientiarum Math. Hung.* **1967**, *2*, 329–339.
101. Csiszár, I. A class of measures of informativity of observation channels. *Priodica Math. Hung.* **1972**, *2*, 191–213.
102. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032–2066.
103. Lindsay, B.G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114.
104. Esteban, M.D.; Morales, D. A summary of entropy statistics. *Kybernetica* **1995**, *31*, 337–346.
105. Itakura, F.; Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968.

106. Fevotte, 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.
107. Teboulle, M.; Vajda, I. Convergence of best ϕ-entropy estimates. *IEEE Trans. Inf. Theory* **1993**, *39*, 297–301.
108. Basu, A.; Lindsay, B.G. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Stat. Math.* **1994**, *46*, 683–705.
109. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36.
110. Broniatowski, M.; Vajda, I. Several applications of divergence criteria in continuous families. *Kybernetika* **2012**, *48*, 600–636.
111. Toma, A. Optimal robust M-estimators using divergences. *Stat. Probab. Lett.* **2009**, *79*, 1–5.
112. Marazzi, A.; Yohai, V. Optimal robust estimates using the Hellinger distance. *Adv. Data Anal. Classif.* **2010**, *4*, 169–179.
113. Toma, A.; Leoni-Aubin, S. Optimal robust M-estimators using Renyi pseudodistances. *J. Multivar. Anal.* **2010**, *115*, 359–373.