Entropy **2008**, *10*, 131-149; DOI: 10.3390/entropy-e10030131

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.org/entropy

Article

Generalised Exponential Families and Associated Entropy Functions

Jan Naudts

Department of Physics, University of Antwerpen, Groenenborgerlaan 171, 2020 Antwerpen, Belgium
E-mail: jan.naudts@ua.ac.be

Received: 26 February 2008; in revised form: 1 July 2008 / Accepted: 14 July 2008 / Published: 16 July 2008

Abstract: A generalised notion of exponential families is introduced. It is based on the variational principle, borrowed from statistical physics. It is shown that inequivalent generalised entropy functions lead to distinct generalised exponential families. The well-known result that the inequality of Cramér and Rao becomes an equality in the case of an exponential family can be generalised. However, this requires the introduction of escort probabilities.

Keywords: generalised entropy, maximum entropy principle, variational principle, generalised exponential family, Bregman divergence, generalised Fisher information, escort probability.

---

1. Introduction

Generalised entropy functions have been studied intensively in the second half of the past century. They have been called quasi-entropies in [1]. Every entropy function is in fact minus a relative entropy, also called a divergence. It is relative to some reference measure c. Consider the f-divergence [2, 3]

$$I(p||c) = \sum_a c_a f(p_a/c_a), \tag{1}$$

with $f(u)$ a convex function defined for $u > 0$ and strictly convex at $u = 1$. It is minus the entropy of p, relative to c. Taking $c_a = 1$ for all a and $f(u) = u \ln u$ one obtains the Boltzmann-Gibbs-Shannon entropy

$$I(p) = -\sum_a p_a \ln p_a. \tag{2}$$

Note that throughout the paper discrete probabilities are considered, with events a belonging to a finite or countable alphabet A.

Recent interest in these generalised entropies within statistical physics goes back to the introduction by Tsallis [4] of the q-entropy

$$I_q(p) = \frac{1}{1-q}\left(\sum_a p_a^q - 1\right), \tag{3}$$

with $q > 0$. In the limit $q = 1$ it converges to (2). It has been studied before in the mathematics literature by Havrda and Charvat [5], and by Daróczy [6]. Investigations within the physics community have lead to some interesting developments. One of them is the introduction of deformed logarithmic and exponential functions [7, 8] — see the Section 13. They have been very useful to generalise common concepts, like that of an exponential family or of a Gaussian distribution. They also helped to clarify the pitfalls of the generalisation process. One of the surprises is the necessity to introduce escort probability functions [9] — see Section 11. In a series of papers, including [10, 11], the present author has elaborated a formalism based on deformed logarithms. In the present work, it is shown that slightly more general results are obtained when abandoning these deformed logarithms.

Independent of the developments in statistical physics was the progress made in the context of game theory. A link, known to exist between maximising entropy and minimising losses [12], was generalised to arbitrary entropies by Grünwald and Dawid [13]. This lead to the introduction of the notion of generalised exponential families, notion which is also essential in [11], and which extents Lafferty's notion of additive models [14].

In Sections 2 to 6 the maximum entropy principle and the variational principle are discussed in the context of generalised entropies. In particular, a characterisation of the maximising probability distributions is given. This is used in Section 7 to define a generalised exponential family. In Section 8 it is shown that the intersection of distinct generalised exponential families is empty and that there exists a one-to-one relation with generalised entropy functions. Sections 9 tot 12 discuss geometric aspects, starting with concepts from thermodynamics and introducing escort families and a generalised Fisher information matrix. Sections 13 and 14 discuss non-extensive thermostatistics and the percolation problem as examples of the generalised formalism. The paper ends with a short discussion in Section 15.

2. Generalised entropies

Let us fix some further notations. The space of probability distributions is denoted $\mathcal{M}_1^+(A)$. Expectation values are denoted $\langle p, X\rangle = \sum_{a\in A} p_a X(a)$. Here we follow the physics tradition to put the elements of the dual space at the l.h.s..

It is rather common to define a generalised entropy as any function $I(p)$ of the form

$$I(p) = \sum_{a\in A} h\left(p_a\right), \tag{4}$$

where $h(u)$ is a continuous strictly concave function, defined on $[0, 1]$, which vanishes when $u = 0$ or $u = 1$. This is a special case of minus the f-divergence (1), with weights $c_a = 1$. The entropy function

$I(p)$ is defined for any $p \in \mathcal{M}_1^+(A)$ and has values in $[0, +\infty]$. In the present paper it is allowed that the function $h(u)$ is stochastic, this means, depends also on a in A. But for convenience of notation, this dependence will not be made explicit.

Throughout the paper it is assumed that the derivative

$$\frac{\mathrm{d}h}{\mathrm{d}u} = -f(u) \tag{5}$$

exists on the interval $(0, 1)$ and defines a continuous function on the halfopen interval $(0, 1]$. Because $h(u)$ is strictly concave, $f(u)$ is strictly increasing. Note that it is allowed to diverge to $-\infty$ at $u = 0$. This is indeed the case when $h(u) = -u \ln u$ and $f(u) = 1 + \ln u$.

The function $f(u)$ can be used to rewrite the entropy $I(p)$ as

$$I(p) = \sum_{a \in A} \int_{p_a}^{1} \mathrm{d}u\, f(u) = -\sum_{a \in A} \int_{0}^{p_a} \mathrm{d}u\, f(u) = -\sum_{a \in A} p_a \int_{0}^{1} \mathrm{d}v\, f(p_a v). \tag{6}$$

Note that the latter expression implies that

$$I(p) \geq -\sum_{a \in A} p_a f(p_a). \tag{7}$$

The standard definition of the Bregman divergence [15] reads (see for instance Section 3 of [14])

$$D(p||q) = I(q) - I(p) - \sum_{a \in A} (p_a - q_a) f(q_a). \tag{8}$$

In the case that $f(u)$ diverges at $u = 0$ it is only well defined when $q_a = 0$ implies $p_a = 0$. It is a convex function of the first argument. Note that one can write

$$D(p||q) = \sum_{a \in A} \int_{q_a}^{p_a} \mathrm{d}u\, \left[f(u) - f(q_a) \right]. \tag{9}$$

From the latter expression it is immediately clear that $D(p||q) \geq 0$, with equality if and only if $p = q$.

3. Maximum entropy principle

Let be given a finite number of real functions $H_1(a)$, $H_2(a)$, $\cdots$, $H_n(a)$. Assume they are bounded from below. In a physical context these functions may be called Hamiltonians. The maximum entropy problem deals with finding the probability distribution p that maximises $I(p)$ under the constraint that the expectation values of the Hamiltonians H_j attain given values U_j, called energies. Introduce the notation

$$\mathcal{P}_U = \{p \in \mathcal{M}_1^+ : \langle p, H_j \rangle = U_j, j = 1, 2, \cdots, n\}. \tag{10}$$

Then one looks for the probability distribution $p \in \mathcal{P}_U$ which maximises $I(p)$.

Definition 1 *A probability distribution $p^* \in \mathcal{P}_U$ is said to satisfy the maximum entropy principle if it satisfies*

$$I(p) \leq I(p^*) < +\infty \quad \textit{for all } p \in \mathcal{P}_U. \tag{11}$$

In what follows a stronger condition is needed. It was introduced some 40 years ago [16] — see Theorem 7.4.1 of [17] — and is in fact a stability criterion.

Definition 2 *A probability distribution p^* is said to satisfy the variational principle if there exist parameters $\theta_1, \theta_2, \cdots, \theta_n$ such that*

$$+\infty > I(p^*) - \sum_{j=1}^{n} \theta_j \langle p^*, H_j \rangle \geq I(p) - \sum_{j=1}^{n} \theta_j \langle p, H_j \rangle \quad \textit{for all } p \in \mathcal{M}_1^+. \tag{12}$$

In statistical physics, a probability distribution satisfying the variational principle is called an equilibrium state. The quantity in (12) is minus the free energy. The well-known interpretation of (12) is then that the free energy is minimal at thermodynamic equilibrium.

4. Lagrange multipliers

A popular way to solve the maximum entropy problem is by the introduction of Lagrange parameters. However, a difficulty arises, known as the cutoff problem. It is indeed possible that some of the probabilities p_a of the optimising probability distribution vanish. Let us see how this problem arises. The Lagrangean reads

$$\mathcal{L} = I(p) - \alpha \sum_{a \in A} p_a - \sum_{j=1}^{n} \theta_j \langle p, H_j \rangle. \tag{13}$$

Here, α is the parameter introduced to fix the normalisation condition $\sum_{a \in A} p_a = 1$, the θ_j are introduced to cope with the constraints (10). Variation of $\mathcal{L}$ w.r.t. the p_a yields

$$f(p_a) = -\alpha - \sum_{j=1}^{n} \theta_j H_j(a). \tag{14}$$

The existence of parameters α and θ_j, so that (14) holds, is known as one of the Karush-Kuhn-Tucker conditions — these are sufficient conditions for the existence of a global maximum. The problem that can arise is that it may well happen that the r.h.s. of this expression does not belong to the range of the function $f(u)$. This situation is particularly likely to occur when $f(u)$ does not tend to $-\infty$ when u tends to 0. If the r.h.s. is in the range of $f(u)$ then p_a is determined uniquely by (14) because of the assumption that $f(u)$ is a strictly increasing function.

The above problem is well known in optimisation theory. Because the constraints, defining $\mathcal{P}_U$, are affine, the set $\mathcal{P}_U$ forms a simplex. Its faces are obtained by putting some of the probabilities p_a equal to zero. Because the entropy function $I(p)$ is concave it attains its maximum within one of these faces. This observation leads to the ansatz that the probability distribution p, which maximises $I(p)$ with p in $\mathcal{P}_U$, if it exists, is determined by a subset $A_0 = \{a \in A : p_a = 0\}$, and by the values of the parameters α and θ_j, which determine the remaining probabilities via (14). Let us now try to prove this statement.

5. Characterisation

The Theorems 1 and 2 below give a characterisation of the probability distributions satisfying the variational principle. This is done separately for the cases that $f(0)$ is finite or infinite. Of course, both

theorems could have been taken together into one single theorem. But it is instructive to emphasise the complications which arise in the case of finite $f(0)$. For the same reason the proofs do not rely on the results of [13], but are worked out independently. In the present section it is assumed that $f(0) = -\infty$.

Lemma 1 *Assume $f(0) = -\infty$. Let $p^* \in \mathcal{M}_1^+$ satisfy the variational principle. Then $p_a^* > 0$ holds for all $a \in A$.*

Proof

The inverted statement is proved.

Because of the normalisation, there exists at least one $a \in A$ for which $p_a^* > 0$. Assume $b \in A$ such that $p_b^* = 0$. Let us show that this implies that p^* does not satisfy the variational principle.

Fix $0 < \epsilon << 1$. Introduce a new probability distribution p which coincides with p^* except that

$$p_a = (1-\epsilon)p_a^* \quad \text{and} \quad p_b = \epsilon p_a^*. \tag{15}$$

Let

$$M(\epsilon) = I(p) - \sum_{j=1}^{n} \theta_j \langle p, H_j \rangle. \tag{16}$$

Then one has

$$\frac{\mathrm{d}M}{\mathrm{d}\epsilon} = f((1-\epsilon)p_a^*) - f(\epsilon p_a^*) - \sum_{j=1}^{n} \theta_j p_a^* \left[H_j(a) - H_j(b) \right]. \tag{17}$$

From the assumption $f(0) = -\infty$ then follows that

$$\lim_{\epsilon \downarrow 0} \frac{\mathrm{d}M}{\mathrm{d}\epsilon} = +\infty. \tag{18}$$

This proves that p^* does not satisfy the variational principle because for ϵ sufficiently small $M(\epsilon)$ is strictly larger than $M(0)$.

□

Theorem 1 *Assume $f(0) = -\infty$. A probability distribution p^* satisfies the variational principle if and only if there exists α and $\theta_1, \theta_2, \cdots, \theta_n$ such that (14) holds for all $a \in A$.*

Proof

First assume that p^* satisfies (14). This implies that $p_a^* > 0$ for all $a \in A$ because $f(0)$ is not defined. Hence, the divergence $D(p||p^*)$ is well defined for all p. Next one calculates

$$\begin{aligned} D(p||p^*) &= I(p^*) - I(p) - \sum_{a \in A} (p_a - p_a^*) f(p_a^*) \\ &= I(p^*) - I(p) - \sum_{a \in A} (p_a - p_a^*) \left[-\alpha - \sum_{j=1}^{n} \theta_j H_j(a) \right] \\ &= I(p^*) - I(p) + \sum_{j=1}^{n} \theta_j \langle p - p^*, H_j \rangle. \end{aligned} \tag{19}$$

Because $D(p||p^*) \geq 0$ with equality if and only if $p = p^*$ there follows that p^* satisfies the variational principle.

Next assume that p^* satisfies the variational principle (12). From the lemma then follows that $p^*_a > 0$ for all $a \in A$. Hence, the divergence $D(p||p^*)$ is well-defined for all $p \in \mathcal{M}_1^+$. It follows from the variational principle that

$$\begin{aligned} D(p||p^*) &= I(p^*) - I(p) - \sum_{a\in A}(p_a - p^*_a)f(p^*_a) \\ &\geq \sum_{j=1}^{n}\theta_j\langle p^* - p, H_j\rangle - \sum_{a\in A}(p_a - p^*_a)f(p^*_a). \end{aligned} \tag{20}$$

Now, the function $p \to D(p||p^*)$ is convex with continuous derivatives. The r.h.s. of the above expression is affine. Both l.h.s. and r.h.s. vanish for $p = p^*$. One then concludes that the r.h.s. is tangent to the convex function and must be identically zero. One concludes that for all p

$$\sum_{a\in A}(p_a - p^*_a)f(p^*_a) = \sum_{j=1}^{n}\theta_j\langle p^* - p, H_j\rangle. \tag{21}$$

This implies that $f(p^*_a)$ is of the form (14) — take $p_a = \delta_{a,b}$ for some fixed b to see this.

□

6. The case with cutoff

Assume now that $f(0) = \lim_{u\downarrow 0} f(u)$ converges. Then the divergence $D(p||q)$ is well defined for any pair of probability distributions p, q.

Theorem 2 *Assume that $f(0) = \lim_{u\downarrow 0} f(u)$ converges. Are equivalent*

1. *p^* satisfies the variational principle;*
2. *there exist parameters α and θ_1, θ_2, $\cdots$, θ_n, and a subset A_0 of A such that*
   - *(14) is satisfied for all $a \in A \setminus A_0$;*
   - $p^*_a = 0$ *for all $a \in A_0$;*
   - $f(0) + \sum_{j=1}^{n}\theta_j H_j(a) \geq -\alpha$ *for all $a \in A_0$.*

Note that this last condition expresses that the r.h.s. of (14) is out of the range of $f(u)$ because it takes a value less than $f(0)$.

Proof

1) implies 2) As in the proof of the previous Theorem, one shows that (20) holds for all p. But now one cannot conclude (21) because some of the p^*_a may vanish so that p^* lies in one of the faces of the simplex $\mathcal{M}_1^+$. But one can still derive (14) for all a for which $p^*_a \neq 0$.

Assume now that $p_a^* = 0$ for some given $a \in A$. Let

$$p_b = (1-\epsilon)p_b^* + \epsilon\delta_{b,a}. \tag{22}$$

Then the l.h.s. of (20) becomes

$$\begin{aligned} D(p||p^*) &= \sum_{b\in A}^{\neq a} \int_{(1-\epsilon)p_b^*}^{p_b^*} \mathrm{d}u\ [f(p_b^*) - f(u)] + \int_0^\epsilon \mathrm{d}u\ [f(u) - f(0)] \\ &\leq \epsilon \sum_{b\in A} p_b^* \left[f(p_b^*) - f((1-\epsilon)p_b^*)\right] + \int_0^\epsilon \mathrm{d}u\, f(u) - \epsilon f(0) \\ &= \mathrm{O}(\epsilon^2). \end{aligned} \tag{23}$$

On the other hand, the r.h.s. of (20) becomes

$$\text{r.h.s.} = \epsilon \sum_{j=1}^n \theta_j \sum_{b\in A} p^*(b) H_j(b) - \epsilon \sum_{j=1}^n \theta_j H_j(a) + \epsilon \sum_{b\in A} p_b^* f(p_b^*) - \epsilon f(0). \tag{24}$$

From the inequality (20) then follows

$$0 \geq \sum_{j=1}^n \theta_j \langle p^*, H_j\rangle - \sum_{j=1}^n \theta_j H_j(a) + \sum_{b\in A} p_b^* f(p_b^*) - f(0). \tag{25}$$

This implies the desired inequality because

$$-\alpha = \sum_{b\in A} p_b^* f(p_b^*) + \sum_{j=1}^n \theta_j \langle p^*, H_j\rangle. \tag{26}$$

2) implies 1) One calculates

$$\begin{aligned} I(p) - \sum_{j=1}^n \theta_j \langle p, H_j\rangle &= -D(p||p^*) + I(p^*) - \sum_{a\in A}(p_a - p_a^*) f(p_a^*) - \sum_{j=1}^n \theta_j \langle p, H_j\rangle \\ &\leq I(p^*) - f(0) \sum_{a\in A_0} p_a + \sum_{a\in A\setminus A_0} (p_a - p_a^*) \left[\alpha + \sum_{j=1}^n \theta_j H_j(a)\right] - \sum_{j=1}^n \theta_j \langle p, H_j\rangle \\ &= I(p^*) - \sum_{j=1}^n \theta_j \langle p^*, H_j\rangle - \sum_{a\in A_0} p_a \left[f(0) + \alpha + \sum_{j=1}^n \theta_j H_j(a)\right]. \end{aligned} \tag{27}$$

The variational principle now follows using the third assumption of the Theorem.

□

7. Statistical models

In the definition of the variational principle there is given a set of Hamiltonians $H_1(a)$, $H_2(a)$, $\cdots$, $H_n(a)$, this means, real functions over the alphabet A, bounded from below. The equilibrium distribution p^* is then characterised by a normalisation constant α, by parameters θ_1, θ_2, $\cdots$, θ_n, and by a subset A_0 of the alphabet A — see (14). The emphasis now shifts towards these parameters.

Theorem 3 *Let be given Hamiltonians $H_1(a)$, $H_2(a)$, $\cdots$, $H_n(a)$. For each θ in $\mathbb{R}^n$ there exists at most one probability distribution p^* satisfying the variational principle (12) with these parameters θ.*

Proof

If p^* and q^* both satisfy the variational principle (12) with the same parameters θ then also the convex combination $r^* = \frac{1}{2}p^* + \frac{1}{2}q^*$ has the same property because the entropy function is concave. But then one can conclude from the inequalities (12) that $I(r^*) = \frac{1}{2}I(p^*) + \frac{1}{2}I(q^*)$. Because the entropy function is strictly concave there follows $p^* = q^*$.

□

The set of θ for which a p^* exists, satisfying the variational principle (12), is denoted $\mathcal{D}$. The probability distribution is denoted p_θ instead of p^*. The constant α appearing in (14) is replaced by $\alpha(\theta)$.

A statistical model is a parametrised set of probability distributions. The above Theorem implies that the set $(p_\theta)_{\theta\in\mathcal{D}}$, of probability distributions satisfying the variational principle, is a statistical model. One can say that such a model belongs to the generalised exponential family.

Definition 3 *Let be given a generalised entropy function $I(p)$ of the form (4). A statistical model $(p_\theta)_{\theta\in\mathcal{D}}$ belongs to the generalised exponential family if there exist real functions $H_1(a)$, $H_2(a)$, $\cdots$, $H_n(a)$, bounded from below, such that each member p_θ of the model satisfies the variational principle (12) with these Hamiltonians and with this set of parameters.*

This definition corresponds with the notion of *natural generalised exponential family* as introduced by Grünwald and Dawid [13]. It extends slightly the notion of phi-exponential family found in [11].

Clearly, entropy functions which differ only by a scalar factor determine the same generalised exponential family.

8. Uniqueness theorem

Let us now turn to the question whether a given model $(p_\theta)_{\theta\in\mathcal{D}}$ can belong to two different generalised exponential families.

Theorem 4 *Let be given a model $(p_\theta)_{\theta\in\mathcal{D}}$. Assume that there exists an open subset $\mathcal{D}_0$ of $\mathcal{D}$ with the property that the set of values of $p_{\theta,a}$ covers the open interval $(0,1)$*

$$(0,1) \subset \{p_{\theta,a} : \theta \in \mathcal{D}_0, a \in A\}. \tag{28}$$

If the model belongs to two different generalised exponential families, one with entropy function $I_1(p)$, the other with entropy function $I_2(p)$, then there exists a constant λ such that $I_2(p) = \lambda I_1(p)$ for all p.

Proof

Take any point u in $(0,1)$ and a corresponding $\theta \in \mathcal{D}_0$ and a such that $p_{\theta,a} = u$. From the previous theorems follows that there exist functions $\alpha_i(\theta)$ and Hamiltonians $H_{i1}(a)$, $H_{i2}(a)$, $\cdots$, $H_{in}(a)$, with $i = 1, 2$, such that

$$p_{\theta,a} = f_{i,a}^{-1}\left(-\alpha_i(\theta) - \sum_{j=1}^{n} \theta_j H_{i,j}(a)\right). \tag{29}$$

Let $F_a = f_{2,a} \circ f_{1,a}^{-1}$. Note that this is a strictly increasing continuous function. Then one has

$$F_a\left(-\alpha_1(\theta) - \sum_{j=1}^{n} \theta_j H_{1,j}(a)\right) = -\alpha_2(\theta) - \sum_{j=1}^{n} \theta_j H_{2,j}(a). \tag{30}$$

This relation holds also on a vicinity of $\theta \in \mathcal{D}_0$. It therefore implies the existence of λ_a and $K_{i,j}$ such that

$$H_{2,j}(a) - K_{2,j} = \lambda_a(H_{1,j}(a) - K_{1,j}), \quad j = 1, 2, \cdots, n. \tag{31}$$

Then one can rewrite (30) as

$$F_a(v) = \gamma_a(\theta) + \lambda_a v, \tag{32}$$

with

$$\gamma_a(\theta) = -\alpha_2(\theta) - \sum_{j=1}^{n} \theta_j K_{2,j} + \lambda_a \left[\alpha_1(\theta) + \sum_{j=1}^{n} \theta_j K_{1,j}\right], \tag{33}$$

valid for some neighbourhood of the given θ. Using the definition of $F_a(v)$ one obtains

$$f_{2,a}(u) \quad = \quad \gamma_a(\theta) + \lambda_a f_{1,a}(u), \tag{34}$$

valid on some neighbourhood of the given $u \in (0, 1)$. Because u is arbitrary and the functions f_{ia} are continuous, the same expression must hold on all of $(0, 1]$. From $0 = h_{i,a}(0) = \int_0^1 \mathrm{d}u\, f_{i,a}(u)$ now follows that $\gamma_a(\theta) = 0$. Therefore (33) becomes

$$\lambda_a = \frac{\alpha_2(\theta) + \sum_{j=1}^{n} \theta_j K_{2,j}}{\alpha_1(\theta) + \sum_{j=1}^{n} \theta_j K_{1,j}}. \tag{35}$$

In particular, λ_a does not depend on $a \in A$. One concludes therefore that there exists λ so that $f_{2,a}(u) = \lambda f_{1,a}(u)$. This implies $I_2(p) = \lambda I_1(p)$.

□

9. Thermodynamics

Throughout this Section, let be given a statistical model $(p_\theta)_{\theta \in \mathcal{D}}$ belonging to the generalised exponential family.

Note that if p_θ and p_η both belong to the same set $\mathcal{P}_U$ then they satisfy $I(p_\theta) = I(p_\eta)$. Hence, a function $S(U)$ can be defined by

$$S(U) = I(p_\theta) \quad \text{whenever } \langle p_\theta, H_j \rangle = U_j \text{ for } j = 1, 2, \cdots, n. \tag{36}$$

In the physics literature, this function is called the thermodynamic entropy (it was called specific entropy in [13]; but note that specific entropy has a different meaning in thermodynamics). The concept of

thermodynamic entropy was first introduced by Clausius around 1850. The Legendre transform of the thermodynamic entropy is given by

$$\Phi(\theta) = \sup\{S(U) - \sum_{j=1}^{n} \theta_j U_j\}. \tag{37}$$

This function was introduced by Massieu in 1869. The suprememum is taken over all U for which $S(U)$ is defined by (36). The function is convex — this is a well-known property of Legendre transforms.

Proposition 1 *One has*

$$\Phi(\theta) = I(p_\theta) - \sum_{j=1}^{n} \theta_j \langle p_\theta, H_j \rangle, \qquad \theta \in \mathcal{D}. \tag{38}$$

Proof

Given $\theta \in \mathcal{D}$ there exists p_θ for which the variational principle holds. Then one has, with $U_j = \langle p_\theta, H_j \rangle$,

$$I(p_\theta) - \sum_{j=1}^{n} \theta_j \langle p_\theta, H_j \rangle = S(U) - \sum_{j=1}^{n} \theta_j U_j \leq \Phi(\theta). \tag{39}$$

This proves the inequality in one direction. Next, fix $\epsilon > 0$ and let U be such that

$$\Phi(\theta) \leq S(U) - \sum_{j=1}^{n} \theta_j U_j + \epsilon, \tag{40}$$

with U such that $S(U)$ is defined by (36). Then, there follows from the definition of $S(U)$ that $\eta \in \mathcal{D}$ exists such that $S(U) = I(p_\eta)$ with $\langle p_\eta, H_j \rangle = U_j$, $j = 1, 2, \cdots, n$. The variational principle now implies that

$$\begin{aligned} I(p_\theta) - \sum_{j=1}^{n} \theta_j \langle p_\theta, H_j \rangle &\geq I(p_\eta) - \sum_{j=1}^{n} \theta_j \langle p_\eta, H_j \rangle \\ &= S(U) - \sum_{j=1}^{n} \theta_j U_j \\ &\geq \Phi(\theta) - \epsilon. \end{aligned} \tag{41}$$

Because $\epsilon > 0$ is arbitrary, the inequality in the other direction follows now.

□

The inverse Legendre transformation reads

$$\overline{S}(U) = \inf_{\theta}\{\Phi(\theta) + \sum_{j=1}^{n} \theta_j U_j\}. \tag{42}$$

It is a concave function.

Proposition 2 *One has $S(U) = \overline{S}(U)$ for all U for which $S(U)$ is defined by (36).*

Proof

From the definition of the Massieu function $\Phi(\theta)$ there follows that

$$\Phi(\theta) \geq S(U) - \sum_{j=1}^{n} \theta_j U_j \quad \text{for all } \theta \in \mathbb{R}^n. \tag{43}$$

This implies that $S(U) \leq \overline{S}(U)$. On the other hand, from the definition (36) of $S(U)$ follows that

$$S(U) = \Phi(\theta) + \sum_{j=1}^{n} \theta_j U_j, \tag{44}$$

where θ is such that $p_\theta \in \mathcal{P}_U$. This implies $S(U) \geq \overline{S}(U)$. The two inequalities together establish the desired equality.

□

10. Thermodynamic relations

Like in the previous Section, there is given a statistical model $(p_\theta)_{\theta\in\mathcal{D}}$ belonging to the generalised exponential family. In addition, let $\mathcal{D}_0$ be an open subset of $\mathcal{D}$ on which the map $\theta \to \langle p_\theta, H_j\rangle$ is continuous.

The following results are typical properties of Legendre transforms. For completeness, proofs are given.

Proposition 3 *The first derivative of the Massieu function $\Phi(\theta)$ exists for θ in $\mathcal{D}_0$. It satisfies*

$$\frac{\partial \Phi}{\partial \theta_j} = -\langle p_\theta, H_j\rangle, \quad \theta \in \mathcal{D}_0. \tag{45}$$

Proof

From the definitions one has for θ and $\theta + \eta$ in $\mathcal{D}_0$

$$\begin{aligned} \Phi(\theta+\eta) &= I(p_{\theta+\eta}) - \sum_{j=1}^{n} (\theta_j + \eta_j)\langle p_{\theta+\eta}, H_j\rangle \\ &\geq I(p_\theta) - \sum_{j=1}^{n} (\theta_j + \eta_j)\langle p_\theta, H_j\rangle \\ &= \Phi(\theta) - \sum_{j=1}^{n} \eta_j \langle p_\theta, H_j\rangle, \end{aligned} \tag{46}$$

and

$$\begin{aligned} \Phi(\theta) &= I(p_\theta) - \sum_{j=1}^{n} \theta_j \langle p_\theta, H_j\rangle \\ &\geq I(p_{\theta+\eta}) - \sum_{j=1}^{n} \theta_j \langle p_{\theta+\eta}, H_j\rangle \end{aligned}$$

$$= \quad \Phi(\theta+\eta) + \sum_{j=1}^{n} \eta_j \langle p_{\theta+\eta}, H_j \rangle. \tag{47}$$

Expression (45) now follows using the continuity of the map $\theta \to \langle p_\theta, H_j \rangle$.

□

Introduce the metric tensor

$$g_{i,j}(\theta) = \frac{\partial^2 \Phi}{\partial \theta_i \partial \theta_j}. \tag{48}$$

Because the Massieu function $\Phi(\theta)$ is convex the matrix $g(\theta)$ is positive definite, whenever it exists. By the previous Proposition one has

$$g_{i,j}(\theta) = -\frac{\partial}{\partial \theta_i} \langle p_\theta, H_j \rangle \tag{49}$$

for those θ in $\mathcal{D}_0$ for which the derivative exists.

In thermodynamics, the derivative of $S(U)$ equals the inverse of the absolute temperature T. Here, the analogous property becomes

Proposition 4 *Let $\theta \in \mathcal{D}_0$ and define U by $U_j = \langle p_\theta, H_j \rangle$. Then one has*

$$\frac{\partial S}{\partial U_j} = \theta_j, \quad j = 1, 2, \cdots, n. \tag{50}$$

Proof

On a vicinity of θ is $S(U) = \Phi(\theta) + \sum_{j=1}^{n} \theta_j U_j$. Hence, one can write

$$\frac{\partial S}{\partial \theta_j} \quad = \quad \sum_{k=1}^{n} \left(\frac{\partial \Phi}{\partial \theta_k} + U_k \right) \frac{\partial \theta_k}{\partial U_j} + \theta_j. \tag{51}$$

But the first term in the r.h.s. vanishes because the previous Proposition holds. Hence, the desired result follows.

□

The two relations (45) and (50) are dual in the sense of Amari [18]. In thermodynamics, the entropy $S(U)$ and Massieu's function $\Phi(\theta)$ are state functions, the energies U_j are extensive thermodynamic variables, the parameters θ_j are the intensive thermodynamic variables.

11. Escort probabilities

Let us now make the additional assumption that the function $f(u)$, which enters the definition (6) of the generalised entropy, has a derivative $f'(u)$. Because $f(u)$ was supposed to be strictly increasing, one can write

$$f(u) = f(1) - \int_u^1 \mathrm{d}v \, \frac{1}{\phi(v)}, \quad u \in (0, 1], \tag{52}$$

where $\phi(v) = 1/(\mathrm{d}f/\mathrm{d}v)$ is a strictly positive function.

As before, there is given a statistical model $(p_\theta)_{\theta\in\mathcal{D}}$ belonging to the generalised exponential family, and $\mathcal{D}_0$ is an open subset of $\mathcal{D}$ on which the map $\theta \to \langle p_\theta, H_j\rangle$ is continuous. The set $A_0(\theta)$ is the set of $a \in A$ for which $p_\theta(a) = 0$. From theorems 1 and 2 now follows

$$\frac{\partial}{\partial\theta_j} p_{\theta,a} = \phi(p_{\theta,a})\left(-\frac{\partial\alpha}{\partial\theta_j} - H_j(a)\right), \quad \theta \in \mathcal{D}_0, a \in A \setminus A_0(\theta). \tag{53}$$

This expression was used in [11] as a condition under which a generalisation of the well-known bound of Cramér and Rao is optimal. An immediate consequence of (53) is

Proposition 5 *Assume the regularity condition*

$$0 = \sum_a \frac{\partial}{\partial\theta_j} p_\theta(a). \tag{54}$$

Assume in addition that

$$z(\theta) = \sum{}' \phi(p_{\theta,a}) < +\infty, \tag{55}$$

where $\sum'$ *denotes the sum over all* $a \in A \setminus A_0(\theta)$. *Then one has*

$$\frac{\partial\alpha}{\partial\theta_j} = -\frac{1}{z(\theta)} \sum{}' \phi(p_{\theta,a}) H_j(a). \tag{56}$$

Proof

On a vicinity of the given θ one has (53). Hence, by summing (53) over $a \in A \setminus A_0(\theta)$ one obtains using (54)

$$0 \;=\; \sum{}' \phi(p_{\theta,a})\left(-\frac{\partial\alpha}{\partial\theta_j} - H_j(a)\right), \quad \theta \in \mathcal{D}_0, a \in A \setminus A_0(\theta). \tag{57}$$

□

The probability distribution

$$\begin{aligned} P_{\theta,a} &= \frac{1}{z(\theta)}\phi(p_{\theta,a}), \quad p_{\theta,a} \neq 0, \\ &= 0, \qquad \text{otherwise}, \end{aligned} \tag{58}$$

when it exists, is called the escort of the model $(p_\theta)_{\theta\in\mathcal{D}}$. With this notation, one can write the result of the Proposition as

$$\frac{\partial\alpha}{\partial\theta_j} = -\langle P_\theta, H_j\rangle. \tag{59}$$

12. Generalised Fisher information

Let be given a model $(p_\theta)_{\theta\in\mathcal{D}}$ for which $z(\theta)$, as given by (55), converges. The escort probabilities $P_{\theta,a}$ are defined by (58). Then one can define a generalised Fisher information matrix by

$$I_{i,j}(\theta) \quad = \quad \langle P_\theta, X_i(\theta)X_j(\theta)\rangle, \tag{60}$$

where the score variables are defined by

$$X_{i,a}(\theta) \equiv \frac{1}{P_{\theta,a}}\frac{\partial}{\partial\theta_i}p_{\theta,a}. \tag{61}$$

Note that in the standard case of $h(u) = -u\ln u$ one has $\phi(u) = u$ so that the escort probabilities P_θ coincide with the p_θ. Then (60) reduces to the conventional definition.

Fix now a set of Hamiltonians $H_1(a)$, $H_2(a)$, $\cdots$, $H_n(a)$. Then one can define a covariance matrix $\sigma(\theta)$ by

$$\sigma_{i,j}(\theta) = \langle P_\theta, H_iH_j\rangle - \langle P_\theta, H_i\rangle\,\langle P_\theta, H_j\rangle. \tag{62}$$

Proposition 6 *Assume a finite alphabet A. Then one has*

$$I_{i,j}(\theta) = z(\theta)g_{i,j} = z^2(\theta)\sigma_{i,j}. \tag{63}$$

Proof

From (53) follows

$$X_{j,a}(\theta) \quad = \quad z(\theta)\left(-\frac{\partial\alpha}{\partial\theta_j} - H_j(a)\right) \tag{64}$$

for all $\theta\in\mathcal{D}_0$ and $a\in A\setminus A_0(\theta)$. Hence, the Fisher information matrix becomes

$$I_{i,j}(\theta) \quad = \quad z^2(\theta)\sum_{a\in A}P_{\theta,a}\left(-\frac{\partial\alpha}{\partial\theta_i} - H_i(a)\right)\left(-\frac{\partial\alpha}{\partial\theta_j} - H_j(a)\right). \tag{65}$$

Using (59) there follows $I_{i,j}(\theta) = z^2(\theta)\sigma_{i,j}$.

On the other hand, from (49) and (53) there follows

$$\begin{aligned} g_{i,j}(\theta) \quad &= \quad -\frac{\partial}{\partial\theta_i}\sum_{a\in A}p_{\theta,a}H_j(a) \\ &= \quad -\sum_{a\in A}P_{\theta,a}\left(-\frac{\partial\alpha}{\partial\theta_i} - H_i(a)\right)H_j(a). \end{aligned} \tag{66}$$

Using (56) there follows $g_{i,j}(\theta) = z(\theta)\sigma_{i,j}$.

□

The assumption of a finite alphabet is made to ensure that the conditions of Proposition 5 are fulfilled and that the sum and derivative may be interchanged in (66).

The generalised inequality of Cramér and Rao, in the present notations, reads [11]

$$\left(\sum_{kl} \sigma_{kl} u_k u_l\right)\left(\sum_{kl} I_{kl} v_l v_k\right) \geq \left(\sum_{kl} g_{kl} u_k v_l\right)^2, \tag{67}$$

with u and v arbitrary real vectors. The previous Proposition then implies that the inequality becomes an equality when $u = v$, when P is related to p via (58), and when p_θ belongs to a generalised exponential family.

13. Non-extensive thermostatistics

Define the q-deformed logarithm by [7, 19]

$$\ln_q(u) = \frac{1}{1-q}\left(u^{1-q} - 1\right). \tag{68}$$

It is a strictly increasing function, defined for $u > 0$. Indeed, its derivative equals

$$\frac{\mathrm{d}}{\mathrm{d}u} \ln_q(u) = \frac{1}{u^q} > 0. \tag{69}$$

In the limit $q = 1$ the q-deformed logarithm converges to the nature logarithm $\ln u$.

The deformed logarithm can be used in more than one way to define an entropy function. The q-entropy (3) can be written as

$$I_q(p) = \sum_{a \in A} p_a \ln_q\left(\frac{1}{p_a}\right). \tag{70}$$

Comparison with (4) gives

$$h(u) = \frac{u}{1-q}\left(u^{q-1} - 1\right) = u \ln_q\left(\frac{1}{u}\right). \tag{71}$$

One has $h(0) = h(1) = 0$. Taking the derivative gives

$$f(u) = -\frac{\mathrm{d}h}{\mathrm{d}u} = \frac{1}{q-1}\left(qu^{q-1} - 1\right). \tag{72}$$

It is a strictly increasing function on $(0, 1]$ when $q > 0$. The function $\phi(u)$ is given by

$$\phi(u) = \frac{1}{q} u^{2-q}. \tag{73}$$

The probability distributions belonging to the generalised exponential family, corresponding with (70), are

$$p_a = q^{1/(1-q)} \left[1 - (q-1)\alpha - (q-1)\sum_j \theta_j H_j(a)\right]_+^{1/(q-1)}, \tag{74}$$

with $[u]_+ = \max\{0, u\}$. This is indeed the kind of probability distribution discussed in the original paper of Tsallis [4]. However, more often used is the alternative of [9]. In the latter paper the concept of escort probability distributions was introduced into the literature. They were defined by

$$P_a = \frac{1}{Z} p_a^q, \tag{75}$$

which in the present notations corresponds with $\phi(u)$ proportional to u^q. This can be obtained by replacing the constant q by $2 - q$ in (70). The entropy function then reads

$$I(p) = -\sum_a p_a \ln_q(p_a), \tag{76}$$

which is *not* the expression that one would write down based on the information theoretical argument that $\ln(1/p_a)$ is the amount of information (counted in units of $\ln 2$), gained from an event occurring with probability p_a. Note that with this definition of entropy function the condition $q < 2$ is needed in order to satisfy the requirements that the function $f(u) = \frac{\mathrm{d}}{\mathrm{d}u}(u \ln_q(u)$ is an increasing function.

14. The percolation problem

This example has been treated in [20]. It is a genuine example of an important model of statistical physics which does not belong to the exponential family. In addition, it is an example which fits into the present generalised context provided that one allows that the function $h(u)$ appearing in the definition (4) of the generalised entropy function is stochastic.

In the site percolation problem [21], the points of a lattice are occupied with probability q, independent of each other. The point at the origin is either unoccupied, with probability $p_\emptyset$, or it belongs to a cluster of shape i, with probability p_i. This cluster is finite with probability 1, provided that $0 \leq q \leq q_c$, where q_c is the percolation threshold. The probability p_∞ that the origin belongs to an infinite cluster is strictly positive for $q > q_c$. However, for the sake of simplicity of the presentation, $0 < q < q_c$ will be assumed — see [20] for the general case.

These probabilities are given by

$$p_i = c_i q^{s(i)} (1-q)^{t(i)}, \tag{77}$$

where c_i is the number of different clusters of shape i, $s(i)$ is the number of occupied sites in the cluster, and $t(i)$ is the number of perimeter sites, this is, of unoccupied neighbouring sites. Note that (77) also holds when the origin is not occupied, provided that one convenes that $c(\emptyset) = 1$, $s(\emptyset) = 0$ and $t(\emptyset) = 1$.

Choose the Hamiltonian

$$H(i) = \frac{t(i)}{t(i) + s(i)}. \tag{78}$$

and introduce the parameter θ by

$$\theta = \ln \frac{q}{1-q}, \qquad q = \frac{1}{1 + e^{-\theta}}. \tag{79}$$

Then one can write

$$\ln \frac{p_i}{c_i} = [-\alpha(\theta) - \theta H(i)]\ [s(i) + t(i)]\,, \tag{80}$$

with

$$\alpha(\theta) = \ln(1 + e^{-\theta}) \tag{81}$$

This looks like an exponential family, except for the extra factor $[s(i) + t(i)]$ in the r.h.s.. Introduce the stochastic function

$$f_i(u) \quad = \quad \frac{\ln u}{s(i) + t(i)}. \tag{82}$$

Then the above expression is of the form (14). By integrating $f_i(u)$ one obtains

$$h_i(u) \quad = \quad -\frac{u \ln u}{s(i) + t(i)}. \tag{83}$$

It is now straightforward to verify that the percolation problem belongs to a generalised exponential family. The relevant entropy function for the percolation model in the non-percolating region $0 < q < q_c$ is therefore

$$I(p) = -\sum_i \frac{p_i \ln p_i}{s(i) + t(i)}. \tag{84}$$

15. Discussion

Sections 3 to 6 of the present paper discuss the variational principle, which is stronger than the maximum entropy principle. It is shown that the method of Lagrange multipliers leads to the correct result, even in the context of generalised entropy functions. The difficulty that arises is known as the cut-off problem: the optimising probability distribution may assign vanishing probabilities to some of the events. To cope with this situation the two cases have been considered separately. Theorem 1 treats the standard case, Theorem 2 copes with the vanishing probabilities.

In Section 7, a generalised definition of an exponential family is given. It identifies the members of the generalised exponential family with the solutions of the variational principle, given a generalised entropy function of the usual form (4). The definition of the standard exponential family corresponds of course with the Boltzmann-Gibbs-Shannon entropy. Entropy functions $I(p)$ and $\lambda I(p)$, with $\lambda > 0$, determine the same exponential family. Assuming some technical condition, the intersection of different generalised exponential families is empty — see Theorem 4. As a consequence, a one-to one relation has been established between generalised exponential families and classes of equivalent entropy functions.

In [11], the notion of phi-exponential family was introduced. The 'phi' in this name refers to the function $\phi(v)$, introduced in (52). It is one divided by the derivative of the function $f(v)$ appearing in the expression (6) for the entropy function $I(p)$. The assumption that the derivative of $f(v)$ exists for all $v > 0$ has been eliminated in the present paper. More important is that the definition of a generalised exponential family is now given directly in terms of the entropy function $I(p)$, via the variational principle, without relying on the notion of deformed exponential functions.

Sections 9 to 12 discuss the geometric properties of a generalised exponential family, using a terminology coming from 150 year old thermodynamics. The main result is (63), proving the equality of the three quantities generalised Fisher information, metric tensor times partition sum $z(\theta)$, and covariance matrix multiplied with $z^2(\theta)$. The covariance matrix is calculated using the escort family of probability distributions.

Many applications of generalised exponential families are found in the literature, in the context of nonextensive thermostatistics. The latter has been discussed in Section 13. A completely different kind of example is found in percolation theory — see Section 14. It illustrates the possibility that the function $f(u)$, which determines the entropy function $I(p)$ via (6), is of a stochastic nature. One can expect that many other applications will be found in the near future.

Finally note that the present work has a quantum analogue. Let be given a strictly increasing function $f(u)$, continuous on $(0, 1]$. The expression (6) can be generalised to

$$I(\rho) = -\int_0^1 \mathrm{d}v \ \mathrm{Tr}\, \rho f(v\rho), \tag{85}$$

where ρ is any density operator in a Hilbert space. The Bregman divergence (8) generalises to

$$D(\rho||\rho') = I(\rho') - I(\rho) - \ \mathrm{Tr}\,(\rho - \rho')f(\rho). \tag{86}$$

The basic inequality $D(\rho||\rho') \geq 0$ is proved using Klein's inequality — see 2.5.2. of [17].

Acknowledgements

This work has benefitted from a series of discussions with Flemming Topsøe. I am grateful to the anonymous referees for their constructive remarks and for pointing out the references [2, 13, 14].

References

1. Petz, D., Quasi-entropies for finite quantum systems *Rep. Math. Phys.* **1986**, *23*, 57–65.
2. Csiszár, I., Information type measure of difference of probability distributions and indirect observations *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
3. Csiszár, I., A class of measures of informativity of observation channels *Per. Math. Hung.* **1972**, *2*, 191–213.
4. Tsallis, C., Possible Generalization of Boltzmann-Gibbs Statistics *J. Stat. Phys.* **1988**, *52*, 479–487.
5. Havrda, J.; Charvat, F., Quantification method of classification processes, the concept of structural a-entropy *Kybernetica* **1967**, *3*, 30–35.
6. Daróczy, Z., Generalized Information Functions *Information and Control* **1970**, *16*, 36–51.
7. Tsallis, C., What are the numbers that experiments provide? *Quimica Nova* **1994**, *17*, 468.
8. Naudts, J., Deformed exponentials and logarithms in generalized thermostatistics *Physica A* **2002**, *316*, 323–334.
9. Tsallis, C.; Mendes, R.; Plastino, A., The role of constraints within generalized nonextensive statistics *Physica A* **1998**, *261*, 543–554.

10. Naudts, J., Continuity of a class of entropies and relative entropies *Rev. Math. Phys.* **2004**, *16*, 809–822.
11. Naudts, J., Estimators, escort probabilities, and phi-exponential families in statistical physics *J. Ineq. Pure Appl. Math.* **2004**, *5*, 102.
12. Topsøe, F., Information-theoretical optimization techniques *Kybernetika* **1979**, *15*, 8–27.
13. Grünwald, P.; Dawid, A., Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory *Ann. Statist.* **2004**, *32*, 1367–1433.
14. Lafferty, J. Additive Models, Boosting, and Inference for Generalized Divergences. *Additive models, boosting, and inference for generalized divergences*, 1999; pp 125–133.
15. Bregman, L., The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming *USSR Comp. Math. Math. Phys.* **1967**, *7*, 200–217.
16. Ruelle, D., A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule *Commun. Math. Phys.* **1967**, *5*, 324–329.
17. Ruelle, D. *Statistical mechanics*; W.A. Benjamin: New York, 1969.
18. Amari, S. *Differential-geometrical methods in statistics*; Springer: New York, Berlin, 1985; Vol. 28.
19. Tsallis, C. Nonextensive statistical mechanics: construction and physical interpretation. *Nonextensive Entropy*, Oxford, 2004; pp 1–53.
20. Naudts, J., Parameter estimation in nonextensive thermostatistics *Physica A* **2006**, *365*, 42–49.
21. Stauffer, D. *Introduction to percolation theory*; Plenum Press: New York, 1985.