On Some Entropy Functionals derived from Rényi Information Divergence

J.-F. Bercher [1]

Laboratoire des Signaux et Systèmes, CNRS-Univ Paris Sud-Supelec, 91192 Gif-sur-Yvette cedex, France

Abstract

We consider the maximum entropy problems associated with Rényi Q-entropy, subject to two kinds of constraints on expected values. The constraints considered are a constraint on the standard expectation, and a constraint on the generalized expectation as encountered in nonextensive statistics. The optimum maximum entropy probability distributions, which can exhibit a power-law behaviour, are derived and characterized.

The Rényi entropy of the optimum distributions can be viewed as a function of the constraint. This defines two families of entropy functionals in the space of possible expected values. General properties of these functionals, including nonnegativity, minimum, convexity, are documented. Their relationships as well as numerical aspects are also discussed. Finally, we work out some specific cases for the reference measure $Q(x)$ and recover in a limit case some well-known entropies.

Key words:
Rényi entropy, Rényi divergences, maximum entropy principle, nonextensivity, Tsallis distributions

1. Introduction

Consider two univariate continuous probability distributions with densities P and Q with respect to the Lebesgue measure. The Rényi information divergence introduced in [32] has the form

$$D_\alpha(P||Q) = -H_Q^{(\alpha)}(P) = \frac{1}{\alpha - 1} \log \int_{\mathcal{D}} P(x)^\alpha Q(x)^{1-\alpha} dx, \tag{1}$$

where α is a positive real and $\mathcal{D}$ the domain of definition of the integral. In the discrete case, the continuous sum is replaced by a discrete one which extends on a subset $\mathcal{D}$ of integers. The opposite $H_Q^{(\alpha)}(P)$ of the Rényi information divergence can be viewed as a Rényi entropy relative to the reference measure Q, and can be called Q-entropy. By L'Hospital's rule, Kullback divergence is recovered in the limit $\alpha \to 1$.

Applications and areas of interest in Rényi entropy are plentiful: communication and coding theory [10], data mining, detection, segmentation, classification [29,5], hypothesis testing [23], characterization of signals and sequences [38,19], signal processing [5,3], image matching and registration [29,15]. Connection with the log-likelihood has been

Email address: jf.bercher@esiee.fr (J.-F. Bercher).

[1] On sabbatical leave from ESIEE-Paris, France

outined in [33], where is also defined a measure of the intrinsic shape of a distribution which can serve as a measure of tail heaviness [27]. Rényi entropies for large families of univariate and bivariate distributions are given in [25,26]. Divergence measures based on entropy functions can be used in the process of inference [12], in clustering or partionning problems [22,2,7].
Rényi entropy also plays a central role in the theory of multifractals, see the review [18] and [4]. In statistical physics, following Tsallis proposal [34,35] of another entropy (which is simply related to Rényi entropy), there has been a high interest on these alternative entropies and the development of a community in "nonextensive thermostatistics". Indeed, the associated maximum entropy distributions exhibit a power-law behaviour, with a remarkable agreement with experimental data, see for instance [6,35] and references therein. These optimum distributions, called Tsallis distributions, are similar to Generalized Pareto Distributions, which also have an high interest in other fields, namely reliability theory [1], climatology [24], radar imaging [21] or actuarial sciences [8].
Jaynes' maximum entropy principle [16,17] suggests that the least biased probability distribution that describes a partially-known system is the probability distribution with maximum entropy compatible with all the available prior information. When prior information is available in the form of constraints on expected values, the maximum entropy method amounts to minimize Kullback information divergence $D(P||Q)$ (or equivalently maximizing Shannon Q-entropy) subject to normalization and these an observation constraints. In the case of a single constraint on the mean of the distribution, say $E_P[X] = m$, the minimum of Kullback information in the set of all probability distributions with expectation m is of course a function of m, denoted $\mathcal{F}(m)$ as follows

$$\mathcal{F}(m) = \left\{ \begin{array}{l} \min_P D(P||Q) \\ \text{s.t.} \quad m = E_P[X] \\ \text{and} \quad \int_{\mathcal{D}} P(x)dx = 1 \end{array} \right. \tag{2}$$

It is a 'contracted' version of Shannon Q-entropy and is called a level-1 entropy functional, or rate function, in the theory of large deviations, e.g. [11]. The maximum entropy method is a widely and successful method extensively used in a large variety of problems and contexts.

We focus here on solutions and properties of maximum entropy problems analog to (2) for the Rényi information divergence (1), and on the associated entropy functionals. The maximum Rényi-Tsallis entropy distribution, with its power law behavior, is at the heart of nonextensive statistics, but have also be considered in [13,14]. In nonextensive statistics, one still consider the usual classical mean constraint, but also a 'generalized' α-expectation constraint. This 'generalized' α-expectation is in fact the expectation with respect to the distribution

$$P^*(x) = \frac{P(x)^\alpha Q(x)^{1-\alpha}}{\int_{\mathcal{D}} P(x)^\alpha Q(x)^{1-\alpha} dx}, \tag{3}$$

that is a weighted geometric mean of P and Q. It is nothing else but the 'escort' or zooming distribution of nonextensive statistics [35] and multifractals. Of course, with $\alpha = 1$, the escort distribution P^* reduces to P and the generalized mean $E_{P^*}[X]$ reduces to the standard one.
Therefore, the maximum entropy problems associated to Rényi information divergence (1), subject to normalization and to a classical (C) or generalized (G) mean constraint states as:

$$\mathcal{F}_\alpha^{(C \text{ resp. } G)}(m) = \left\{ \begin{array}{l} \min_P D_\alpha(P||Q) \\ \text{s.t.} \quad (C) m = E_P[X] \\ \quad \text{or } (G) m = E_{P^*}[X] \\ \quad \text{and} \quad \int_{\mathcal{D}} P(x)dx = 1 \end{array} \right. \tag{4}$$

where $\mathcal{F}_\alpha^{(C)}(m)$ and $\mathcal{F}_\alpha^{(G)}(m)$ are the level-one entropy functionals associated to Rényi Q-entropy for the classical an generalized constraints respectively. Since Rényi entropy reduces to Shannon's for $\alpha = 1$, functionals $\mathcal{F}_\alpha^{(.)}(m)$ will reduce to $\mathcal{F}(m)$ when $\alpha \to 1$.

Hence, in this paper, we consider the forms and properties of maximum entropy solutions associated to Rényi Q-entropy, subject to two kind of constraints, as explained above. The value of the maximum entropy problems at the optimum define classes of entropy functionals $\mathcal{F}_\alpha^{(.)}(m)$ associated to each choice of reference Q, and indexed by the parameter α. The introduction of the reference measure Q, and therefore the definition of functionals $\mathcal{F}_\alpha^{(.)}(m)$ is, to the best of our knowledge, new in this setting. In section 2, the exact form of the probability distributions P that realize the minimum of the Rényi information divergence in the right side of (4) are first derived. Then we give some properties of these distributions and of their partition functions. We show that the entropy functionals $\mathcal{F}_\alpha^{(.)}(m)$ are simply linked to these partition functions. General properties of the entropy functionals, including nonnegativity, convexity, are established. We also indicate how the problems (4) can be tackled numerically, for specific values of the constraints, even thouh the maximum entropy distributions exhibit implicit relationships. A divergence in the object space, that reduces to a Bregman divergence for $\alpha \to 1$ is defined. These results are illustrated in section 3 where we study four special cases of reference Q, and characterize the associated entropy functionals. It is then shown that some well-known entropies are recovered.

2. The minimum of Rényi divergence

Let us define by

$$P_\nu(x) = \frac{[1+\gamma\,(x-\bar{x})]^\nu}{Z_\nu(\gamma,\bar{x})} Q(x), \tag{5}$$

a probability density function on a subset $\mathcal{D}$ of $\mathbb{R}$, where $\mathcal{D}$ ensure that the numerator of (5) is always nonnegative and its integral finite. The normalization $Z_\nu(\gamma,\overline{x})$ is the partition function defined by

$$Z_\nu(\gamma,\overline{x}) = \int_{\mathcal{D}} [1+\gamma(x-\overline{x})]^\nu\, Q(x)dx \tag{6}$$

The density P_ν depends of three parameters: the exponent ν which can be considered as a shape parameter, a scale parameter γ and a location parameter $\bar{x}$. But these parameters can be also be linked. For instance, $\bar{x}$ might be a function of ν and γ. When non ambigous, we may also denote by $E_\nu[X]$ the statistical mean with respect to $P_\nu(x)$.

With these notations, we have the following result.

Theorem 1

(C) The distribution $P_C(x)$ in the family (5) with $\nu = \xi = \frac{1}{\alpha-1}$ and $\overline{x} = E_P[X] = E_\xi[X]$, has the minimum Rényi divergence to Q

$$D_\alpha(P||Q) \geq D_\alpha(P_C||Q) \tag{7}$$

for all probability distributions $P(x)$ absolutely continuous with respect to $P_C(x)$ with a given (classical) expectation $\overline{x}$.

(G) The distribution $P_G(x)$ in the family (5) with $\nu = -\xi = \frac{1}{1-\alpha}$ and $\overline{x} = E_{P_G^*}[X] = E_{-(\xi+1)}[X]$, has the minimum Rényi divergence to Q

$$D_\alpha(P||Q) \geq D_\alpha(P_G||Q) \tag{8}$$

for all probability distributions $P(x)$ absolutely continuous with respect to $P_G(x)$ with a given generalized expectation $\overline{x}$.

Corollary 2 *The solution to the minimization of Rényi divergence in (4) is as given in theorem 1 for the particular values γ^* of γ such that $\overline{x} = m$.*

It is important to emphasize that $\overline{x}$ is here a statistical mean, and not the constraint m, and as such a function of γ.

Proof. See Appendix A ■

Remark 3 *When α tends to 1, $|\nu|$ tends to $+\infty$. Let us introduce $\tilde{\gamma}$ such that $\gamma = \tilde{\gamma}/\nu$. Then*

$$P_\nu(x) = e^{\nu \log\left[1+\frac{\tilde{\gamma}}{\nu}(x-\bar{x})\right] - \log Z_\nu(\tilde{\gamma},\bar{x})} Q(x), \tag{9}$$

and

$$\lim_{|\nu|\to+\infty} P_\nu(x) = e^{\tilde{\gamma}(x-\bar{x}) - \log Z_\nu(\tilde{\gamma},\bar{x})} Q(x), \tag{10}$$

that is the standard exponential, which is the well-known solution of the minimisation of Kullback-Leibler divergence subject to a constraint on an expected value [20, Theo 2.1, page 38]. In this case, the log-partition function becomes

$$\lim_{|\nu|\to+\infty} \log Z_\nu(\tilde{\gamma},\bar{x}) = \tilde{\gamma}\bar{x} - \log \int_{\mathcal{D}} e^{\tilde{\gamma}x} Q(x)dx \tag{11}$$

Properties of entropy functionals $\mathcal{F}_\alpha^{(C)}(m)$ and $\mathcal{F}_\alpha^{(G)}(m)$ are of course linked to the properties of the optimum distribution (5) and its partition function (6). In Property 4, we characterize partition functions of successive exponents, which enables to derive the expression of the Rényi entropy associated to the optimum distribution. In Proposition 6, we give the expression of the derivative of the partition function with respect to γ. Since the optimum distribution (5) is 'self-referential' (because it depends of its mean, which gives an implicit relation), direct determination of its parameters is difficult. It could rely on tabulation or on iterative techniques [36], that still suppose that the solution is an attractive fixed point. We define in Proposition 9 two functionals whose maximization provide the γ parameter of the optimum distributions associated to the classical and generalized mean constraint. Then general properties of nonnegativity, minimum, convexity are then given in Proposition 11. We also show that the two entropy optimization problems are related and that functionals $\mathcal{F}_\alpha^{(.)}(m)$ obey a special symmetry. Finally, we define a divergence in the space of possible means.

Property 4 *Partition functions of successive exponents are linked by*

$$Z_{\nu+1}(\gamma,\overline{x}) = E_{\nu+1-k}\left[(\gamma(x-\overline{x})+1)^k\right] Z_{\nu+1-k}(\gamma,\overline{x}). \tag{12}$$

An interesting particular case is for k=1:

$$Z_{\nu+1}(\gamma,\overline{x}) = E_\nu\left[\gamma(x-\overline{x})+1\right] Z_\nu(\gamma,\overline{x}). \tag{13}$$

This is easily checked by direct calculation. As a direct consequence, we may also observe that $Z_{\nu+1}(\gamma,\overline{x}) = Z_\nu(\gamma,\overline{x})$ if and only if $\overline{x} = E_\nu[x]$. When $\overline{x}$ is a fixed parameter m, this will be only true for a special value γ^* such that $E_\nu[x] = m$.

Now, using (13) in Property 4, it is possible to give the expression of the Rényi divergence associated to the distribution (5) and in particular to the solutions P_C and P_G of problems (4):

Property 5 *The Rényi information divergence associated to the optimum distributions (5) in theorem 1 is (C)* $D_\alpha(P||Q) = -\log Z_\xi(\gamma,\overline{x}) = -\log Z_{\xi+1}(\gamma,\overline{x})$*, and (G)* $D_\alpha(P||Q) = -\log Z_{-\xi}(\gamma,\overline{x}) = -\log Z_{-(\xi+1)}(\gamma,\overline{x})$.

Proof.

The Rényi entropy associated to (5) writes

$$\begin{aligned} D_\alpha(P||Q) &= \frac{1}{\alpha-1}\log\int P(x)^\alpha Q(x)^{1-\alpha}dx \\ &= \frac{1}{\alpha-1}\log\int \left(1+\gamma\left(x-\bar{x}\right)\right)^{\alpha\nu} Q(x)dx - \frac{\alpha}{\alpha-1}\log Z_\nu(\gamma,\bar{x}), \end{aligned}$$

that simply reduces to

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log Z_{\alpha\nu}(\gamma,\bar{x}) - \frac{\alpha}{\alpha-1}\log Z_\nu(\gamma,\bar{x}).$$

(C) In one hand, if $\nu = \xi = \frac{1}{\alpha-1}$, then $\alpha\nu = \frac{\alpha}{\alpha-1} = \xi+1$, and $D_\alpha(P||Q) = \frac{1}{\alpha-1}\log Z_{\xi+1}(\gamma,\bar{x}) - \frac{\alpha}{\alpha-1}\log Z_\xi(\gamma,\bar{x})$ Therefore, when $\bar{x} = E_\xi[x]$, then (13) gives $Z_{\xi+1}(\gamma,\bar{x}) = Z_\xi(\gamma,\bar{x})$, and it simply remains

$$D_\alpha(P||Q) = -\log Z_\xi(\gamma,\overline{x}) = -\log Z_{\xi+1}(\gamma,\overline{x}).$$

(G) In the other hand, if $\nu = -\xi = \frac{1}{1-\alpha}$, then $\alpha\nu = \frac{\alpha}{1-\alpha} = -\xi-1$, and $D_\alpha(P||Q) = \frac{1}{\alpha-1}\log Z_{-(\xi+1)}(\gamma,\bar{x}) - \frac{\alpha}{\alpha-1}\log Z_{-\xi}(\gamma,\bar{x})$. When $\bar{x} = E_{-(\xi+1)}[x]$, we have $Z_{-\xi}(\gamma,\bar{x}) = Z_{-(\xi+1)}(\gamma,\bar{x})$ according to (13) and it remains

$$D_\alpha(P||Q) = -\log Z_{-\xi}(\gamma,\overline{x}) = -\log Z_{-(\xi+1)}(\gamma,\overline{x}).$$

■

Since the Rényi information divergence of distributions (5) is simply the log-partition function, it will be useful to examine the behaviour of the partition function with respect to the parameter γ. Hence, the following proposition gives the expression of the derivative of the partition function.

Proposition 6 *For the partition function (6) with domain of definition $\mathcal{D}$, the derivative with respect to γ of the partition function with characteristic exponent ν is given by*

$$\frac{d}{d\gamma} Z_\nu(\gamma, \overline{x}) = \nu \left(E_{\nu-1}\left[x - \overline{x}\right] - \gamma \frac{d\overline{x}}{d\gamma} \right) Z_{\nu-1}(\gamma, \overline{x}). \tag{14}$$

if (a) the domain $\mathcal{D}$ does not depend of γ, or (b) on subsets of γ such that the domain increment $\delta\mathcal{D}$ associated to the variation $\delta\gamma$ remains empty, or (c) for $\nu > 0$ in the continuous case or $\nu > 1$ in the discrete case.

Proof. See Appendix B ■

Using this proposition on the derivative of the partition function and Property 4 on the link between partitions functions of succesive exponents, we readily have

Property 7 *If $\overline{x} = E_{\nu-1}\left[X\right]$, then, with the same conditions as in proposition 6:*

$$\frac{d}{d\gamma} \log Z_\nu(\gamma, \overline{x}) = -\gamma\nu \frac{d\overline{x}}{d\gamma}, \tag{15}$$

and

$$\frac{d}{d\overline{x}} \log Z_\nu(\gamma, \overline{x}) = -\gamma\nu. \tag{16}$$

This is immediately checked using (13) and (14) with $\overline{x} = E_{\nu-1}[X]$. It is now interesting to consider the special case where $\overline{x}$ is a fixed value, say m. Then, it is immediate to check that the extrema of the function $\log Z_\nu(\gamma, m)$ occur for γ^* such that $m = E_{\nu-1}[X]$:

Property 8 *If $\overline{x}$ is a fixed value m, then*

$$\frac{d}{d\gamma} \log Z_\nu(\gamma, m)\Big|_{\gamma=\gamma^*} = 0. \tag{17}$$

if and only if γ^* is such that $m = E_{\nu-1}\left[X\right]$.

This result is important because it provides an easy way to find the value of the parameter γ of the optimum distributions (5) that solves the maximum entropy problems (4).

Proposition 9 *The values γ^* of the parameter γ of the optimum distributions that solve the maximum entropy problems (4) are the minimum of the maximizers of*

$$D_C(\gamma) = -\log Z_{\xi+1}(\gamma, m) \tag{18}$$

$$D_G(\gamma) = -\log Z_{-\xi}(\gamma, m) \tag{19}$$

where the two partitions functions involved are convex, possibly on several well defined intervals. Then, the entropy functionals $\mathcal{F}_\alpha^{(.)}$ are simply given by

$$\mathcal{F}_\alpha^{(C \text{ resp. } G)}(m) = D_{C \text{ resp. } G}(\gamma^*). \tag{20}$$

Proof. Indeed, Theorem 1 and its corrolary indicates that the solution for the classical constraint (C) is obtained for $\overline{x} = m = E_\xi\left[X\right]$ and by $\overline{x} = m = E_{-\xi-1}\left[X\right]$ for the generalized constraint (G). Then by Property 8 it suffices to look for the extrema of $D_C(\gamma) = -\log Z_{\xi+1}(\gamma, m)$ in the first case or of $D_G(\gamma) = -\log Z_{-\xi}(\gamma, m)$ in the second case. With similar conditions of derivation as in Proposition 6 the second derivative of the partition function with respect to γ writes

$$\frac{d^2 Z_\nu(\gamma, m)}{d\gamma^2} = \nu(\nu-1) \int_{\mathcal{D}} (x-m)^2 \left[1 + \gamma\left(x - \bar{x}_\gamma\right)\right]^{\nu-2} Q(x) dx \tag{21}$$

$$= \nu(\nu-1) E_{\nu-2}\left[(X-m)^2\right] Z_{\nu-2}(\gamma, m). \tag{22}$$

For $\nu = \xi + 1$ and $\nu = -\xi$, the factor $\nu(\nu-1)$ reduces to $\frac{\alpha}{(\alpha-1)^2}$. Since α is positive, the second derivative is always positive and the partition functions $Z_{\xi+1}(\gamma, m)$ and $Z_{-\xi}(\gamma, m)$ are convex on their domain of definition. On these domains, the functionals in (18) and (19) are then unimodal and their extrema are maxima.

In the discrete case and for $\nu < 0$, $Z_\nu(\gamma, m)$ has singularities for all $\gamma = \frac{1}{m-k}$, where k is an integer in the support of the distribution. Therefore, $Z_\nu(\gamma, m)$ is only defined on segments $\left(\frac{1}{m-k}, \frac{1}{m-k-1}\right)$, for $m \notin (k+1, k)$), and $\left(\frac{1}{m-k-1}, \frac{1}{m-k}\right)$ for $m \in (k+1, k)$. In such a case, $-\log Z_\nu(\gamma, m)$ may present several maxima. The situation

$\nu < 0$ occurs for the classical constraint when $\alpha \in (0,1)$ (since the index $\xi + 1 = \alpha/(\alpha-1)$ is negative), and for the generalized constraint when $\alpha > 1$. An example of functional $D_C(\gamma)$ with $\alpha = 0.5$ in the case of a Poisson distribution is reported in Fig. 6. In the $\nu > 0$ discrete case or in the continuous case, there is a single maximum.
Finally, since the expression of the Rényi information divergence of the optimum distributions is precisely the opposite of the log-partition function as indicated in Property 5, the value of functionals (18) and (19) at their optima γ^* such that $\bar{x} = m$ is precisely the value of entropy functionals $\mathcal{F}_\alpha^{(1)}(m)$ and $\mathcal{F}_\alpha^{(\alpha)}(m)$. ■

Remark 10 *When α tends to 1, the parameter $\tilde{\gamma}^*$ is thus the maximizer of (11), and we obtain*

$$\lim_{\alpha \to 1} \mathcal{F}_\alpha^{(.)} = \sup_{\tilde{\gamma}} \left\{ \tilde{\gamma}\bar{x} - \log \int_{\mathcal{D}} e^{\tilde{\gamma}x} Q(x) dx \right\}, \tag{23}$$

that is the Cramér transform of $Q(x)$.

With the help of these different results it is now possible to characterize more precisely the entropy functionals

Proposition 11 *Entropy functionals $\mathcal{F}_\alpha^{(C)}(m)$ and $\mathcal{F}_\alpha^{(G)}(m)$ are nonnegative, with an unique minimum at m_Q , the mean of Q, and $\mathcal{F}_\alpha^{(.)}(m_Q) = 0$. Furthermore, $\mathcal{F}_\alpha^{(C)}(m)$ is strictly convex for $\alpha \in [0,1]$.*

Proof. Rényi information divergence $D_\alpha(P||Q)$ is always nonnegative, and equal to zero for $P = Q$. Since functionals $\mathcal{F}_\alpha^{(.)}(x)$ are defined as the minimum of $D_\alpha(P||Q)$, they are always nonnegative. If $P = Q$, we have also $P^* = Q$ and $m = E_P[X] = E_{P^*}[X] = m_Q$. Therefore $\mathcal{F}_\alpha^{(.)}(m_Q) = 0$ and m_Q is a global minimum.
From (16), we have $\frac{d}{d\overline{x}} \log Z_{\nu+1}(\gamma, \overline{x}) = -\gamma(\nu+1)$. Then, functionals $\mathcal{F}_\alpha^{(.)}(x)$ are only minimum if $\gamma = 0$, and the corresponding optimum probability distributions are simply $P = Q$, and $D_\alpha(Q||Q) = 0$. Therefore, $\mathcal{F}_\alpha^{(.)}(x)$ have an unique minimum for $x = m_Q$, the mean of Q, and $\mathcal{F}_\alpha^{(.)}(m_Q) = 0$.
Finally, we examine the convexity of $\mathcal{F}_\alpha^{(C)}(m)$, for $\alpha \in [0,1]$.
Let P_1 and P_2 be the distributions that achieve the minimization of $D_\alpha(P||Q)$ subject to the constraints $x_1 = E_P[X]$ and $x_2 = E_P[X]$ respectively. Then, $\mathcal{F}_\alpha^{(C)}(x_1) = D_\alpha(P_1||Q)$, and $\mathcal{F}_\alpha^{(C)}(x_2) = D_\alpha(P_2||Q)$. In the same way, denote $\mathcal{F}_\alpha^{(C)}(\mu x_1 + (1-\mu)x_2) = D_\alpha(\hat{P}||Q)$, where $\hat{P}$ denotes the optimum distribution with mean $\mu x_1 + (1-\mu)x_2$. Distributions $\hat{P}(u)$ and $\mu P_1(u) + (1-\mu)P_2(u)$ have the same mean $\mu x_1 + (1-\mu)x_2$. Hence, when $D_\alpha(P||Q)$ is a convex function of P, that is for $\alpha \in [0,1]$, we have $D_\alpha(P^*||Q) \leq \mu D_\alpha(P_1(u)||Q) + (1-\mu)D_\alpha(P_2(u)||Q)$, that is $\mathcal{F}_\alpha^{(C)}(\mu x_1 + (1-\mu)x_2) \leq \mu \mathcal{F}_\alpha^{(C)}(x_1) + (1-\mu)\mathcal{F}_\alpha^{(C)}(x_2)$ and $\mathcal{F}_\alpha^{(C)}(x)$ is a convex function. ■
Up to now the two optimization problems have been considered in parallel. But here is a special symmetry that enables to relate the solutions of the minimization of Rényi divergence subject to classical and generalized mean constraints. Then, there exists a simple relationship between the entropy functionals $\mathcal{F}_\alpha^{(C)}(x)$ and $\mathcal{F}_\alpha^{(G)}(x)$.
Let us consider our original Rényi divergence minimization problem, on one hand with index α_1 and subject to a classical mean constraint m, and on the other hand with index α_2 and subject to a generalized mean constraint m. The associated functionals, by Property 9, are $D_C(\gamma) = -\log Z_{\xi_1+1}(\gamma, m)$ and $D_G(\gamma) = -\log Z_{-\xi_2}(\gamma, m)$. Thus, we will have pointwise equality of these functions if $\xi_1 + 1 = -\xi_2$, that is if indexes α_1 and α_2 satisfy $\alpha_1 = 1/\alpha_2$. In this case, we will of course have equality of the optimum parameters γ, and the two optimization problems will have the same optimum value. Because of the pointwise equality functions $D_G(\gamma)$ and $D_G(\gamma)$, it is clear that the associated divergences are equal at the optimum, that is $D_{\alpha_1}(P_C||Q) = D_{\alpha_2}(P_G||Q)$. Besides this is easily checked in the general case: for the escort distribution $P^*(x)$ in (3), we *always* have the equality $D_{\frac{1}{\alpha}}(P^*||Q) = D_\alpha(P_1||Q)$. Hence, the minimization of the α Rényi divergence subject to the generalized mean constraint is exactly equivalent to the minimization of the $1/\alpha$ Rényi divergence subject to the classical mean constraint

$$\begin{cases} \inf_{P_1} D_\alpha(P_1||Q) \\ s.t \;\; E_{P^*}[X] = m \end{cases} = \begin{cases} \inf_{P^*} D_{\frac{1}{\alpha}}(P^*||Q) \\ s.t \;\; E_{P^*}[X] = m \end{cases}, \tag{24}$$

so that generalized and classical mean constraints can always be swapped, provided the index α is changed into $1/\alpha$, as was argued in [31,28]. Hence, equality (24) enables us to complete the characterization of entropy functionals $\mathcal{F}_\alpha^{(C)}(m)$ and $\mathcal{F}_\alpha^{(G)}(m)$:

Property 12 *Entropy functionals $\mathcal{F}_\alpha^{(C)}(m)$ and $\mathcal{F}_\alpha^{(G)}(m)$ admit the symmetry $\mathcal{F}_\alpha^{(G)}(x) = \mathcal{F}_{1/\alpha}^{(C)}(x)$. Besides, $\mathcal{F}_\alpha^{(C)}(m)$ is strictly convex for $\alpha \in [0,1]$ and $\mathcal{F}_\alpha^{(G)}(m)$ is strictly convex for $\alpha \in [1,+\infty]$.*

Interestingly, it is also possible to define a *divergence in the object space*, that is a kind of generalized distance between two "objects". These divergences may be used for instance in clustering [30]. The objects are here considered as generalized means of distributions with minimum divergence to a reference measure $Q(x)$.

Proposition 13 *If P_1 and P_2 are two distributions in (5) with exponent $\nu = -\xi$ (generalized constraint), with $P_2 \ll P_1$, and with respective parameters γ_1, γ_2 and means m_1, m_2, then*

$$\mathcal{F}_\alpha^{(G)}(m_2, m_1) = D_\alpha(P_2||P_1) = \mathcal{F}_\alpha^{(G)}(m_2) - \mathcal{F}_\alpha^{(G)}(m_1) + \frac{1}{\alpha - 1} \log \left(1 - (\alpha - 1) \frac{d\mathcal{F}_\alpha^{(G)}}{dm}(m_1)(m_2 - m_1) \right), \tag{25}$$

and $\mathcal{F}_\alpha^{(G)}(m_2, m_1) \geq 0$, with equality if and only if $m_2 = m_1$.

Proof. The result is obtained by simple computations. First, we have

$$D_\alpha(P_2||P_1) = \frac{1}{\alpha - 1} \log \int \frac{[1 + \gamma_2(x - m_2)]^{\frac{\alpha}{1-\alpha}}}{Z_{-\xi}(\gamma_2, m_2)^\alpha} \frac{[1 + \gamma_1(x - m_1)]}{Z_{-\xi}(\gamma_1, m_1)^{1-\alpha}} Q(x)dx$$

which can be rewritten as

$$D_\alpha(P_2||P_1) = \frac{1}{1-\alpha} \left(\alpha \log Z_{-\xi}(\gamma_2, m_2) + (1-\alpha) \log Z_{-\xi}(\gamma_1, m_1) - \log Z_{-\xi-1}(\gamma_2, m_2) \right) \tag{26}$$

$$+ \frac{1}{\alpha - 1} \log \left[1 + \gamma_1 \int (x - m_1) \frac{[1 + \gamma_2(x - m_2)]^{-\xi-1}}{Z_{-\xi-1}(\gamma_2, m_2)} Q(x)dx \right] \tag{27}$$

In the first line, we have $Z_{-(\xi+1)}(\gamma_2, m_2) = Z_{-\xi}(\gamma_2, m_2)$ by Property 4, eq. (13), and we recognize from Proposition 9 that $\mathcal{F}_\alpha^{(G)}(m) = -\log Z_{-\xi}(\gamma, m)$. In the second line, the integral reduces to $(m_2 - m_1)$ since m_2 is the generalized mean of the distribution P_2. Finally, γ_1 can be expressed as the derivative of the log-partition function as stated by (16) in Property 7.

By definition, $\mathcal{F}_\alpha^{(G)}(m_2, m_1)$ is the Rényi information divergence $D_\alpha(P_2||P_1)$ which is always greater or equal to zero, with equality if and only if $P_2 = P_1$, which implies $m_2 = m_1$. ■

For $\alpha \to 1$, $\mathcal{F}_\alpha^{(G)}(m_2, m_1)$ reduces to a standard *Bregman divergence*. Indeed, using $\log(1 - x) \simeq -x$, we have simply

$$\lim_{\alpha \to 1} \mathcal{F}_\alpha^{(G)}(m_2, m_1) = \mathcal{F}_\alpha^{(G)}(m_2) - \mathcal{F}_\alpha^{(G)}(m_1) - \frac{d\mathcal{F}_\alpha^{(G)}}{dm}(m_1)(m_2 - m_1).$$

3. Examples of entropy functionals

We now examine 4 special cases for the reference mesure $Q(x)$: a uniform and an exponential distribution that model systems with continuous states; and then a Bernoulli (two-levels) and a Poisson distribution which may model systems with discrete states. The minima of the Rényi divergence, that is the entropies $\mathcal{F}_\alpha^{(C \text{ or } G)}(x)$, are attained for the values γ^* that maximize the functionals $D_C(\gamma)$ and $D_G(\gamma)$ in Proposition 9. This involves the computation of $Z_\nu(\gamma, m)$ for all reference measures Q considered, and the resolution of $\frac{d}{d\gamma} Z_{\nu+1}(\gamma, m) = 0$. The case $\alpha = 1$ is obtained in the limit $|\nu| \to +\infty$, since $|\xi| \to +\infty$ when α tends to 1. Results of numerical evaluations for varying α are provided.

3.1. *Uniform reference*

Let us first consider the case of the uniform reference $Q(x)$ on $[0, 1$. The partition function is given by $Z_\nu(\gamma, m) = \int_{\mathcal{D}} [\gamma(x - m) + 1]^\nu \, dx$, where the domain $\mathcal{D}$ is defined by $\mathcal{D} = \mathcal{D}_Q \cap \mathcal{D}_\gamma$, with $\mathcal{D}_Q = \{x : x \in [0, 1]\}$ and $\mathcal{D}_\gamma = \{x : \gamma(x - m) + 1 \geq 0\}$.

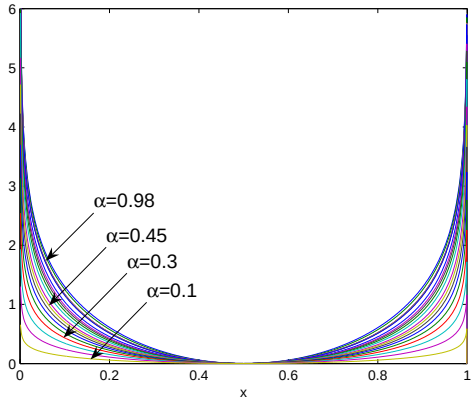

Fig. 1. Entropy functional $\mathcal{F}_\alpha^{(C)}(x)$ for a uniform reference measure and $\alpha \in (0,1)$.

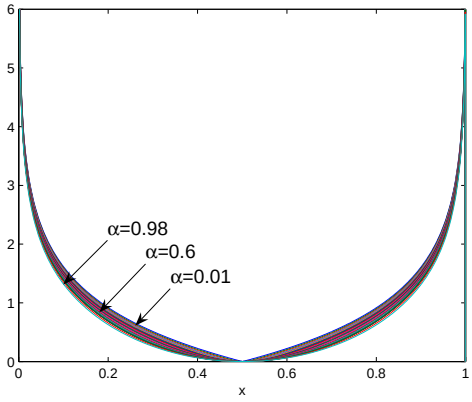

Fig. 2. Entropy functional $\mathcal{F}_\alpha^{(G)}(x)$ for a uniform reference measure and $\alpha \in (0,1)$.

Computation of the partition function in the different domains together with the fact that $m \in [0,1]$ leads to

$$Z_\nu(\gamma,m) = \frac{1}{\gamma\,(1+\nu)}\left((\gamma-\gamma m+1)^{\nu+1}\,U(\gamma-\frac{1}{m-1}) - (-\gamma m+1)^{\nu+1}\,U(-\gamma+\frac{1}{m})\right),$$

$$\text{for all } \gamma \text{ if } \nu \geq 0, \text{ for } \gamma \in \left(\frac{1}{m-1},\frac{1}{m}\right) \text{ if } \nu < 0, \text{ and } Z_\nu(\gamma,m) = +\infty \text{ otherwise,}$$

where U denotes the Heaviside distribution: $U(t) = 0$ for $t < 0$ and $U(t) = 1$ for $t > 0$.
The first derivative of the partition function is given by

$$\frac{d}{d\gamma}Z_\nu(\gamma,m) = -\frac{\nu\gamma\,(m-1)+1}{\gamma^2\,(\nu+1)}\,(\gamma\,(m-1)+1)^\nu\,U(\gamma-\frac{1}{m-1}) + \frac{\gamma m(\nu)+1}{\gamma^2\,(\nu+1)}\,(1-\gamma m)^\nu\,U(-\gamma+\frac{1}{m}). \quad (28)$$

We next have to look for the expression of entropy functionals $\mathcal{F}_\alpha^{(.)}(x)$. Unfortunately, no analytical solution can be exhibited here, but the two functionals still can be evaluated numerically. For the classical mean constraint (C) we can check that $\mathcal{F}_\alpha^{(C)}(x)$ is a family of convex functions on $(0,1)$, minimum for the mean of the reference measure Q, as was indicated in Proposition 11. In the same way, we can check that for the generalized mean constraint (G) $\mathcal{F}_\alpha^{(G)}(x)$ is a family of nonnegative functions on $(0,1)$, also minimum for the mean of the reference measure Q. The entropies $\mathcal{F}_\alpha^{(C)}(x)$ and $\mathcal{F}_\alpha^{(G)}(x)$ were evaluated numerically and are given in Figs. 1 and 2 for $\alpha \in (0,1)$. Of course, the $\alpha \leftrightarrow 1/\alpha$ duality given in Property 12 enables to extend these two functionals for $\alpha > 1$.
Hence, it is apparent that the minimization of $\mathcal{F}_\alpha^{(.)}(x)$ under some constraint would automatically lead to a solution on (0,1). Moreover, the parameter α may serve to tune the curvature of the functional and the degree of penalization of bounds.

3.2. *Exponential reference*

The exponential probability density function is $Q(x) = \beta e^{-\beta x}$, for $x \geq 0$ and $\beta > 0$. The partition function is given by

$$Z_\nu(\gamma,m) = \beta\int_{\mathcal{D}}\left[\gamma(x-m)+1\right]^\nu e^{-\beta x}dx \quad (29)$$

where $\mathcal{D} = \left\{x : x \geq \max\left\{0, m-\frac{1}{\gamma}\right\} \text{ if } \gamma > 0 \text{ or } x \in [0, m-\frac{1}{\gamma}] \text{ if } \gamma < 0\right\}$, ensures that the integrand $[\gamma(x-m)+1]$ is nonnegative and the integral finite.
The evaluation of $Z_\nu(\gamma,m)$ on the different domains gives:

$$Z_\nu(\gamma,m)=\begin{cases} e^{-\beta\frac{\gamma m-1}{\gamma}}\left(\frac{\gamma}{\beta}\right)^\nu\Gamma(\nu+1) & \text{if } \gamma>\frac{1}{m}>0,\ \nu\geq 0 \\ e^{-\beta\frac{\gamma m-1}{\gamma}}\left(\frac{\gamma}{\beta}\right)^\nu\Gamma(\nu+1,\beta\frac{1-\gamma m}{\gamma}) & \text{if } \frac{1}{m}>\gamma>0 \\ e^{-\beta\frac{\gamma m-1}{\gamma}}\left(\frac{\beta}{\gamma}\right)^{-\nu}\left(\Gamma\left(\nu+1,\beta\frac{-\gamma m+1}{\gamma}\right)-\Gamma(\nu+1)\right) & \text{if } \gamma<0<\frac{1}{m},\ \nu\geq 0 \end{cases} \tag{30}$$

and $Z_\nu(\gamma,m)=+\infty$ for $\gamma<0$ or $\gamma>\frac{1}{m}$ if $\nu<0$.

Let us now examine the behavior of the entropies $\mathcal{F}_\alpha^{(.)}(x)$ when $\alpha\to 1$. This amounts to study $Z_\nu(\gamma,m)$ and its maximum when $|\nu|\to+\infty$.

The simplest derivation is as follows. As in Remark 3, let $\gamma=\tilde{\gamma}/\nu$, so that $(1+\gamma(x-m))^\nu\sim\exp(\tilde{\gamma}(x-m))$. In this case, one easily obtain that

$$\log Z_\nu(\tilde{\gamma},m)\simeq\log\beta-\tilde{\gamma}m-\log(\beta-\tilde{\gamma}), \tag{31}$$

whose derivative is equal to zero for

$$\tilde{\gamma}^*=\beta-\frac{1}{m}. \tag{32}$$

We shall also note that if $\nu<0$, the sign of $\gamma=\tilde{\gamma}/\nu$ is the sign of $(1-\beta m)$. Since $Z_\nu(\gamma,m)$ is only defined for $\gamma>0$ when $\nu<0$, it means that we only have a solution for $m<1/\beta$. Indeed, for $\gamma>0$ and $\nu<0$, the factor $(1+\gamma(x-m))^\nu$ is decreasing, and consequently the mean of the optimum distribution (5) cannot be greater than the mean of the reference distribution, $m_Q=E_Q[X]=1/\beta$.

With the optimum value $\tilde{\gamma}^*$, the log partition function becomes

$$\log Z_\nu(\gamma^*,m)\simeq-(\beta m-1)+\log(\beta m)\qquad(\forall m \text{ if } \nu\to+\infty, \text{for } m<1/\beta \text{ if } \nu\to-\infty). \tag{33}$$

Finally, we thus obtain

$$\mathcal{F}^{(C)}_{\alpha\to 1}(x)=-\log Z_{\xi+1}(\gamma^*,x)=(\beta x-1)-\log(\beta x), \tag{34}$$

for $x<1/\beta$ when α tends to 1 by lower values, and for all x if α tends to 1 by higher values. By the duality property 12, this expression is also the limit form of functional $\mathcal{F}^{(G)}_\alpha(x)$.

As was expected, the functional $(\beta x-1)-\log(\beta x)$ is strictly convex, positive and zero for $x=1/\beta$, the mean of the exponential distribution. It was employed in speech processing and is called the *Itakura-Saïto entropy functional*. For $\beta=1$, it reduces to the so-called *Burg entropy* that is well-known in spectrum analysis.

The entropy functionals can be evaluated numerically. For instance, $\mathcal{F}^{(G)}_\alpha(x)$ is given on Fig. 3 for $\alpha>0$. It is a family of nonnegative functions, equal to zero for $x=m_Q=1/\beta$, and convex for $\alpha\in[1,+\infty)$.

3.3. *Bernoulli reference*

Let us now consider the case of the Bernoulli measure $Q(x)=\beta\delta(x)+(1-\beta)\delta(x-1)$. Of course, the (generalized) mean of optimum distributions is somewhere in the interval $[0,1]$. When γ is outside of the interval $(\frac{1}{m-1},\frac{1}{m})$, the probability distribution reduces to a pure state — $\delta(x)$ or $\delta(x-1)$, and its (generalized) mean is 0 or 1. Incorporation of the bounds into the domain depends on the sign of ν : for $\nu<0$, $Z_\nu(\gamma,m)$ diverges to $+\infty$ on the bounds whereas it remains finite for $\nu>0$. The expression of the partition function follows directly from the definition:

$$Z_\nu(\gamma,m)=\beta(1-\gamma m)^\nu+(1-\beta)(1+\gamma(1-m))^\nu. \tag{35}$$

In contrast to the previous case, it is possible here to obtain an explicit expression of the entropy functionals for any α. Indeed, if p denotes the value of the optimum distribution at $x=1$, then the generalized expectation is

$$m=\frac{\sum_{x=0}^1 xP(x)^\alpha Q(x)^{1-\alpha}}{\sum_{x=0}^1 P(x)^\alpha Q(x)^{1-\alpha}}=\frac{(1-\beta)^{1-\alpha}p^\alpha}{\beta^{1-\alpha}(1-p)^\alpha+(1-\beta)^{1-\alpha}p^\alpha} \tag{36}$$

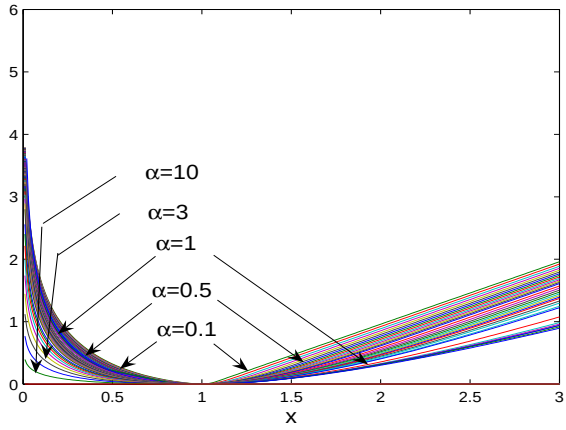

Fig. 3. Entropy functional $\mathcal{F}_\alpha^{(G)}(x)$ for an exponential reference measure with $\beta = 1$ and $\alpha > 0$. By Property 12 it is also $\mathcal{F}_{1/\alpha}^{(C)}(x)$.

and it is therefore possible to express p as a function of m:

$$p = \frac{\left(\beta^{1-\alpha} x\right)^{\frac{1}{\alpha}}}{\left(\beta^{1-\alpha} x\right)^{\frac{1}{\alpha}} + \left((1-\beta)^{1-\alpha}(1-x)\right)^{\frac{1}{\alpha}}}. \tag{37}$$

Now, since the Rényi information divergence is

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log\left[\beta^{1-\alpha}(1-p)^\alpha + (1-\beta)^{1-\alpha} p^\alpha\right] \tag{38}$$

it suffices to replace p by the expression (37) which leads to

$$\mathcal{F}_\alpha^{(G)}(m) = \frac{\alpha}{1-\alpha} \log\left[\beta^{1-\frac{1}{\alpha}}(1-m)^{\frac{1}{\alpha}} + (1-\beta)^{1-\frac{1}{\alpha}} m^1 \alpha\right] \tag{39}$$

The case of the classical mean is even simpler: we have $m = p$, and $\mathcal{F}_\alpha^{(C)}(m)$ has the expression of the divergence in (38) with p replaced by m. It is also interesting to note, and check, that the $\alpha \leftrightarrow 1/\alpha$ duality of Property 12 links these two expressions.
The limit case $\alpha \to 1$ is easily derived using L'Hospital's rule. It comes

$$\mathcal{F}_{\alpha\to 1}^{(.)}(x) = x \ln\left(\frac{x}{1-\beta}\right) + (1-x)\ln\left(\frac{1-x}{\beta}\right). \tag{40}$$

This expression is the celebrated *Fermi-Dirac entropy* that is strictly convex, nonnegative, and equal to zero for $x = E_Q[X] = 1 - \beta$, the mean m_Q of the reference measure.
Plots of the entropy functionals are given in Figs. 4 and 5 for $\alpha \in (0, 1)$ and $\beta = 1/2$. In both cases, we have a family of nonnegative functions, equal to zero for the mean of the reference measure. It can also be checked that $\mathcal{F}_\alpha^{(C)}(x)$ is convex for $\alpha \in (0, 1]$.

3.4. *Poisson reference*

As a final example, let us consider the case of a Poisson measure $Q(x) = \frac{\mu^x}{x!} e^{-\mu}$, for $x \geq 0$. Domain $\mathcal{D}$ is $\mathcal{D} = \mathcal{D}_Q \cap \mathcal{D}_\gamma$, where $\mathcal{D}_Q = \mathbb{N}^+$ and $\mathcal{D}_\gamma = \{x : \gamma(x - m) + 1 \geq 0\}$. The partition function is given by

$$Z_\nu(\gamma, m) = \sum_{\mathcal{D}} \left[\gamma(x-m)+1\right]^\nu \frac{\mu^x}{x!} e^{-\mu}. \tag{41}$$

Three cases appear, according to the value of γ:
(a) if $\frac{1}{m} \geq \gamma \geq 0$, then $\mathcal{D}$ reduces to $\mathcal{D}_1 = \{x : x \in [0, +\infty)\}$;

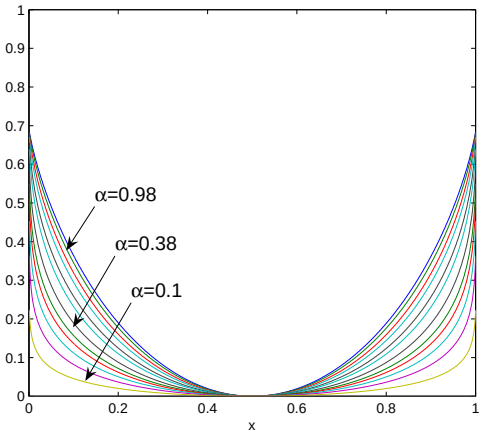

Fig. 4. Entropy functional $\mathcal{F}_\alpha^{(C)}(x)$ for a Bernoulli reference measure and $\alpha \in (0,1)$.

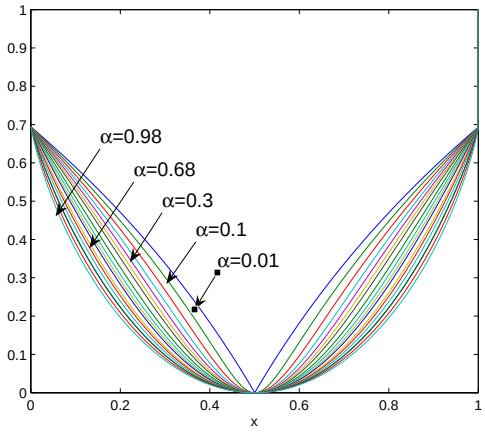

Fig. 5. Entropy functional $\mathcal{F}_\alpha^{(G)}(x)$ for a Bernoulli reference measure and $\alpha \in (0,1)$.

(b) for $\gamma \geq \frac{1}{m}$ the domain is $\mathcal{D}_2 = \left\{ x : x \in [\left\lceil m - \frac{1}{\gamma} \right\rceil, +\infty) \right\}$;

(c) when $\gamma < 0$, $\mathcal{D} = \mathcal{D}_3 = \left\{ x \in [0, \left\lfloor m - \frac{1}{\gamma} \right\rfloor] \right\}$.

In these expressions $\lfloor x \rfloor$ denotes the floor function that returns the largest integer less than or equal to x; and $\lceil x \rceil$ is the ceil function, the smallest integer not less than x.

Closed-form formulas can not be derived in the general case, but only in the case of an integer exponent ν. When ν is not an integer, we will have to resort to the serie (41), possibly truncated for numerical computations. In order to save space, we only sketch the derivation in $\mathcal{D}_1$:

$$Z_\nu(\gamma, m) = (1 - \gamma m)^\nu \, e^{-\mu} \sum_{x=0}^{+\infty} (\theta x + 1)^\nu \, \frac{\mu^x}{x!} \tag{42}$$

with $\theta = \frac{\gamma}{1-\gamma m}$. In the serie above the ratio of successive terms $\frac{(1+\theta x+\theta))^\nu}{(x+1)(1+\theta x)^\nu}\mu$ is the ratio of two completely factored polynomials. This indicates that the serie can be written as a generalized hypergeometric function, when ν is integer. So doing, we obtain

$$Z_\nu(\gamma, m) = (1 - \gamma m)^\nu \, e^{-\mu} \, {}_{|\nu|}F_{|\nu|}(a, ..., a, b, ..., b, \mu)$$

with $a = (1+\theta)/\theta$ and $b = 1/\theta$ for $\nu > 0$; or with $a = 1/\theta$ and $b = (1+\theta)/\theta$ for $\nu < 0$.

The derivative with respect to γ is

$$\frac{d}{d\gamma} Z_{\nu+1}(\gamma, m) = (1 - \gamma m)^\nu \, e^{-\mu} \, (\nu + 1) \sum_{x=0}^{+\infty} (x - m) \, (1 + \theta x)^\nu \, \frac{\mu^x}{x!}, \tag{43}$$

that can also be expressed using hypergeometric functions. Formulas for domains $\mathcal{D}_2$ and $\mathcal{D}_3$ also involve hypergeometric functions. With these formulas, or by direct evaluation of (41), functionals $D_C(\gamma)$ and $D_G(\gamma)$ can be evaluated and maximized on their domains of definition so as to find the optimum value γ^*.

Given the signs of ν and γ, and the supports $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$, it is already possible to deduce that the solution γ^* is necessarily in a specific interval. Hence, we obtain here that for $\nu > 0$ (respectively for $\nu < 0$), solutions associated to a constraint $m > \mu$ corresponds to case (a) (resp. case (c)) and that solutions for $m < \mu$ correspond to case (c) (resp. case (a)). The argumentation relies on the fact that if P_i and P_j are two optimum distributions with supports $\mathcal{D}_i$ and $\mathcal{D}_j$, with the same (generalized) mean but different parameters, then by Theorem 1 $D_\alpha(P_j||Q) \geq D_\alpha(P_i||Q)$ if P_j is dominated by P_i.

In the case $m > \mu$, $\nu < 0$, the solution with minimum divergence is for a distribution P_3 in case (c), and furthermore we have $D_\alpha(P_3\,||Q) \to 0$. This can be seen as follows. Let $x \in \mathcal{D}_3$ and $k = \left\lfloor m - \frac{1}{\gamma} \right\rfloor$, so that $x < k+1$. Let now $\gamma = \frac{1}{m-k} + \epsilon$ with $\epsilon \in \left(0, \frac{1}{m-k-1} - \frac{1}{m-k}\right)$. Then the mean of the distribution is given by

$$E_\nu[X] = \frac{1}{Z_\nu(\gamma, m)} \sum_{x=0}^{k-1} x \left[\frac{k-x}{k-m}\right]^\nu Q(x) + k \frac{[k-m]^\nu}{Z_\nu(\gamma, m)} \epsilon^\nu Q(k), \tag{44}$$

and any value higher than $\mu = E_Q[X]$ can be obtained by tuning ϵ, for many values of k. When k increases, $\gamma = \frac{1}{m-k}$ tends to 0 by lower values and P_3 tends to Q, which results in $D_\alpha(P_3\,||Q) \to 0$.

The $\nu < 0$ case has the specificity that $Z_\nu(\gamma, m)$ exhibits singularities at $\gamma = \frac{1}{m-k}$ for all $k \geq 0$. Then $Z_\nu(\gamma, m)$, with $\nu = -(\xi+1)$ or $\nu = \xi$, is only convex on intervals $[\frac{1}{m-k}, \frac{1}{m-k-1}]$ or $[\frac{1}{m-k-1}, \frac{1}{m-k}]$ (for $k+1 > m > k$), with $Z_\nu(\gamma, m) = +\infty$ on the bounds of each interval. Consequently, $-\log Z_\nu(\gamma, m)$ may present several maxima. This is illustrated in Fig. 6 where function $D_C(\gamma)$ with $\alpha = 0.5$ presents many extrema. The solution with minimum Rényi divergence corresponds to the minimum of these maxima.

The limit case $\alpha \to 1$ is obtained with $|\nu| = |\xi| \to +\infty$. According to the discussion above, the optimum γ corresponds to case (a) for $\{m > \mu, \nu > 0\}$ and $\{m < \mu, \nu < 0\}$, and to case (c) for $\{m < \mu, \nu > 0\}$. For case (a), the support is $\mathcal{D}_1$, and the derivative of the partition function $Z_\nu(\gamma, m)$ is given by (43). In this derivative, the sum can be rewritten as

$$\sum_{x=0}^{+\infty} (x-m)(1+\theta x)^\nu \frac{\mu^x}{x!} = \sum_{x=0}^{+\infty} \left(\mu(1+\theta x+\theta)^\nu - m(1+\theta x)^\nu\right) \frac{\mu^x}{x!}, \tag{45}$$

so that $Z_{\nu+1}(\gamma, m)$ is minimum when the RHS of (45) is equal to zero. We have to solve this equation in θ. Suppose that θ is small and that $\theta x \ll 1$ for the significative values of the probability distribution. In this case, we use the approximation $(1+\theta x)^\nu = e^{\nu \log(1+\theta x)} \approx e^{\nu\theta x}$, that leads to

$$\sum_{x=0}^{+\infty} \left(\mu e^{\nu\theta(x+1)} - m e^{\nu\theta x}\right) \frac{\mu^x}{x!} = e^{\mu e^{\nu\theta}} \left(\mu e^{\nu\theta} - m\right) = 0 \tag{46}$$

The solution is given by $\theta^* = \frac{1}{\nu} \log(\frac{m}{\mu})$, that in turn provides

$$\gamma^* = \frac{\ln \frac{m}{\mu}}{\nu + m \ln \frac{m}{\mu}}. \tag{47}$$

In case (a), γ is positive, and this will be true for γ^* if $\{m > \mu, \nu > 0\}$ or $\{m < \mu, \nu < 0\}$. For the log-partition function, when $|\nu| \to +\infty$, this leads to

$$-\log\ Z_{\nu+1}(\gamma^*, m) \approx m \log \frac{m}{\mu} + (\mu - m). \tag{48}$$

In domain $\mathcal{D}_3$, the derivative of the partition function $Z_\nu(\gamma, m)$ is equal to zero if

$$\sum_{x=0}^{k} (x-m)(1+\theta x)^\nu \frac{\mu^x}{x!} = 0, \text{ with } k = \left\lfloor m - \frac{1}{\gamma} \right\rfloor, \gamma < 0.$$

If γ is small enough, $k \to +\infty$ and we obtain for $\nu > 0$ the same formulation and solution as in $\mathcal{D}_1$ The solution γ^* in (47) is now negative, that imposes $m < \mu$ for $\nu > 0$. Finally, we have shown above that if $m > \mu$ with $\nu < 0$ then $D_\alpha(P_3\,||Q) \to 0$.

Hence, we obtain that the entropy functionals converge to

$$\mathcal{F}^{(.)}_{\alpha\to 1}(x) = x \ln \frac{x}{\mu} + (\mu - x) \tag{49}$$

with the restriction that $\mathcal{F}^{(.)}_\alpha(x) = 0$ for $x > \mu$ if (C) $\alpha < 1$ or (G) $\alpha > 1$.

This functional is simply the cross-entropy between x and μ or Kullback-Leibler (Shannon) entropy functional with respect to μ [9]. It measures a 'distance' between a possible mean (observable) and a reference mean μ, and it

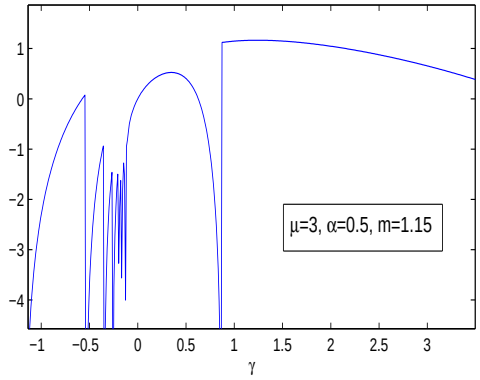

Fig. 6. Example of functional $D_C(\gamma)$ for the Poisson reference with classical mean constraint, with $\mu = 3, \alpha = 0.5$ ($\xi = -2$) and $m = 1.15$. It presents singularities at $1/(m-k)$, $\forall k$, and maxima at $\gamma = 0.35$ and $\gamma = 1.24$.

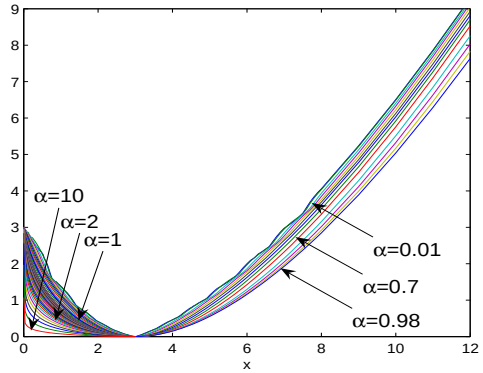

Fig. 7. Entropy functional $\mathcal{F}_\alpha^{(G)}(x)$ for a Poisson reference measure with $\mu = 3$ and $\alpha \geq 0$. For $\alpha \to 1$, $\mathcal{F}_\alpha^{(.)}(x)$ converges to $x \ln \frac{x}{\mu} + (\mu - x)$.

has been used as a regularization functional in several applied problems, such as astronomy, tomography, RMN, and spectrometry.

As in the previous cases, the entropy functionals $\mathcal{F}_\alpha^{(C)}(x)$ and $\mathcal{F}_\alpha^{(G)}(x)$ can be evaluated numerically. For instance, $\mathcal{F}_\alpha^{(G)}(x)$ is given on Fig. 7 for $\mu = 3$. It presents an unique minimum for $m = \mu$, and we note that it is is not convex for small values of α.

4. Conclusion and future work

By weakening one of the postulates that lead to the definition of Shannon entropy, Rényi [32] introduced a one parameter family of entropy and divergence. Shannon entropy and Kullback-Leibler divergence are recovered in the limiting case for the parameter $\alpha \to 1$. In this work, we considered the maximum entropy problems associated with Rényi Q-entropies. We characterized the solutions for a standard mean constraint and for the generalized mean constraint of nonextensive statistics. We defined and discussed the entropy functionals as a function of the constraints. These entropies were characterized and various properties and relationships were highlighted. We also discussed numerical aspects. Finally we illustrated this setting through some specific examples and recovered some well-kown entropy functionals.

Future work will consider the extension of this setting in the multivariate case. An issue that should be examined is the fact that the direct multivariate extension of (5) is not separable in the case of a separable reference $Q(x)$; which means that some dependances are implicitely introduced in the maximum entropy solution.

We also intend to investigate a possible underlying geometrical structure of the maximum entropy distributions (5). This structure should extend the geometrical structure of exponential families and involve the Bregman-like divergence introduced by (25).

Finally, maximum entropy methods have been successfully employed for solving inverse problems. We intend to consider the potential of Rényi entropies and divergence in this field. A simple contribution would be to examine the interest of a Rényi entropy functional, e.g. (39), as a potential in a Markov field for image deconvolution or restoration.

Appendix A. Proof of Theorem 1

Let us begin with the classical constraint (C). In this first case, we follow the approach of [37]. Consider the functional Bregman divergence :

$$B_h(f,g) = \int d(f,g)h(x)dx = \int -\left(f(x)^\alpha - g(x)^\alpha - \alpha\left(f(x) - g(x)\right)g(x)^{\alpha-1}\right)h(x)dx$$

where $h(x)$ is a nonnegative functional, associated to the (pointwise) Bregman divergence $d(f,g)$ built upon the strictly convex function $-x^\alpha$ for $\alpha \in (0,1)$. Then

$$B_{Q^{1-\alpha}}(P,P_C) = -\int_{\mathcal{S}} P(x)^\alpha - P_C(x)^\alpha - \alpha(P(x)P_C(x)^{\alpha-1} - P_C(x)^\alpha)Q(x)^{1-\alpha}dx \tag{A.1}$$

$$= -\int_{\mathcal{S}} P(x)^\alpha Q(x)^{1-\alpha}dx + \int_{\mathcal{S}} P_C(x)^\alpha Q(x)^{1-\alpha}dx. \tag{A.2}$$

with $h(x) = Q(x)^{1-\alpha}$ and where $\mathcal{S}$ denotes the support of $P_C(x)$. The second line follows from the fact that when P and P_C have the same mean $\bar{x} = E_{P_C}[X] = E_P[X]$, then using the expression in (5) with $\nu = \xi = \frac{1}{\alpha-1}$ it is possible to check that

$$\int_{\mathcal{S}} P(x)P_C(x)^{\alpha-1}Q(x)^{1-\alpha}dx = \int_{\mathcal{S}} P_C(x)^\alpha Q(x)^{1-\alpha}dx = Z_\xi(\gamma,\bar{x})^{-\alpha}$$

provided the whole support of $P(x)$ is included in $\mathcal{S}$, which is the case by the absolute continuity of $P(x)$ with respect to $P_C(x)$.

The Bregman divergence $B_{Q^{1-\alpha}}(P,P_C)$ being always positive and equal to zero if and only if $P = P_C$, the equality (A.2) implies that, for $\alpha \in (0,1)$,

$$D_\alpha(P||Q) \geq D_\alpha(P_C||Q) \tag{A.3}$$

which means that P_C is the distribution with minimum Rényi (Tsallis) divergence to Q, in the set of all distributions $P \ll P_C$ with a given mean $\bar{x}$, for $\alpha \in (0,1)$. The case $\alpha > 1$ can be derived accordingly, beginning with the Bregman divergence associated to the strictly convex function x^α.

As far as the generalized mean constraint (G) is concerned, let us now consider the Rényi information divergence $D_\alpha(P||P_G)$ from P to P_G, with P_G given in (5) with $\nu = -\xi = \frac{1}{1-\alpha}$

$$(\alpha-1)D_\alpha(P||P_G) = \log \int_{\mathcal{S}} P(x)^\alpha P_G(x)^{1-\alpha}dx, \tag{A.4}$$

with $\mathcal{S}$ the support of $P_G(x)$, and which can be rearranged as

$$(\alpha-1)D_\alpha(P||P_G) = \log \int_{\mathcal{S}} \frac{P(x)^\alpha Q(x)^{1-\alpha}}{\int_{\mathcal{S}} P(x)^\alpha Q(x)^{1-\alpha}dx}\left[\gamma(x-\overline{x})+1\right]dx \tag{A.5}$$

$$+ \log \int_{\mathcal{S}} P(x)^\alpha Q(x)^{1-\alpha}dx - (1-\alpha)\log Z_{\frac{1}{1-\alpha}}(\gamma,\overline{x}). \tag{A.6}$$

The generalized mean with respect to P appears in the first term, and cancels if P and P_G have the same generalized mean $\bar{x}$ and $P_G \gg P$. In such a case, we obtain

$$D_\alpha(P||P_G) = \frac{1}{(\alpha-1)}\log \int P(x)^\alpha Q^{1-\alpha}dx + \log Z_{\frac{1}{1-\alpha}}(\gamma,\overline{x}) \tag{A.7}$$

$$= D_\alpha(P||Q) - D_\alpha(P_G||Q), \tag{A.8}$$

where we used the fact that $D_\alpha(P_G||Q) = -\log Z_{\frac{1}{1-\alpha}}(\gamma,\overline{x})$ as stated in Proposition 5. Since the Rényi information divergence is always greater or equal to zero, we have

$$D_\alpha(P||Q) \geq D_\alpha(P_G||Q) \tag{A.9}$$

and conclude that P_G is the distribution with minimum Rényi (Tsallis) divergence to Q, in the set of all distributions $P \ll P_G$ with a given generalized α-mean $\bar{x}$.

Finally, it is easy to check, given the expression of P_G and the fact that $\alpha\xi = \xi + 1$, that the generalized mean of P_G is also the standard mean of the distribution with exponent $\nu = -(\xi+1)$, that is $E^{(\alpha)}_{P_G}[X] = E_{P^*_G}[X] = E_{-(\xi+1)}[X]$. Note that the equality in (A.8), $D_\alpha(P||Q) = D_\alpha(P||P_G) + D_\alpha(P_G||Q)$, is a pythagorean equality, which means that P_G is the orthogonal projection of P on the set of probability distributions with fixed generalized mean $\bar{x}$.

Appendix B. Proof of Proposition 6

The exact behaviour depends on the reference distribution $Q(x)$ and on the sign of the exponent ν. Because the domain of definition $\mathcal{D}$ might depend on γ, the derivative of the partition function writes

$$\frac{dZ_\nu(\gamma,\bar{x}_\gamma)}{d\gamma} = \lim_{\delta\gamma\to 0}\frac{1}{\delta\gamma}\left(Z_\nu(\gamma+\delta\gamma,\bar{x}_{\gamma+\delta\gamma}) - Z_\nu(\gamma,\bar{x}_\gamma)\right)$$

where $\bar{x}_\gamma$ and $\bar{x}_{\gamma+\delta\gamma}$ now denote the parameter $\bar{x}$ for distributions with parameter γ and $\gamma+\delta\gamma$. Let us begin with the continuous case. If $\delta\mathcal{D}$ denotes the domain increment associated to the variation $\delta\gamma$, it remains

$$\frac{dZ_\nu(\gamma,\bar{x}_\gamma)}{d\gamma} = \int_{\mathcal{D}}\frac{d}{d\gamma}\left(1+\gamma\left(x-\bar{x}_\gamma\right)\right)^\nu Q(x)dx \tag{B.1}$$

$$+\lim_{\delta\gamma\to 0}\frac{1}{\delta\gamma}\int_{\delta\mathcal{D}}\left(1+(\gamma+\delta\gamma)\left(x-\bar{x}_{\gamma+\delta\gamma}\right)\right)^\nu Q(x)dx \tag{B.2}$$

Of course, when $\mathcal{D}$ does not depend on γ, we only have the first term, and it is easy to obtain (14). Otherwise, in order to satisfy the positivity of the integrand, the domain $\mathcal{D}$ is bounded above by $\left(\bar{x}_\gamma - \frac{1}{\gamma}\right)$ for $\gamma<0$ and below by the same value for $\gamma>0$. Then, the second integral, say G, can be expressed as

$$G = \operatorname{sign}(\gamma)\int_{\bar{x}_{\gamma+\delta\gamma}-\frac{1}{\gamma+\delta\gamma}}^{\bar{x}_\gamma-\frac{1}{\gamma}}\left(1+(\gamma+\delta\gamma)\left(x-\bar{x}_{\gamma+\delta\gamma}\right)\right)^\nu Q(x)dx \tag{B.3}$$

$$= \frac{\operatorname{sign}(\gamma)}{\gamma+\delta\gamma}\int_0^a y^\nu Q\left(\frac{y-1}{\gamma+\delta\gamma}+\bar{x}_{\gamma+\delta\gamma}\right)\mathrm{d}y \tag{B.4}$$

with $a=(\gamma+\delta\gamma)\left(\bar{x}_{\gamma+\delta\gamma}-\bar{x}_\gamma\right)-\frac{\delta\gamma}{\gamma}$, that tends to zero with $\delta\gamma$ if $\bar{x}_\gamma$ is continuous. At first order, we then obtain

$$G = \operatorname{sign}(\gamma)\frac{Q\left(\bar{x}_{\gamma+\delta\gamma}-\frac{1}{\gamma+\delta\gamma}\right)}{\gamma+\delta\gamma}\int_0^a y^\nu \mathrm{d}y \propto \frac{a^{1+\nu}}{1+\nu}$$

for $\nu>-1$. Then, it is readily checked that $\lim_{\delta\gamma\to 0}\frac{1}{\delta\gamma}G=0$ for $\nu>0$, so that (B.2) is always zero for $\nu>0$ and (14) is true.

In the discrete case, the partition function is

$$Z_\nu(\gamma,\bar{x}_\gamma) = \sum_{x\in\mathcal{D}}\left(1+\gamma\left(x-\bar{x}_\gamma\right)\right)^\nu Q(x)$$

There exists singular isolated values of γ such that $1+\gamma\left(x-\bar{x}_\gamma\right)=0$, for x integer. For such values, the corresponding term in the partition function diverges for $\nu<0$. Contrary to the continuous case where the domain of γ is contiguous, the domain of values of γ ensuring that the partition function is finite will be interrupted by isolated values of γ: the domain of possible γ will be constituted of segments.

As in the continous case, the derivative of the partition function writes as the sum of two terms, the second one involving a domain increment

$$\frac{dZ_\nu(\gamma,\bar{x}_\gamma)}{d\gamma} = \sum_{\mathcal{D}}\frac{d}{d\gamma}\left(1+\gamma\left(x-\bar{x}_\gamma\right)\right)^\nu Q(x) \tag{B.5}$$

$$+\lim_{\delta\gamma\to 0}\frac{1}{\delta\gamma}\sum_{\delta\mathcal{D}}\left(1+(\gamma+\delta\gamma)\left(x-\bar{x}_{\gamma+\delta\gamma}\right)\right)^\nu Q(x) \tag{B.6}$$

If $\mathcal{D}$ does not depend on γ, there is no domain increment and the derivative is given by (B.5). When the bounds of $\mathcal{D}$ depend of γ, the domain increment is given by the integers in the interval $\left(\lceil\bar{x}_{\gamma+\delta\gamma}-\frac{1}{\gamma+\delta\gamma}\rceil,\lceil\bar{x}_\gamma-\frac{1}{\gamma}\rceil\right)$ ($\gamma>0$) or $\left(\lfloor\bar{x}_\gamma-\frac{1}{\gamma}\rfloor,\lfloor\bar{x}_{\gamma+\delta\gamma}-\frac{1}{\gamma+\delta\gamma}\rfloor\right)$ ($\gamma<0$); where $\lfloor x\rfloor$ is the floor function that returns the largest integer less than or equal to x; and $\lceil x\rceil$ is the ceil function, the smallest integer not less than x. If γ belongs in some interval such that

the domain increment remains empty, then the derivative is of course simply (B.5). An extension will occur for an infinitesimal variation $\delta\gamma$ if $\bar{x}_\gamma - \frac{1}{\gamma}$ is precisely an integer, say k,
Then, the second sum reduces to

$$G = \left(1 + (\gamma + \delta\gamma)\left(k - \bar{x}_{\gamma+\delta\gamma}\right)\right)^{\nu} Q(k) \tag{B.7}$$

$$= \left(-\frac{\delta\gamma}{\gamma} - (\gamma + \delta\gamma)\left(\bar{x}_{\gamma+\delta\gamma} - \bar{x}_\gamma\right)\right)^{\nu} Q(k), \tag{B.8}$$

and finally

$$\lim_{\delta\gamma\to 0} \frac{1}{\delta\gamma} G = \lim_{\delta\gamma\to 0} \delta\gamma^{\nu-1} \left((\gamma + \delta\gamma)\frac{\left(\bar{x}_{\gamma+\delta\gamma} - \bar{x}_\gamma\right)}{\delta\gamma} - \frac{1}{\gamma}\right)^{1+\nu} = 0 \quad \text{for } \nu > 1. \tag{B.9}$$

since all terms in the parenthesis remains finite when $\delta\gamma \to 0$. In such case the derivative reduces to (B.5) and (14) is true.

References

[1] M. Asadi, I. Bayramoglu, The mean residual life function of a k-out-of-n structure at the system level, IEEE Transactions on Reliability 55 (2006) 314–318.
[2] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res 6 (2005) 1705–1749.
[3] R. Baraniuk, P. Flandrin, A. Janssen, O. Michel, Measuring time-frequency information content using the Rényi entropies, IEEE Transactions on Information Theory 47 (2001) 1391–1409.
[4] A. G. Bashkirov, On maximum entropy principle, superstatistics, power-law distribution and Rényi parameter, Physica A 340 (2004) 153–162.
[5] M. Basseville, Distance measures for signal processing and pattern recognition, Signal Processing 18 (1989) 349–369.
[6] C. Beck, Generalized statistical mechanics of cosmic rays, Physica A 331 (2004) 173–181.
[7] D. Bhandari, N. R. Pal, Some new information measures for fuzzy sets, Information Sciences 67 (1993) 209–228.
[8] A. C. Cebrian, M. Denuit, P. Lambert, Generalized pareto fit to the society of actuaries' large claims database, North American Actuarial Journal 7 (2003) 18–36.
[9] I. Csiszár, Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems, Annals of Statistics 19 (1991) 2032–2066.
[10] I. Csiszár, Generalized cutoff rates and Rényi's information measures, IEEE Transactions on Information Theory 41 (1995) 26–34.
[11] R. S. Ellis, Entropy, Large Deviations, and Statistical Mechanics, vol. 271 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, 1985.
[12] M. D. Esteban, Divergence statistics based on entropy functions and stratified sampling, Information Sciences 87 (1995) 185–203.
[13] A. Golan, J. M. Perloff, Comparison of maximum entropy and higher-order entropy estimators, Journal of Econometrics 107 (2002) 195 – 211.
[14] M. Grendar, M. Grendar, Maximum entropy method with non-linear moment constraints: challenges, AIP, 2004.
[15] Y. He, A. Hamza, H. Krim, A generalized divergence measure for robust image registration, IEEE Transactions on Signal Processing, [see also Acoustics, Speech, and Signal Processing 51 (2003) 1211–1220.
[16] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 108 (1957) 171.
[17] E. T. Jaynes, On the rationale of maximum entropy methods, Proc. IEEE 70 (1982) 939–952.
[18] P. Jizba, T. Arimitsu, The world according to Rényi: thermodynamics of multifractal systems, Annals of Physics 312 (2004) 17–59.
[19] A. Krishnamachari, V. moy Mandal, Karmeshu, Study of dna binding sites using the rényi parametric entropy measure, Journal of Theoretical Biology 227 (2004) 429–436.
[20] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[21] B. LaCour, Statistical characterization of active sonar reverberation using extreme value theory, Oceanic Engineering, IEEE Journal of 29 (2004) 310–316.
[22] M. M. Mayoral, Rényi's entropy as an index of diversity in simple-stage cluster sampling, Information Sciences 105 (1998) 101–114.
[23] I. Molina, D. Morales, Rényi statistics for testing hypotheses in mixed linear regression models, Journal of Statistical Planning and Inference 137 (2007) 87–102.
[24] M. A. J. V. Montfort, J. V. Witter, Generalized Pareto distribution applied to rainfall depths, Hydrological Sciences Journal 31 (1986) 151–162.
[25] S. Nadarajah, K. Zografos, Formulas for Rényi information and related measures for univariate distributions, Information Sciences 155 (2003) 119–138.
[26] S. Nadarajah, K. Zografos, Expressions for Rényi and shannon entropies for bivariate distributions, Information Sciences 170 (2005) 173–189.
[27] A. K. Nanda, S. S. Maiti, Rényi information measure for a used item, Information Sciences 177 (2007) 4161–4175.
[28] J. Naudts, Dual description of nonextensive ensembles, Chaos, Solitons, and Fractals 13 (2002) 445–450.

[29] H. Neemuchwala, A. Hero, P. carson, Image matching using alpha-entropy measures and entropic graphs, Signal Processing 85 (2005) 277–296.
[30] R. Nock, F. Nielsen, On weighting clustering, IEEE Trans. Pattern Anal. Mach. Intell 28 (2006) 1223–1235.
[31] G. A. Raggio, On equivalence of thermostatistical formalisms, http://arxiv.org/abs/cond-mat/9909161 (1999).
[32] A. Rényi, On measures of entropy and information, Univ. California Press, Berkeley, Calif., 1961.
[33] K.-S. Song, Rényi information, loglikelihood and an intrinsic distribution measure, Journal of Statistical Planning and Inference 93 (2001) 51–69.
[34] C. Tsallis, Possible generalization of boltzmann-gibbs statistics, Journal of Statistical Physics 52 (1988) 479–487.
[35] C. Tsallis, Entropic nonextensivity: a possible measure of complexity, Chaos, Solitons,& Fractals 13 (2002) 371–391.
[36] C. Tsallis, R. S. Mendes, A. R. Plastino, The role of constraints within generalized nonextensive statistics, Physica A 261 (1998) 534–554.
[37] C. Vignat, A. Hero, J. A. Costa, About closedness by convolution of the Tsallis maximizers, Physica A 340 (2004) 147–152.
[38] S. Vinga, J. S. Almeida, Rényi continuous entropy of DNA sequences, Journal of Theoretical Biology 231 (2004) 377–388.