Entropies and entropic criteria

Jean-François Bercher

Chapitre 1

Entropies and entropic criteria

1.1. Introduction

This chapter focuses on the notions of entropy and of maximum entropy distribution which will be characterized according to different perspectives. Beyond links with applications in engineering and physics, it will be shown that it is possible to build regularization functionals based on the use of a maximum entropy technique, which can then possibly be employed as *ad hoc* potentials in data inversion problems.

The chapter begins with an overview of the key properties of information measures, and with the introduction of various concepts and definitions. In particular, the Rényi divergence is defined, the concept of escort distribution is presented, and the principle of maximum entropy that will be subsequently used will be commented on. A conventional engineering problem is then presented, the problem of source coding, and it shows the benefit of using measures with a different length than the standard measure, and in particular an exponential measure, which leads to a source coding theorem whose minimum bound is a Rényi entropy. It is also shown that optimal codes can be easily calculated with escort distributions. In Section 1.4, a simple state transition model is introduced and examined. This model leads to an equilibrium distribution defined as a generalized escort distribution, and as a by-product leads once again to a Rényi entropy. The Fisher information flow along the curve defined by the generalized escort distribution is examined and connections with the Jeffreys divergence are achieved. Finally, various arguments are obtained which, in this framework, lead to an inference method based on the minimization of the Rényi entropy under a generalized mean constraint, that is to say, taken with

---

Chapitre rédigé par Jean-François BERCHER.

regard to the escort distribution. From subsection 1.5.3, the main concern is about the minimization of the Rényi divergence subject to a generalized average constraint. The optimal density that solves this problem, and the value of the corresponding optimal divergence are given and characterized. The main properties of any entropy that may be related are defined and characterized. Finally, it is shown how to practically calculate these entropies and how it can be envisaged to use them for solving linear problems.

1.2. Some entropies in information theory

The concept of information plays a major role in a number of scientific and technical fields and in their applications. Moreover, information theory, "the Shannon way", meets the theories of physics, mutually fertilizing each other; these interactions have been exploited by Jaynes [JAY 57a, JAY 57b] since 1957, are discussed for example by Brillouin [BRI 62] and more recently in the fascinating work of [MER 10]. We will further give a simple model of phase transition that yields a Rényi entropy.

A fundamental question in information theory is of course the measure, or the definition, of the information. Several approaches are possible. The first is pragmatic and accepts as measure of valid information the measures that appear by themselves when solving a practical problem. The second is axiomatic, which starts with a certain number of reasonable properties or postulates, and then carries on with the mathematical derivation of the functions that exhibit these properties. This is the point of view adopted originally by Shannon, in his fundamental article [SHA 48a, SHA 48b], and that has led to a number of subsequent developments, among which [ACZ 75] and [ACZ 84] will be cited (where the author warns against the excesses of generalizations: *"I wish to urge here caution with regard to generalizations in general, and in particular with regard to those introduced through characterizations. (...) There is a large number of "entropies" and other "information measures" and their "characterizations", mostly formal generalizations of (1), (19), (16), (24), (17), (23) etc. popping up almost daily in the literature. It may be reassuring to know that most are and will in all probability be completely useless.*"

Similarly, Rényi himself [CSI 06, RÉN 65] stressed that only quantities that can actually be used in concrete problems should be considered as information measures, in agreement with the pragmatic approach (*As a matter of fact, if certain quantities are deduced from some natural postulates (from "first principles") these certainly need for their final justification the control whether they can be effectively used in solving concrete problems*).

1.2.1. *Main properties and definitions*

We will however remind here the main properties used for the caracterizations of information measures. If P, Q, R refer to discrete probability distributions for n events, with p_k the probability associated to the k-th event $k = 1, \ldots, n$, then noting $H(P) = H(p_1,p_2, \ldots, p_n)$ the information measure related to distribution events P, the main properties are as follows:

– (P1) symmetry: $H(p_1,p_2, \ldots, p_n)$ does not depend on the order of the events;

– (P2) $H(p,1-p)$ is a continuous function of p;

– (P3) $H(1/2,1/2) = 1$;

– (P4) recursion (branching):
$H_{n+1}(p_1q_1,p_1q_2,p_2, \ldots, p_n) = H_n(p_1,p_2, \ldots, p_n) + p_1H_2(q_1,q_2)$;

– (P5) expansibility: $H_{n+1}(p_1,p_2, \ldots, p_n,0) = H_n(p_1,p_2, \ldots, p_n)$;

– (P6) subadditivity: $H(PQ) \leqslant H(P) + H(Q)$
(and additivity in the independent case: $H(PQ) = H(P) + H(Q)$);

– (P7) conditional subadditivity: $H(PQ|R) \leqslant H(P|R) + H(Q|R)$;

– (P8) generalized recursion:
$H_{n+1}(p_1q_1,p_1q_2,p_2, \ldots, p_n) = H_n(p_1,p_2, \ldots, p_n) + m(p_1)H_2(q_1,q_2)$.

Simple consequences

The first four postulates are Faddeev's axioms [FAD 56], that suffice to uniquely characterize the Shannon entropy:

$$H(P) = -\sum_{i=1}^{n} p_i \ln p_i \qquad [1.1]$$

If the recursion postulate is evaluated but an additivity requirement is added, then the class of possible solutions is much wider, and includes in particular the Rényi entropy, which will be referred further in the text. The replacement of the P4 recursion by a general recursion postulate, P8, with $m(p_1p_2)$ multiplicative $m(p_1p_2) = m(p_1)m(p_2)$ and especially $m(p) = p^q$ leads to the entropy of order q:

$$H_q(P) = \frac{1}{2^{1-q} - 1} \left(\sum_{i=1}^{n} p_i^q - 1 \right) \qquad [1.2]$$

which was introduced by [HAV 67], independently by Daróczy [DAR 70], and then rediscovered in the field of statistical physics by C. Tsallis [TSA 88]. For $q \geqslant 1$, these entropies are subadditive, but are not additive. In the case of $q = 1$, by l'Hôpital's rule, the entropy of order $q = 1$ is none other than the Shannon entropy. In statistical physics, a significant community has formed around the study of non-extensive thermodynamics (non-additive in fact) [TSA 09] based on the use of the

Tsallis entropy, on associated maximum entropy distributions and on the extension of classical thermodynamics.

In Faddeev's axiomatics, Rényi [RÉN 61] has proposed to replace the recursion postulate by the additivity property, and add a property of mean entropy, which specifies that the entropy of the union of two incomplete probability distributions is equal to the weighted average of the two entropy distributions. When the mean being used is an arithmetic mean, the only solution is the Shannon entropy. On the other hand, by using an exponential mean, the entropy that appears is a Rényi entropy:

$$H_q(P) = \frac{1}{1-q} \ln \sum_{i=1}^{n} p_i^q \qquad [1.3]$$

Another way to apprehend the Rényi entropy is to note that the Shannon entropy is the arithmetic mean, with weights p_i, of the basic information $I_i = -\ln p_i$ associated to the different events. By replacing the arithmetic mean by a Kolmogorov-Nagumo average, the entropy becomes:

$$H_\psi(p_1, \ldots, p_n) = \psi^{-1}\left(\sum p_i \psi(-\ln p_i)\right)$$

Under an additional additivity condition and under the condition $\lim_{p\to 0} H_\psi(p, 1-p) = 0$, this entropy is either the Shannon entropy, the Rényi entropy, with $q \geqslant 0$. Again, by l'Hospital's rule, the Shannon entropy is met once again for $q = 1$. Furthermore, for $q = 0$, the Rényi entropy becomes the Hartley entropy, the logarithm of the number of events of non-zero probability.

1.2.2. *Entropies and divergences in the continuous case*

In the continuous case, the definition used for the Shannon entropy associated with a density $f(x)$ is:

$$H[f] = -\int f(x) \ln f(x)\, \mathrm{d}x \qquad [1.4]$$

However, it should be noted that this expression only results from the transition to the limit of the discrete case up to an additive constant tending to infinity (see for example [PAP 81]). Therefore, the concern is rather about differential entropy. However, Jaynes has lucidly noted that, since [JAY 63, p. 202], it is necessary to introduce a measure $m(x)$ accounting for "points density" shifting the procedure to the limit; this measure conferring in addition a coordinate change invariance to the resulting information, which is not the case of [1.4]. The corresponding differential entropy then takes the form:

$$H[f] = -\int f(x) \ln \frac{f(x)}{m(x)}\, \mathrm{d}x \qquad [1.5]$$

This form is similar to a Kullback-Leibler divergence [KUL 59] (or I-divergence in Csiszár's terminology) between two probability distributions with densities $f(x)$ and $g(x)$ relatively to a common measure $\mu(x)$, and which is defined by:

$$D(f||g) = \int f(x) \ln \frac{f(x)}{g(x)} \, \mathrm{d}\mu(x) \qquad [1.6]$$

by assuming g absolutely continuous relatively to f, and with the convention $0 \ln 0 = 0$. When g is uniform, with respect to μ, the Kullback divergence becomes, in absolute value, a μ-entropy. In the case where μ is the Lebesgue measure, the differential Shannon [1.5] entropy appears once again; in the discrete case, if μ is the counting measure, then [1.1] will appear again. It is easily shown, by application of Jensen's inequality that the Kullback divergence is defined as non-negative, $D(f||g) \geqslant 0$ with equality if and only if $f = g$. It can thus be understood as a distance between distributions, although it is not symmetric and does not check not the triangle inequality.

In the same way, continuous versions of Rényi and Tsallis entropies can be defined. For an entropy index $q \neq 1$:

$$S_q[f] = \frac{1}{1-q} \left(\int f(x)^q \, \mathrm{d}\mu(x) - 1 \right) \qquad [1.7]$$

is the Tsallis entropy and:

$$H_q[f] = \frac{1}{1-q} \ln \int f(x)^q \, \mathrm{d}\mu(x) \qquad [1.8]$$

that of Rényi. These two entropies are tantamount to the Shannon entropy for $q = 1$. Divergence can also be associated to them; for example, the Rényi divergence:

$$D_q(f||g) = \frac{1}{q-1} \ln \int f(x)^q g(x)^{1-q} \, \mathrm{d}\mu(x) \qquad [1.9]$$

which is also defined as non-negative (by Jensen's inequality), and is reduced to the Kullback divergence when $q \to 1$.

1.2.3. *Maximum entropy*

The principle of maximum entropy is widely used in physics, and can rely on a large number of arguments: counts, axioms, etc. The principle has been particularly highlighted by Jaynes [JAY 57a] "*Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally*

noncommittal with regard to missing information", and we will confine ourselves here to recall its relevance in terms of statistics, by following Ellis [ELL 99] (theorem 2 for example).

If f_N is the empirical distribution corresponding to the collection of N random variables according to a density distribution g relatively to μ, then the probability Q of finding f_N in a set $\mathcal{B}$ is roughly (see Ellis [ELL 99] for more correct formulations), and for large N:

$$Q\left(f_N \in \mathcal{B}\right) \approx \exp\Big(-N \inf_{P\in\mathcal{B}} D(f||g)\Big) \qquad [1.10]$$

It can be thus derived, by iterating reasoning on subsets of $\mathcal{B}$, that the absolutely predominant distribution in $\mathcal{B}$ is the one that achieves the minimum Kullback distance to g: there is concentration of all the probability on the closest distribution to g. A minimum distance Kullback principle can thus be derived, or equivalently, if g is uniform, a principle of maximum entropy. Among all the distributions of a set $\mathcal{B}$, the density that minimizes $D(f||g)$ should be selected. When the point of interest is, as in statistical physics, the probability of finding an empirical mean x_N, that is to say, the mean under f_N, in a set $\mathcal{C}$, then a result with large level 1 deviations is obtained, which indicates that:

$$Q\left(x_N \in \mathcal{C}\right) \approx \exp\Big(-N \inf_{x\in\mathcal{C}} \mathcal{F}(\boldsymbol{x})\Big) \qquad [1.11]$$

where $\mathcal{F}(\boldsymbol{x})$ is the rate function $\mathcal{F}(\boldsymbol{x}) = \inf_{P:\boldsymbol{x}=\mathrm{E}[\boldsymbol{X}]} D(P||\mu)$. This result suggests thus to select the most probable element, that which achieves the minimum of $\mathcal{F}(\boldsymbol{x})$ on $\mathcal{C}$. The shift from a problematics of distributions to a problematics of means is known as the contraction principle.

1.2.4. *Escort distributions*

We will also use in the remainder of this chapter the notion of escort distribution. These escort distributions have been introduced as a tool in the context of multifractals [BEC 93, CHH 89], with interesting connections with standard thermodynamics. Escort distributions are proving useful in source coding, where they enable optimal code words to be obtained whose mean length is bounded by a Rényi entropy [BER 09]. This is what we will present in 1.3.3. We will then find these escort distributions in the framework of a state transition problem, Section 1.4.

If $f(x)$ is a probability density, then its escort of order $q \geqslant 0$ is:

$$f_q(x) = \frac{f(x)^q}{\int f(x)^q \,\mathrm{d}\mu(x)} \qquad [1.12]$$

provided that the informational generating function $M_q[f] = \int f(x)^q \, \mathrm{d}\mu(x)$ is finite. We can easily see that if $f_q(x)$ is the escort of $f(x)$, then $f(x)$ is itself the escort of order $1/q$ of $f_q(x)$. When q decreases, the escort comes closer to a uniform distribution whereas when q increases, density modes are amplified. This can be specified: as a matter of fact, it can be shown, in the compact support case that $D(f_q||U) > D(f||U)$ for $q > 1$, and that $D(f_q||U) < D(f||U)$ for $q < 1$, which means that f_q is further away from the uniform than f when $q > 1$ and closer otherwise.

The concept of escort distribution can also be expanded in order to take into account two densities $f(x)$ and $g(x)$ according to:

$$f_q(x) = \frac{f(x)^q g(x)^{1-q}}{\int f(x)^q g(x)^{1-q} \, \mathrm{d}\mu(x)} \qquad [1.13]$$

when $M_q[f, g] = \int f(x)^q g(x)^{1-q} \, \mathrm{d}\mu(x) < \infty$. This generalized escort distribution is simply a weighted geometrical mean of $f(x)$ et $g(x)$. Of course, if $g(x)$ is a uniform measure whose support includes that of $f(x)$, then the generalized escort is reduced to the standard escort [1.12]. This generalized escort appears in the analysis of the effectiveness of hypotheses tests [CHE 52] and allows the best possible exponent to be defined in the error probability [COV 06, chapitre 11]. When q varies, the generalized escort describes a curve that connects $f(x)$ and $g(x)$. Finally, we will call generalized moments the moments taken with respect to an escort distribution: the generalized of order p associated to the standard escort of order q will be:

$$m_{p,q}[f] = \int |x|^p f_q(x) \, \mathrm{d}x = \frac{\int |x|^p f(x)^q \, \mathrm{d}\mu(x)}{\int f(x)^q \, \mathrm{d}\mu(x)} \qquad [1.14]$$

1.3. Source coding with escort distributions and Rényi bounds

In this section, the advantage of the Rényi entropy and escort distributions is illustrated within the framework of source coding, one of the fundamental problems of information theory. After a very brief reminder of the context of source coding, a source coding theorem is described linking a new measure of mean length and the Rényi entropy. It is then shown that it is possible to practically calculate the optimal codes by using the concept of escort distribution. Details about these elements as well as other results are given in [BER 09].

1.3.1. *Source coding*

In source coding, considering a set $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ of symbols generated by a source with respective probabilities p_i where $\sum_{i=1}^{N} p_i = 1$. The role of source

coding is to associate to each symbol x_i a code word c_i, with length l_i, expressed with an alphabet of D elements. It is well known that if the lengths verify Kraft-Mac Millan inequality:

$$\sum_{i=1}^{N} D^{-l_i} \leqslant 1 \qquad [1.15]$$

then there exists a uniquely decodable code with these elementary lengths. In addition, any uniquely decodable code satisfies Kraft-Mac Millan inequality [1.15]. Shannon's source coding theorem indicates that the mean length $\bar{L}$ of word codes is bounded from below by the source entropy, $H_1(p)$, and that the best uniquely decodable code satisfies:

$$H_1(p) \leqslant \bar{L} = \sum_i p_i l_i < H_1(p) + 1 \qquad [1.16]$$

where the logarithm used in the Shannon entropy is calculated in base D, and noted $\log_D$. This result indicates that the Shannon entropy $H_1(p)$ is a fundamental limit to the minimal mean length for any code built for the source. The lengths of the optimal word codes are given by:

$$l_i = -\log_D p_i \qquad [1.17]$$

The characteristic of these optimal codes is that they assign the shortest words to the most probable symbols and the longest words to the rarest symbols.

1.3.2. *Source coding with Campbell measure*

It is well known that the Huffman algorithm provides a prefix code that minimizes the mean length and approaches the optimal length limits $l_i = -\log_D p_i$. However other forms of length measurement have also been considered. In particular the first, that of Campbell [CAM 65], is fundamental. It has been seen, by relation [1.17], that the lowest probabilities lead to longer word codes. However, the cost of using a code is not necessarily a linear function of its length, and it is possible that the addition of a letter to a long word is much more expensive than the addition of the same letter to a short word. This led Campbell to propose a new measure of mean length, by introducing an exponential penalty of the lengths of the word codes. This length, the Campbell length, is a generalized Kolmogorov-Nagumo average associated with an exponential function:

$$C_\beta = \frac{1}{\beta} \log_D \sum_{i=1}^{N} p_i D^{\beta l_i} \qquad [1.18]$$

with $\beta > 0$. The remarkable result of Campbell is that, in the same way as the Shannon entropy places a lower bound on the average length of the word codes, the Rényi

entropy of order q, with $q = 1/(\beta + 1)$, is the lower bound of the mean Campbell length [1.18]:

$$C_\beta \geqslant H_q(p) \qquad [1.19]$$

A simple demonstration of the result is given in [BER 09]. It is easy to see that equality is obtained for:

$$l_i = -\log_D P_i = -\log_D \left(\frac{p_i^q}{\sum_{j=1}^N p_j^q} \right) \qquad [1.20]$$

Clearly, the lengths l_i obtained in this way can be made smaller than the optimal Shannon's lengths, by choosing a quite small parameter q, which then tends to standardize the distribution, then actually enhancing the the lowest probabilities. Thus, the procedure effectively penalizes the longest word codes and provides word codes of different lengths than Shannon's, with eventually shorter word codes associated with the low probabilities.

1.3.3. *Source coding with escort mean*

For the usual mean length measure $\bar{L} = \sum_i p_i l_i$, there is a linear combination of the elementary length, weighted by probabilities p_i. In order to increase the impact of the most important lengths associated with low probabilities, the Campbell length uses an exponential of the elementary lengths. Another idea is to modify the weights in the linear combination, such that to increase the importance of the words with low probabilities. A simple way to achieve this is to standardize the initial probability distribution, and to use the weights achieved by this new distribution rather than p_i. Naturally, this leads to use a mean taken with an escort distribution:

$$M_q = \sum_{i=1}^N \frac{p_i^q}{\sum_j p_j^q} l_i = \sum_{i=1}^N P_i l_i \qquad [1.21]$$

In the case of the imaginary source that would have a distribution P the standard statistics mean is M_q, and Shannon's classical source coding theorem can immediately be applied:

$$M_q \geqslant H_1(P) \qquad [1.22]$$

with equality if:

$$l_i = -\log_D P_i \qquad [1.23]$$

or exactly the lengths [1.20] obtained for Campbell's measure. The simple relation $l_i = -\log_D P_i$ obtained for the minimization of M_q under the constraint supplied by Kraft-Mac Millan's inequality has an immediate but important application. As a matter of fact, it simply suffices to provide the escort distribution P rather than the initial distribution p to an standard encoding algorithm, for example a Huffman algorithm, to obtain an optimized code for the Campbell length C_β, or in a similar manner, for the measurement of length M_q. Table 1.1 gives a simple example with $D = 2$: we have used a standard Huffman algorithm, with the initial distribution, then its escorts of order $q = 0.7$ and $q = 0.4$.

p_i	$q = 1$	$q = 0{,}7$	$q = 0{,}4$
0,48	0	0	00
0,3	10	10	01
0,1	110	1100	100
0,05	1110	1101	101
0,05	11110	1110	110
0,01	111110	11110	1110
0,01	111111	11111	1111

Tableau 1.1. *Example of binary codes, for different values of q*

It is important to note that specific algorithms have been developed for the mean Campbell length. The above connection provides an easy and immediate alternative. Another important point is that these codes have practical applications: they are optimal for the minimization of the probability of buffer overflowing [HUM 81] or, with $q > 1$, for maximizing the probability of receiving a message in single send of limited size.

1.4. A simple transition model

In the previous section, we have seen emerging, and appreciated the significance of the Rényi entropy and escort distributions for a source coding problem. In this section, we will show that these two quantities are also involved in an equilibrium, or a transition model framework, between two states. It has actually been noted

that extended thermodynamics, associated to the Tsallis and Rényi entropies, seems particularly relevant in the case of deviations from the conventional Boltzmann-Gibbs equilibrium. This suggests then to amend the conventional formulation of the conventional approach of maximum entropy (or of the the minimum of divergence) and to imagine an equilibrium characterized by two (and no longer a single) distributions: rather than selecting the nearest distribution of a reference distribution under a mean constraint, an intermediary distribution $p_q(x)$ is desired, in a sense that needs clarification, between two references $p_0(x)$ et $p_1(x)$. This construction, as well as some of its consequences, are also described in [BER 12].

1.4.1. *The model*

Considering two density states with probabilities $p_0(x)$ and $p_1(x)$ at point x of the phase space, and searching for an intermediate state according to the following scenario. The initial state system p_0, subject to a generalized force, is moved and held at a distance $\eta = D(p||p_0)$ of p_0. On the other hand, the system is attracted towards a final state p_1. As a result, the new intermediate state p_q is chosen such that it minimizes its divergence from the attractor p_1 while being maintained at a distance η of p_0. As illustrated in figure 1.1, the intermediate probability density is "aligned" with p_0 and p_1 and at the intersection with the set $D(p||p_0) = \eta$, a circle of radius η centered on p_0. More specifically, by taking densities relatively to the Lebesgue measure, the problem can be formulated as follows:

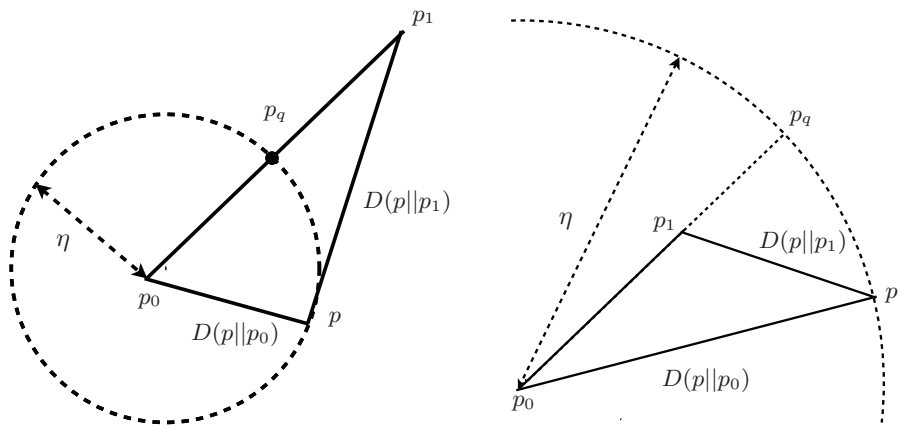
$$\begin{cases} \min_p \ D(p||p_1) \\ \text{under } D(p||p_0) = \eta \\ \text{and } \int p(x)\,\mathrm{d}x = 1 \end{cases} \qquad [1.24]$$

The solution is given by the following proposition.

PROPOSITION 1.1.– *If q is a real positive that $D(p_q||p_0) = \eta$ and if $M_q(p_1,p_0) = \int p_1(x)^q p_0(x)^{1-q}\,\mathrm{d}x < \infty$, then, the solution of the problem [1.24] is given by:*

$$p_q(x) = \frac{p_1(x)^q p_0(x)^{1-q}}{\int p_1(x)^q p_0(x)^{1-q}\,\mathrm{d}x} \qquad [1.25]$$

REMARQUE.– When p_0 is uniform and with compact support, the standard escort distribution [1.12] is met once again. If media is not compact and uniform distribution improper, it is possible to simply change the formulation by taking as a constraint a fixed entropy $H(p) = -\eta$, and then the escort distribution is obtained.



(a) Cas $\eta < D(p_1||p_0)$ (b) Cas $\eta > D(p_1||p_0)$

Figure 1.1. *Equilibrium between the states p_0 and p_1: the search for the equilibrium distribution is carried out in all of the distributions within a difference fixed at p_0, $D(p||p_0) = \eta$, and at a minimal Kullback distance of p_1. The resulting equilibrium distribution p_q, the generalized escort distribution is "aligned" with p_0 and p_1, and at the intersection of the set $D(p||p_0) = \eta$.*

Let us evaluate the divergence $D(p||p_q)$. For all densities p such that constraint $D(p||p_0) = \eta$ is satisfied, it yields:

$$\begin{aligned} D(p||p_q) &= \int p(x) \ln \frac{p(x)}{p_q(x)} \, \mathrm{d}x = \int p(x) \ln \frac{p(x)^q p(x)^{1-q}}{p_1(x)^q p_0(x)^{1-q}} \, \mathrm{d}x + \ln M_q(p_1,p_0) \\ &= q \int p(x) \ln \frac{p(x)}{p_1(x)} \, \mathrm{d}x + (1-q) \int p(x) \ln \frac{p(x)}{p_0(x)} \, \mathrm{d}x + \ln M_q(p_1,p_0) \\ &= q \, D(p||p_1) + (1-q)\eta + \ln M_q(p_1,p_0) \end{aligned} \quad [1.26]$$

By taking $p = p_q$, the last equality becomes:

$$D(p_q||p_q) = q \, D(p_q||p_1) + (1-q)\eta + \ln M_q(p_1,p_0) \quad [1.27]$$

Finally by subtracting [1.26] and [1.27], it gives:

$$D(p||p_q) - D(p_q||p_q) = q \, (D(p||p_1) - D(p_q||p_1)) \quad [1.28]$$

Since $q \geqslant 0$ and $D(p||p_q) \geqslant 0$ with equality if and only if $p = p_q$, it finally yields $D(p||p_1) \geqslant D(p_q||p_1)$ which proves the proposition 1.1.

When η varies, the function $q(\eta)$ is increasing, with still $D(p_q||p_0) = \eta$. For $\eta = 0$ it gives $q = 0$ and for $\eta = D(p_1||p_0)$ it gives $q = 1$. Therefore, when q varies, p_q defines a curve that links p_0 ($q = 0$) to p_1 (q =1), and beyond for $q > 1$, see figure 1.1.

REMARQUE.– It is interesting to also note that results have shown that work dissipated during a transition can be expressed as a Kullback-Leibler divergence [PAR 09]. In this context, with a Halmitonian pair following the impulse, the constraint $D(p||p_k) = \eta$, $k = 0$ where 1, can be interpreted as a bound on the average work dissipated during the transition from p to p_k.

1.4.2. *The Rényi divergence as a consequence*

Finally, it is interesting to note that the Rényi divergence appears as a byproduct of our construction. As a matter of fact, as a direct consequence of [1.27] and of the definition of the Rényi divergence [1.9], the minimum Kullback information can be expressed as:

$$D(p_q||p_1) = \left(1 - \frac{1}{q}\right)(\eta - D_q(p_1||p_0)) \qquad [1.29]$$

By taking a uniform measure for p_0, the Rényi entropy is revealed.

$$D(p_q||p_1) = \left(1 - \frac{1}{q}\right)(\eta + H_q[p_1]) \qquad [1.30]$$

The Kullback-Leibler divergence is not symmetrical. Since the beginning, Kullback and Leibler have introduced a symmetrical version, returning again to the Jeffreys divergence. In our case, this Jeffreys divergence is a simple affine function of the Rényi divergence:

$$J(p_1,p_q) = D(p_1||p_q) + D(p_q||p_1) = \frac{(q-1)^2}{q}(D_q(p_1||p_0) - \eta) \qquad [1.31]$$

This equality is a simple consequence of the relation [1.26], with $p = p_1$, and the relation [1.27]. It can be noted, as a significant consequence, that the minimization of the Jeffreys divergence between p_1 and p_q under certain constraints, is therefore equivalent to the minimization of the Rényi divergence with the same constraints.

1.4.3. *Fisher information for the parameter* q

The generalized escort distribution p_q defined a curve indexed by q linking distributions p_0 and p_1 for $q = 0$ and $q = 1$. It is interesting to evaluate the attached information to the parameter q of the generalized distribution. This Fisher information is given by:

$$I(q) = \int \frac{1}{p_q(x)}\left(\frac{dp_q(x)}{dq}\right)^2 \mathrm{d}x = \int \frac{dp_q(x)}{dq} \ln \frac{p_1(x)}{p_0(x)}\, \mathrm{d}x \qquad [1.32]$$

where the right term is obtained using the relation:

$$\frac{\mathrm{d}p_q(x)}{\mathrm{d}q} = p_q(x)\left(\ln\frac{p_1(x)}{p_0(x)} - \frac{\mathrm{d}\ln M_q}{\mathrm{d}q}\right) \quad [1.33]$$

and the fact that:

$$\int \frac{\mathrm{d}p_q(x)}{\mathrm{d}q}\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}q}\int p_q(x)\,\mathrm{d}x = 0$$

by Leibniz's rule. It can also be shown that this Fisher information is equal to the variance, with respect to the distribution p_q, of the likelihood ratio.

Finally, it is possible to identify the integral of the Fisher information along the curve, the "energy" of the curve, at the Jeffreys divergence. More specifically, the following proposal is given.

PROPOSITION 1.2.– *The integral of the Fisher information, from $q = r$ to $q = s$ is proportional to the Jeffreys divergence between p_r and p_s:*

$$(s-r)\int_r^s I(q)\,\mathrm{d}q = J(p_s,p_r) = D(p_s||p_r) + D(p_r||p_s) \quad [1.34]$$

With $r = 0$ et $s = 1$, it therefore yields that:

$$\int_0^1 I(q)\,\mathrm{d}q = J(p_1,p_0) = D(p_1||p_0) + D(p_0||p_1) \quad [1.35]$$

To demonstrate [1.34], it is sufficient to integrate [1.32]:

$$\begin{aligned}\int_r^s I(q)\,\mathrm{d}q &= \int_r^s \int \frac{\mathrm{d}p_q(x)}{\mathrm{d}q}\ln\frac{p_1(x)}{p_0(x)}\;\mathrm{d}x\;\mathrm{d}q \\ &= \int (p_s(x) - p_r(x))\ln\frac{p_1(x)}{p_0(x)}\;\mathrm{d}x\end{aligned}$$

Taking into account the fact that $\ln p_s/p_r = (s-r)\ln p_1/p_0$, we then get [1.34].

Finally, if θ_i, $i = 1..M$ is a set of intensive variables depending on q, then $\frac{d\ln p}{dq} = \sum_{i=1}^{M}\frac{\partial \ln p}{\partial\theta_i}\frac{d\theta_i}{dq}$ and the Fisher information q can be expressed according to the Fisher information matrix of θ. In these conditions, and for the generalized escort distribution, the result is that the "thermodynamic divergence" for the transition is none other than the Jeffreys divergence [1.35]:

$$\mathcal{J} = \int_0^1 I(q)\,\mathrm{d}q = \sum_{i=1}^{M}\sum_{j=1}^{M}\int_0^1 \frac{\mathrm{d}\theta_i}{\mathrm{d}q}\left[I(\theta)\right]_{i,j}\frac{\mathrm{d}\theta}{\mathrm{d}q}\;\mathrm{d}q = D(p_1||p_0)+D(p_0||p_1) \quad [1.36]$$

1.4.4. *Distribution inference with generalized moment constraint*

Assuming now that the p_1 distribution is imperfectly known, but that additional information is available under the form of a mean value, achieved with distribution p_q. This mean is the generalized mean [1.14], which is used in non-extensive statistical physics; It has here the clear interpretation of a mean obtained from the equilibrium distribution p_q. The problem that arises now is then the determination of the most general distribution compatible with this constraint.

The idea of minimizing the divergence to p_1 can be retained as in the problem [1.24] which has led us to the equilibrium distribution with generalized escort. Since the Kullback divergence is directed, the direction will be retained by minimizing $D(p_q||p_1)$ for $q < 1$ and $D(p_1||p_q)$ for $q > 1$. In both cases, the divergence is expressed as an affine function of the Rényi divergence $D_q(p_1||p_0)$, see [1.29], and these minimizations are finally equivalent to the minimization of the Rényi divergence under the generalized mean constraint.

Similarly, the concern could be about the minimization of the symmetric Jeffreys divergence between p_q and p_1. However, we have noted in [1.31] that this is also expressed as a simple affine function of the Rényi divergence: Its minimization is therefore equivalent to the minimization of the Rényi divergence under a generalized mean constraint.

Finally, the Jeffreys divergence $J(p_1,p_q)$ is proportional to the thermodynamic divergence, the integral of the Fisher information, as shown in [1.34], for $q > 1$ as well as for $q < 1$. Therefore, the minimization of the thermodynamic divergence between p_q and p_1 is also equivalent to the minimization of the Rényi divergence.

These different arguments very legitimately lead us to search for distribution p_1 as the distribution minimizing the Rényi divergence of index q, under the generalized mean constraint.

1.5. Minimization of the Rényi divergence and associated entropies

In previous paragraphs we have described a framework enabling the Rényi information, the escort distributions and generalized moments to appear naturally. In addition, we in have derived an inference distribution method: the minimization of the Rényi information, with information available in the form of generalized moments. In this section, we will first give the expression for the density that minimizes the Rényi divergence, then we will describe some properties of the associated partition functions. Finally, we will show how new entropic functionals can be derived of which a few examples will be given. Some of these results, but also some extensions can be referred to in [BER 08].

1.5.1. *Minimisation under generalized moment constraint*

We will first consider a generalized moment of any order [1.14], whose expression is mentioned below:

$$m_{p,q}[f] = \int |x|^p f_q(x)\,\mathrm{d}\mu(x) = \frac{\int |x|^p f(x)^q g(x)^{1-q}\,\mathrm{d}\mu(x)}{\int f(x)^q g(x)^{1-q}\,\mathrm{d}\mu(x)} \qquad [1.37]$$

The problem is then considered:

$$\mathcal{F}_q(m) = \left\{ \begin{array}{c} \min_f D_q(f||g) \\ \text{under} \quad m = m_{p,q}[f] \\ \text{and} \quad \int f(x)\,\mathrm{d}\mu(x) = 1 \end{array} \right. \qquad [1.38]$$

The minimum obtained is of course a function of m, that will be noted $\mathcal{F}_q(m)$. It is a contracted version of the Rényi divergence, which defines an "entropy" in the space of possible means m. In [BER 11], we have considered a more general problem, in which the indices of the generalized moment and of the Rényi divergence are not identical. In any case, the result here obtained is as follows.

PROPOSITION 1.3.– *The density G_γ which achieves the minimum in the problem [1.38] is given by:*

$$G_\gamma(x) = \frac{1}{Z_\nu(\gamma,\bar{x}_p)} \left(1 - (1-q)\gamma\left(|x|^p - \bar{x}_p\right)\right)_+^\nu g(x) \qquad [1.39]$$

or equivalently by:

$$G_{\bar{\gamma}}(x) = \frac{1}{Z_\nu(\bar{\gamma})} \left(1 - (1-q)\bar{\gamma}|x|^p\right)_+^\nu g(x) \qquad [1.40]$$

with $\nu = 1/(1-q)$, $\bar{x}_p$ an eventual translation parameter, γ and $\bar{\gamma}$ selected scaling parameters chosen such that the generalized constraint moment is satisfied, and finally

where $(x)_+ = \max(0,x)$. *Quantities* $Z_\nu(\gamma,\bar{x}_p)$ *and* $Z_\nu(\bar{\gamma})$ *are partition functions that allow the standardization of the density. For* $q = 1$, *density* $G_\gamma(x)$ *becomes an exponential density:*

$$G_\gamma(x) = \frac{1}{Z_\nu(\gamma)} \exp\left(-\gamma\left(|x|^p - \bar{x}_p\right)\right) g(x) \qquad [1.41]$$

with respect to $g(x)$.

In the case $p = 2$, a Gaussian density is thus found once again. Density G_γ is sometimes called "generalized Gaussian". It should noted once more that γ and $\bar{\gamma}$ are determined by the relation:

$$\bar{\gamma} = \frac{\gamma}{1 + \frac{\gamma}{\nu}\bar{x}_p} \qquad [1.42]$$

In the case of expression [1.40], the demonstration is here proposed. The approach is rather similar in the case of density [1.39].

As in [BER 11], let $A(\bar{\gamma}) = 1/Z(\bar{\gamma})$. It immediately yields:

$$\int f^q G_{\bar{\gamma}}^{1-q}\, \mathrm{d}\mu(x) = A(\bar{\gamma})^{1-q} M_q[f,g] \times \int \left(1 - (1-q)\bar{\gamma}|x|^p\right)_+ \frac{f^q g^{1-q}}{M_q[f,g]}\, \mathrm{d}\mu(x)$$

$$\geqslant A(\bar{\gamma})^{1-q}\left(1 - (1-q)\bar{\gamma}\, m_{p,q}[f]\right) M_q[f,g] \qquad [1.43]$$

with $M_q[f,g] = \int f^a g^{1-q}\, \mathrm{d}\mu(x)$, where $m_{p,q}[f]$ refers to the generalized, and where the inequality results from the fact that the support $(1-(1-q)\bar{\gamma}|x|^p)_+$can be included in that of $f^q g^{1-q}$. From [1.43] it directly gives, with $f = G_{\bar{\gamma}}$:

$$M_1[G_{\bar{\gamma}}] = 1 = A(\bar{\gamma})^{1-q}\left(1 - (1-q)\bar{\gamma}\, m_{q,p}[G_{\bar{\gamma}}]\right) M_q[G_{\bar{\gamma}}, g] \qquad [1.44]$$

Thus, for all distributions f of generalized $m_{p,q}[f] = m$ and for $\bar{\gamma}$ such that $G_{\bar{\gamma}}$ has the same moment $m_{p,q}[G_{\bar{\gamma}}] = m$, then the combination of [1.43] and [1.44] results in:

$$\int f^q G_{\bar{\gamma}}^{1-q}\, \mathrm{d}\mu \geqslant \frac{M_q[f,g]}{M_q[G_{\bar{\gamma}}, g]}$$

Finally the Rényi divergence of order q can thus be expressed as:

$$D_q(f||G_{\bar{\gamma}}) = \ln\left(\int f^q G_{\bar{\gamma}}^{1-q}\, \mathrm{d}\mu(x)\right)^{\frac{1}{q-1}} \qquad [1.45]$$

$$\leqslant \ln\left(\frac{M_q[f,g]}{M_q[G_{\bar{\gamma}}, g]}\right)^{\frac{1}{q-1}} = D_q(f||g) - D_q(G_{\bar{\gamma}}||g) \qquad [1.46]$$

By the non-negativity of the divergence, it thus ensues that:

$$D_q(f||g) \geqslant D_q(G_{\tilde{\gamma}}||g) \qquad [1.47]$$

for all distributions f of generalized $m_{p,q}[f] = m_{p,q}[G_{\tilde{\gamma}}] = m$, and with equality if and only if $f = G_{\tilde{\gamma}}$.

1.5.2. *A few properties of the partition functions*

Some important properties of partition functions $Z_\nu(\gamma,\bar{x}_p)$ associated to the optimal density G_γ are given here (see [BER 08]). These properties will be essential for the characterization of entropic functionals $\mathcal{F}_q(x)$. E_ν refers to the statistic mean taken relatively to the optimum density distribution [1.39], with $\nu = 1/(1-q)$. It is also important to realize, from now on, that the escort density of order q of [1.39] is none other than this same density G_γ but with an exponent $\nu - 1$, such that:

$$m_{p,q}[G_{\tilde{\gamma}}] = E_{\nu-1}[X] \qquad [1.48]$$

The successive partition functions are linked by:

$$Z_\nu(\gamma,\overline{x}_p) = E_{\nu-1}\left[1 - \frac{\gamma}{\nu}\left(|x|^p - \overline{x}_p\right)\right] Z_{\nu-1}(\gamma,\overline{x}) \qquad [1.49]$$

As a direct result, it can be seen that $Z_\nu(\gamma,\overline{x}_p) = Z_{\nu-1}(\gamma,\overline{x}_p)$ if and only if $\overline{x}_p = E_{\nu-1}\left[|X|^p\right]$.

Using Leibniz's rule, the derivative with respect to γ can be obtained and is given by:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma,\overline{x}_p) = \left(-E_{\nu-1}\left[|X|^p - \overline{x}_p\right] + \gamma\frac{\mathrm{d}\overline{x}_p}{\mathrm{d}\gamma}\right) Z_{\nu-1}(\gamma,\overline{x}_p) \qquad [1.50]$$

under the condition that $\overline{x}_p$ is really differentiable with respect to γ. Similarly:

$$\frac{\mathrm{d}}{\mathrm{d}\overline{x}_p} Z_\nu(\gamma,\overline{x}_p) = \left(-\frac{\mathrm{d}\gamma}{\mathrm{d}\overline{x}_p} E_{\nu-1}\left[|X|^p - \overline{x}_p\right] + \gamma\right) Z_{\nu-1}(\gamma,\overline{x}_p) \qquad [1.51]$$

Thus, if $\overline{x}_p = E_{\nu-1}\left[|X|^p\right]$, then taking into account the equality of the partition functions of rank ν and $\nu - 1$, it gives:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \ln Z_\nu(\gamma,\overline{x}_p) = \gamma\frac{\mathrm{d}\overline{x}_p}{\mathrm{d}\gamma} \qquad [1.52]$$

or even:

$$\frac{\mathrm{d}}{\mathrm{d}\overline{x}_p} \ln Z_\nu(\gamma,\overline{x}_p) = \gamma \qquad [1.53]$$

On the other hand, when $\overline{x}_p$ is an independent parameter of γ, say $\overline{x}_p = m$, then:

$$\frac{\mathrm{d}^2 Z_\nu(\gamma,m)}{\mathrm{d}\gamma^2} = \frac{\nu-1}{\nu} E_{\nu-2}\left[(X-m)^2\right] Z_{\nu-2}(\gamma,m) \qquad [1.54]$$

and similarly:

$$\frac{\mathrm{d}^2 Z_\nu(\gamma,m)}{\mathrm{d}m^2} = \frac{\nu-1}{\nu} \gamma^2 E_{\nu-2}\left[(X-m)^2\right] Z_{\nu-2}(\gamma,m) \qquad [1.55]$$

which, considering the fact that $(\nu-1)/\nu = q > 0$, the fact that partition functions are strictly positive, shows that if $\overline{x}_p = m$ and γ are independent, then the partition $Z_\nu(\gamma,m)$ is convex in its two variables.

Finally, the solution of the problem [1.38] can be expressed, that is $\mathcal{F}_q(m)$, from the partition function. By direct calculation, it actually yields:

$$D_q(G_\gamma||g) = \frac{1}{q-1} \ln Z_{q\nu}(\gamma,\bar{x}_p) - \frac{q}{q-1} \ln Z_\nu(\gamma,\bar{x}_p) \qquad [1.56]$$

which is simply reduced to:

$$\mathcal{F}_q(m) = D_q(G_\gamma||g) = -\ln Z_\nu(\gamma,m) = -\ln Z_{\nu-1}(\gamma,m) \qquad [1.57]$$

for the value of γ such that the constraint is satisfied, or $m_{p,q}[G_{\bar{\gamma}}] = E_{\nu-1}[X] = \bar{x}_p = m$.

1.5.3. *Entropic functionals derived from the Rényi divergence*

Thus, the solution of the minimization problem of the Renyi divergence of order q viewed as a function of constraint, defines an "entropic functional". Different functionals will be associated with the several specifications of the reference density $g(x)$, as well as with the various values of the index q. We will see that the functions in question present interesting properties. Therefore, a set of functions is potentially available that can be eventually used as objective functions or regularization terms.

The main characterization of $\mathcal{F}_q(m)$ is as follows.

PROPOSITION 1.4.– *The entropy $\mathcal{F}_q(m)$, defined by [1.38], is non-negative, with a single minimum m_g, the average of g, and $\mathcal{F}_q(m_g) = 0$. The entropy is a pseudoconvex function for $q \in [0,1)$ and strictly convex for $q \geqslant 1$.*

The Rényi divergence $D_q(f||g)$ is always non-negative, and zero only for $f = g$. Since functionals $\mathcal{F}_q(m)$ are defined as the minimum of the divergence $D_q(f||g)$, they are always non-negative. Based on [1.53], it gives $\frac{d}{d\overline{x}} \ln Z_\nu(\gamma,\overline{x}) = \gamma$. Thus,

functionals $\mathcal{F}_q(x)$ are only presenting a single singular point in $\gamma = 0$. For this value of γ, it yields $G_{\gamma=0} = g$, and $D_q(g||g) = 0$. Under these conditions, $\mathcal{F}_q(x)$ has a unique minimum for $x = m_g$, the mean of g, and $\mathcal{F}_q(m_g) = 0$. Therefore, it follows that $\mathcal{F}_q(x)$ is unimodal and does not present any points of inflection with a horizontal tangent; this is sufficient to claim that $\mathcal{F}_q(x)$ is pseudo-convex, as referred to by Mangasarian [MAN 87]. Let us now examine the convexity for $q \geqslant 1$. If f_q is the generalized escort distribution given by [1.13], then the equality $D_q(f||g) = D_{1/q}(f_q||g)$ holds. Subsequently, searching for the distribution f that achieves the minimum of $D_q(f||g)$ with a generalized mean constraint, that is to say, taken with respect to f_q, is equivalent to searching the distribution f_q that minimizes $D_{1/q}(f_q||g)$, under a standard moment constraint. In these circumstances, given p_1 and p_2 the two densities that minimize $D_{1/q}(f_q||g)$ under constraints $x_1 = E_{f_q}[X]$ and $x_2 = E_{f_q}[X]$. Then, $\mathcal{F}_q(x_1) = D_{1/q}(p||g)$, and $\mathcal{F}_q(x_2) = D_{1/q}(p_2||g)$. In the same way, given $\mathcal{F}_q(\mu x_1 + (1-\mu)x_2) = D_{1/q}(\hat{P}||Q)$, where $\hat{P}$ is the optimal escort distribution of mean $\mu x_1 + (1-\mu)x_2$. Distributions $\hat{P}$ and $\mu p_1 + (1-\mu)p_2$ then have the same mean. Thus, when $D_{1/q}(f_q||g)$ is a strictly convex function f_q, that is to say for $q \geqslant 1$ it follows that $D_{1/q}(\hat{P}||g) < \mu D_{1/q}(p_1||g) + (1-\mu)D_{1/q}(p_2||g)$, or $\mathcal{F}_q(\mu x_1 + (1-\mu)x_2) < \mu\mathcal{F}_q(x_1) + (1-\mu)\mathcal{F}_q(x_2)$ and entropy $\mathcal{F}_q(x)$ is a strictly convex function.

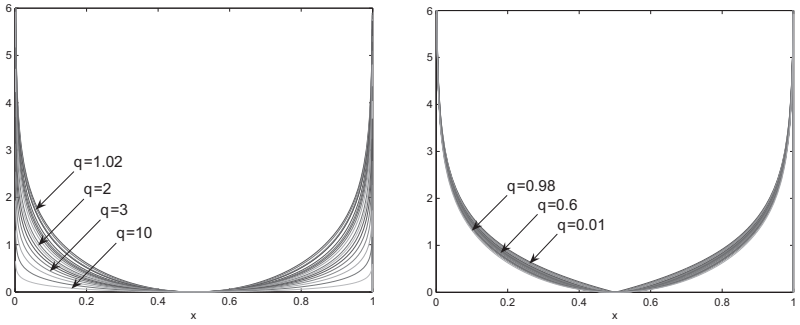
Even provided with this interesting characterization, a practical significant problem still remains: how to analytically or numerically determine the entropies $\mathcal{F}_q(x)$ for a reference density g and a given index entropy q. The problem amounts to determine the parameter γ such that the optimal generalized mean density [1.39] has a specified value m. A simple manner of proceeding consist in recalling the fact that if $\bar{x}_p$ is a fixed parameter m, independent of γ, then the derivative relation [1.50] is reduced to:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma,m) = \left(m - E_{\nu-1}\left[|X|^p\right]\right) Z_{\nu-1}(\gamma,m) \qquad [1.58]$$

Therefore, it can be seen that it suffices to search for the extrema of the partition function $Z_\nu(\gamma,m)$ to obtain a γ such that $m = E_{\nu-1}\left[|X|^p\right]$. Since we have seen that $Z_\nu(\gamma,m)$, with fixed m, is strictly convex, then this extremum is unique and is a minimum. Finally, the value of the entropy is simply given by [1.57]: $\mathcal{F}_q(m) = -\ln Z_\nu(\gamma,m)$.

The search for the expression of $\mathcal{F}_q(m)$ therefore requires to compute the partition function then to solve $\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma,m) = 0$ with respect to γ. With the exception of a few special cases, this resolution does not seem analytically possible, and entropy $\mathcal{F}_q(m)$ is given implicitly. In the particular case where g is a Bernoulli measure, it is possible to obtain an analytic expression for $\mathcal{F}_q(m)$, this for any $q > 0$. For other reference densities g, it is possible to obtain analytic expressions when $q \to 1$. These points are detailed in [BER 08], where in addition different densities of reference g are

studied, and corresponding entropies numerically evaluated according to the scheme previously mentioned. As an example, the numerical results obtained in the case where $p = 1$ and a uniform density in the interval $[0,1]$ are given in Fig. 1.2. For $q \geqslant 1$, a family of convex functions is correctly obtained over $(0,1)$, minimum for the mean of g, or 0.5, as we have indicated above. For $q < 1$, a family of non-negative, unimodal functions, also minimal in $x = m_g = 0.5$ is obtained.



(a) Case $q > 1$ (b) Case $q \in (0,1)$

Figure 1.2. *Entropy $\mathcal{F}_q(x)$ for a uniform reference, respectively for $q \geqslant 1$ and $q \in]0,1[$*

1.5.4. *Entropic criteria*

Based on the previous figures, it is apparent that the minimization of $\mathcal{F}_q(x)$ under certain constraints automatically provides a solution in the interval $(0,1)$. In addition, the parameter q can be used to adjust the curvature of the function or the penalty on the boundaries of the domain. It is thus interesting to use these entropies when the purpose is to solve inverse problems. More specifically, an entropy criterion can be used such as $\mathcal{F}_q(x)$ as objective function. The whole of this section will be restricted to the case $p = 1$. When considering a linear inverse problem $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, with x_k the components of $\boldsymbol{x}$, then this can be formulated as:

$$\begin{cases} \min_{\boldsymbol{x}} \sum_k \mathcal{F}_q(x_k) \\ \text{under } \boldsymbol{y} = \mathbf{A}\boldsymbol{x} \end{cases} \qquad [1.59]$$

This then corresponds to select among possible solutions the solution whose components are of minimum entropy. It should be noted that it is assumed here, implicitly, that the criterion was separable into its components. In reality, if we define $\mathcal{F}_q(\boldsymbol{x})$ as the Rényi divergence under the generalized mean constraint, then, even when assuming that components are independent, it yields a density on $\boldsymbol{x}$ similar to [1.39], which is not separable. In order to obtain a separable criterion, which is both more

consistent with intuition and easier to use, we amend the formulation by searching the density product that achieves the minimum of the Rényi divergence under generalized mean constraint, which leads effectively to the separable criterion. Thus, the previous problem [1.59] can also be read as:

$$\left\{ \begin{array}{l} \min_{\boldsymbol{x}} \left\{ \begin{array}{c} \min_{\boldsymbol{f}} D_q(\boldsymbol{f}||\boldsymbol{g}) \\ \text{under } \boldsymbol{f} = \prod_k f_k \\ \text{and } \boldsymbol{x} = E_{\boldsymbol{f}_q}[X] \end{array} \right. \\ \text{under } \boldsymbol{y} = \mathbf{A}\boldsymbol{x} \end{array} \right. \tag{1.60}$$

where $E_{\boldsymbol{f}_q}[X]$ refers to the generalized mean, that is to say, taken with respect to the escort distribution of order q. The point of concern here is therefore a "maximum entropy" problem which consists in selecting a solution $\boldsymbol{x}$, seen as the generalized mean of a minimum Rényi divergence distribution with a reference density g. Entropies $\sum_k \mathcal{F}_q(x_k)$ being pseudo-convex, it is known that minimization under linear constraints leads to a single minimum (see for example [CAM 08, theorem 4.4.1]). Now let us examine how it is possible to obtain a solution of [1.59], even in the case where the entropies have no explicit expression. The solution corresponds to a stationary point of the Lagrangian $L(\boldsymbol{\lambda},\boldsymbol{x})$ associated to the problem [1.59], and the objective is therefore to solve:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda},\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \sum_k \mathcal{F}_q(x_k) + \boldsymbol{\lambda}^{\mathrm{t}} (\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) \tag{1.61}$$

$$= \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \sum_k \mathcal{F}_q(x_k) - c_k x_k + \boldsymbol{\lambda}^{\mathrm{t}} \boldsymbol{y} \tag{1.62}$$

with $c_k = \left[\boldsymbol{\lambda}^{\mathrm{t}} \mathbf{A}\right]_k$. Using the fact that:

$$\mathcal{F}_q(x_k) = -\ln Z_\nu(\gamma_*,x_k) = -\inf_{\gamma} \ln Z_\nu(\gamma,x_k)$$

as we have seen in Subsection 1.5.3, it thus gives:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda},\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \sum_k -\ln Z_\nu(\gamma_*,x_k) - c_k x_k + \boldsymbol{\lambda}^{\mathrm{t}} \boldsymbol{y} \tag{1.63}$$

$$= \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\mathrm{t}} \boldsymbol{y} - \sum_k \max_{x_k} \left(\ln Z_\nu(\gamma_*,x_k) + c_k x_k\right) \tag{1.64}$$

However, by the relation [1.53], it follows:

$$\frac{\mathrm{d}}{\mathrm{d}x_k} \left(\ln Z_\nu(\gamma_*,x_k) + c_k x_k\right) = \gamma_* + c_k \tag{1.65}$$

which yields $\gamma_* = -c_k$, and x_k is the associated generalized mean. Finally, the concern is therefore about solving:

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\mathrm{t}} \boldsymbol{y} - \sum_k \left(\ln Z_\nu(-c_k,x_k) + c_k x_k\right) \tag{1.66}$$

where, for any c_k, the corresponding generalized mean x_k can be calculated as the unique solution of the problem:

$$x_k = \arg\min_x \left(\ln Z_\nu(-c_k,x) + c_k x\right)$$

It is thus possible to solve the problem [1.59] which provides a unique "maximum Rényi entropy" solution to the inverse linear problem $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, problem where various constraints can be included, including support, through the reference density g, and where the form of the criteria can be adjusted by means of the index entropy q.

In the case where data $\boldsymbol{y}$ would be imperfect, it is possible to minimize the entropy criterion under a constraint provided by a statistic (for example, residual χ^2) rather than with an exact constraint. It is also possible to use the entropy criterion with a data-fidelity term.

In the case where $q = 1$, the Rényi divergence is reduced to the Kullback divergence, the generalized moments to the usual moments, and the optimal density to an exponential density [1.41] with respect to g. Under these conditions, the log-partition function is written $\ln Z_\infty(-c_k,x_k) = -c_k x_k + \ln \int \exp\left(c_k x_k\right) g(x_k)\, \mathrm{d}\mu(x_k)$, the problem [1.66] becomes:

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \sum_k \ln \int \exp\left(c_k x_k\right) g(x_k)\, \mathrm{d}\mu(x_k)$$

and the optimal solution is given by the derivative of the log-partition function with respect to c_k. This latter approach has been developed in works supervised by Guy Demoment [LEB 99].

1.6. Bibliographie

[ACZ 75] ACZÉL J., DAROCZY Z., *On measures of information and their characterizations*, Academic Press, 1975.

[ACZ 84] ACZÉL J., "Measuring information beyond communication theory—Why some generalized information measures may be useful, others not", *Aequationes Mathematicae*, vol. 27, n° 1, p. 1-19, mars 1984.

[BEC 93] BECK C., SCHLOEGL F., *Thermodynamics of Chaotic Systems*, Cambridge University Press, 1993.

[BER 08] BERCHER J.-F., "On some entropy functionals derived from Rényi information divergence", *Information Sciences*, vol. 178, n° 12, p. 2489-2506, juin 2008.

[BER 09] BERCHER J.-F., "Source coding with escort distributions and Rényi entropy bounds", *Physics Letters A*, vol. 373, n° 36, p. 3235-3238, août 2009.

[BER 11] BERCHER J.-F., "Escort entropies and divergences and related canonical distribution", *Physics Letters A*, vol. 375, n° 33, p. 2969-2973, août 2011.

[BER 12] BERCHER J.-F., "A simple probabilistic construction yielding generalized entropies and divergences, escort distributions and q-Gaussians", *Physica A: Statistical Mechanics and its Applications*, vol. 391, n° 19, p. 4460-4469, oct. 2012.

[BRI 62] BRILLOUIN L., *Science and Information Theory*, Academic Press, 2e édition, juin 1962.

[CAM 65] CAMPBELL L. L., "A coding theorem and Rényi's entropy", *Information and Control*, vol. 8, n° 4, p. 423-429, 1965.

[CAM 08] CAMBINI A., MARTEIN L., *Generalized convexity and optimization*, Springer, nov. 2008.

[CHE 52] CHERNOFF H., "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", *Annals of Mathematical Statistics*, vol. 23, n° 4, p. 493-507, 1952.

[CHH 89] CHHABRA A., JENSEN R. V., "Direct determination of the $f(\alpha)$ singularity spectrum", *Physical Review Letters*, vol. 62, n° 12, p. 1327, mars 1989.

[COV 06] COVER T. M., THOMAS J. A., *Elements of Information Theory*, Wiley-InterScience, 2e édition, juil. 2006.

[CSI 06] CSISZÁR I., "Stochastics: Information theory", HORVÁTH J. (DIR.), *A Panorama of Hungarian Mathematics in the Twentieth Century I*, vol. 14, p. 523-535, Springer, Berlin, Heidelberg, 2006.

[DAR 70] DARÓCZY Z., "Generalized information functions", *Information and Control*, vol. 16, n° 1, p. 36-51, mars 1970.

[ELL 99] ELLIS R. S., "The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence", *Physica D*, vol. 133, n° 1-4, p. 106-136, sep. 1999.

[FAD 56] FADDEEV D., "On the concept of entropy of a finite probabilistic scheme", *Uspekhi Matematicheskikh Nauk*, vol. 11, n° 1(67), p. 227-231, 1956, (en Russe).

[HAV 67] HAVRDA J., CHARVÁT F., "Quantification method of classification processes. concept of structural a-entropy", *Kybernetika*, vol. 3, p. 30-35, 1967.

[HUM 81] HUMBLET P., "Generalization of Huffman coding to minimize the probability of buffer overflow", *IEEE Transactions on Information Theory*, vol. 27, n° 2, p. 230-232, 1981.

[JAY 57a] JAYNES E. T., "Information theory and statistical mechanics", *Physical Review A*, vol. 106, n° 4, p. 620-630, mai 1957.

[JAY 57b] JAYNES E. T., "Information theory and statistical mechanics. II", *Physical Review A*, vol. 108, n° 2, p. 171-190, oct. 1957.

[JAY 63] JAYNES E. T., "Information theory and statistical mechanics", *1962 Brandeis Summer Institute in Theoretical Physics*, vol. 3, p. 182-218, K. W. Ford, W. A. Benjamin Inc., New York, 1963, reprinted in: E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, éd. R. D. Rosencrantz, Synthèse Library, Vol. 138, Reidel, 1983.

[KUL 59] KULLBACK S., *Information Theory and Statistics*, John Wiley & Sons, New York, 1959, Republished by Dover Publications, 1997.

[LEB 99] LE BESNERAIS G., BERCHER J.-F., DEMOMENT G., "A new look at entropy for solving linear inverse problems", *IEEE Transactions on Information Theory*, vol. 45, n° 5, p. 1565-1578, 1999.

[MAN 87] MANGASARIAN O. L., *Nonlinear Programming*, SIAM, jan. 1987.

[MER 10] MERHAV N., "Statistical physics and information theory", *Foundations and Trends in Communications and Information Theory*, vol. 6, n° 1-2, p. 1-212, 2010.

[PAP 81] PAPOULIS A., "Maximum entropy and spectral estimation: A review", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, n° 6, p. 1176-1186, déc. 1981.

[PAR 09] PARRONDO J. M. R., VAN DEN BROECK C., KAWAI R., "Entropy production and the arrow of time", *New Journal of Physics*, vol. 11, n° 7, juil. 2009.

[RÉN 61] RÉNYI A., "On measures of entropy and information", *4th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, Etats-Unis, p. 547-561, 1961.

[RÉN 65] RÉNYI A., "On the foundations of information theory", *Review of the International Statistical Institute*, vol. 33, n° 1, p. 1-14, 1965.

[SHA 48a] SHANNON C., "A mathematical theory of communication", *The Bell System Technical Journal*, vol. 27, n° 3, p. 379-423, 1948.

[SHA 48b] SHANNON C., "A mathematical theory of communication", *The Bell System Technical Journal*, vol. 27, n° 4, p. 623-656, 1948.

[TSA 88] TSALLIS C., "Possible generalization of Boltzmann-Gibbs statistics", *Journal of Statistical Physics*, vol. 52, n° 1, p. 479-487, juil. 1988.

[TSA 09] TSALLIS C., *Introduction to Nonextensive Statistical Mechanics*, Springer, avr. 2009.