Towards operational interpretations of generalized entropies

Flemming Topsøe

University of Copenhagen, Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen, Denmark

E-mail: topsoe@math.ku.dk

Abstract. The driving force behind our study has been to overcome the difficulties you encounter when you try to extend the clear and convincing operational interpretations of classical Boltzmann-Gibbs-Shannon entropy to other notions, especially to generalized entropies as proposed by Tsallis. Our approach is philosophical, based on speculations regarding the interplay between *truth, belief* and *knowledge*. The main result demonstrates that, accepting philosophically motivated assumptions, the only possible measures of entropy are those suggested by Tsallis – which, as we know, include classical entropy. This result constitutes, so it seems, a more transparent interpretation of entropy than previously available. However, further research to clarify the assumptions is still needed. Our study points to the thesis that one should never consider the notion of entropy in isolation – in order to enable a rich and technically smooth study, further concepts, such as *divergence*, *score functions* and *descriptors* or *controls* should be included in the discussion. This will clarify the distinction between *Nature* and *Observer* and facilitate a game theoretical discussion. The usefulness of this distinction and the subsequent exploitation of game theoretical results – such as those connected with the notion of *Nash equilibrium* – is demonstrated by a discussion of the *Maximum Entropy Principle.*

1. Introduction

Our aim is to provide transparent operational interpretations of generalized notions of entropy, especially of *Tsallis entropy*, cf. [1] and [2]. Our approach depends on abstract, philosophical considerations centred around notions of *truth, belief* and *knowledge* and their possible interplay. In Section 2 we introduce the main abstract notions. They are not necessarily tied to probabilistic considerations. However, for the present exposition, we tone down a bit the very abstract discussion.

In Section 3, we introduce the probabilistic models we shall work with. We only consider discrete models. For these models, we identify the most natural or acceptable forms of interplay between truth, belief and knowledge. Key notions are related to *interaction* and *description.*

Quantitative reasoning is enabled by the introduction of *score functions.* This type of function is known from statistical decision theory. Here, it determines the *effort* needed by an observer in order to gain knowledge. A variational principle is introduced which is related to the *fundamental inequality of information theory.* The notions introduced are needed for the formulation of two main results, Theorems 1 and 2. In particular, Theorem 2 is presented as an operational interpretation of Tsallis entropy. It singles out Tsallis entropy among other possibilities and may thus be taken to support the view that these notions occupy a unique position in statistical physics.

In Section 4 we introduce concepts closely related to *experiments* and *observation*. The further study depends on game theoretical considerations and this is taken up in Section 5. Two notions of equilibrium are introduced and their relation is established in Theorem 3. This result also contains the *pythagorean inequalities*, well known and celebrated results from information theory.

Section 6 introduces *exponential families* and Theorem 4 establishes their relevance for the easy determination of equilibrium and optimal strategies, in particular *maximum entropy distributions* in cases where the models studied are given by what we call *feasible preparations*.

Section 7, the final section, contains a discussion of various points related to the text with a view also to desirable further research.

Throughout the study, we have emphasized the underlying ideas and though a fair number of proofs are given, we have introduced certain simplifications and tried to avoid mathematical subtleties.

2. Abstract philosophical considerations

The whole is the *world*. We shall consider *situations* from the world which involve *Nature*, without a mind but holder of *truth*, and *Observer*, seeking the truth but restricted to *belief*. In contrast to Nature, Observer has a conscious and creative mind which can be exploited to obtain *knowledge* as effortlessly as possible.

By x we denote a generic *truth instance* associated with a specific situation. We imagine that Nature has "chosen" this instance from a certain set $\mathbb{M}$. To simplify notationally, we shall not express any dependency of $\mathbb{M}$ on the situation. This is justified by the fact that for the present study only one situation will be considered at a time. The set $\mathbb{M}$ is the *set of possible truth instances.*

In any situation, Observer speculates over what Nature is up to, and Observer expresses his belief in the form of an assignment of a *belief instance*, typically denoted by y, to the situation. To simplify, we assume that the belief instance is also chosen from $\mathbb{M}$.

Observers choice of belief instance in any specific situation is considered a subjective choice taking available information into account such as general insight and any specific *prior knowledge*. These thoughts agree with Bayesian thinking, and as such are subject to standard criticism which applies to this line of thought, cf. [3].

Observer acts by designing *experiments* and by making subsequent *observations*. We shall later return in more detail to this aspect. For now we note that the result of observations can be more or less informative, ranging from initial and very limited experience to a final more conclusive stage. It is the final stage we have in mind. We refer to it as *knowledge* and think of it as the *synthesis of extensive experience.*

The end result of Observers endeavours in any particular situation is a *knowledge instance*, typically denoted by z. We reserve the letter $\mathbb{F}$ for the *set of knowledge instances*. We assume that $\mathbb{M} \subseteq \mathbb{F}$. Often, $\mathbb{F} = \mathbb{M}$ will hold.

The connection to extensive experience is just one side of "knowledge". We may also view knowledge as *the way the World presents truth to Observer in any given situation* and, therefore, as the way Observer *perceives* the situation. We assume that knowledge is a function which depends on the situation through the already introduced notions of truth- and belief instances. Formally, we shall operate with a function $\Pi : \mathbb{M} \times \mathbb{M} \to \mathbb{F}$, and interpret $z = \Pi(x, y)$ as the knowledge instance in a situation with x as truth instance and y as belief instance. We call Π the *interactor*, sometimes the *global interactor* in order to distinguish it from a related concept to be introduced later. Two worlds with the same interactor are identified, thus conceived as the *same* world. We use $\mathcal{W}_\Pi$ to denote the world with interactor Π.

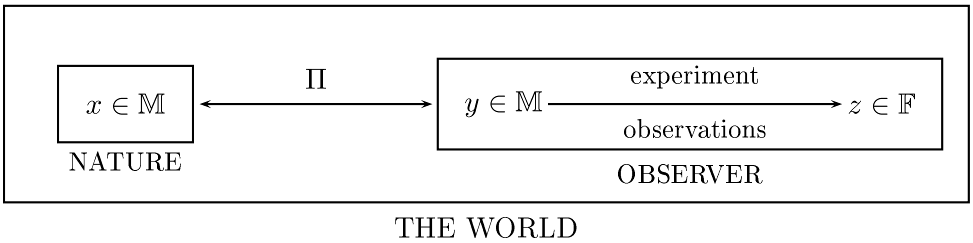

Figure 1. Some key elements of the philosophical considerations

The *classical world*, $\mathcal{W}_1$, is characterized by the interactor Π_1 given by

$$\Pi_1(x, y) = x\,. \tag{1}$$

This is the *world of observable truth* or, expressed somewhat differently, a world where *truth is learnable.* As another extreme, consider $\mathcal{W}_0$, conceived as a *black hole* and characterized by the interactor Π_0 given by

$$\Pi_0(x, y) = y\,. \tag{2}$$

In this world, no matter what Observer does, he will only see a mirror image of himself – it is a world of total narcissism, *what you see is what you believe.* By contrast, in a classical world, *what you see is what is true.*

We also consider mixtures of the two worlds identified above. For this to make sense, we assume that $\mathbb{M}$ is embedded in a linear space. Then, to each $q \in \mathbb{R}$, we may consider the world characterized by the interactor Π_q defined by

$$\Pi_q(x, y) = qx + (1 - q)y\,. \tag{3}$$

The worlds defined in this way are the *Tsallis worlds*, denoted $\mathcal{W}_q$.

All worlds which we will consider will be *sound* in the sense that $\Pi(x, y) = x$ provided there is a *perfect match* (*belief matches truth*), i.e. provided $y = x$.

In order to enable quantitative reasoning, we introduce functions which determine the *effort* needed by Observer in order to acquire knowledge. For reasons discussed in Section 7, these functions are called *score functions.* In principle, any function $\Phi : \mathbb{M} \times \mathbb{M} \to [0, \infty]$ could be a score function [1]. Of course, in order for the intended interpretation to make sense, Φ should be defined in some meaningful way pertaining to the special world and the special situation considered. Only then can $\Phi(x, y)$ be taken to represent the necessary effort needed by Observer in order to gain the knowledge $z = \Pi(x, y)$ in a situation with truth-instance x and belief-instance y. We claim that the appropriate selection of a score function will depend on tools of "description" available to Observer. This view will become more clear when, in the next section, we turn to probabilistic modelling.

We imagine that Observer has many sensible score functions to choose from. Of particular interest are *proper score functions* which are score functions which satisfy the inequality

$$\Phi(x, y) \geq \Phi(x, x) \tag{4}$$

[1] negative values could be allowed and would be convenient for certain wider studies, especially for enabling a smooth treatment of Kullback's *minimum information discrimination principle.*

for any $(x, y) \in \mathbb{M}\times\mathbb{M}$ and for which equality in (4) only holds in case of a perfect match ($y = x$). The implied variational principle, viz. for Observer to choose, whenever possible, a proper score function among the available score functions, we refer to as the *Perfect Match Principle* (PMP). We may allow *singular* cases of PMP for which equality in (4) can take place in other cases than the perfect match case $y = x$. We shall always emphasize if such singular cases are allowed.

The significance of proper score functions can be illuminated by the following considerations for which we change a bit the role of Nature. What we shall imagine is that Nature can communicate with Observer. Then we talk of an *expert* (by the name "Expert") rather than Nature. Consider some situation and assume that Observer seeks the advice of Expert. Experts best advice is x. However, for dubious reasons, Expert may be tempted to give an advice y which differs from x. With the ongoing crisis in world economy in mind, you may think of a bank advisor and a situation involving investment planning. With access to a proper score function, Observer can enter into a contract with Expert which will encourage Expert to be honest, i.e. to give the advice reflecting a perfect match, $y = x$. The contract may be formulated as an agreement for Observer to pay a one-and-for-all sum on signing the contract supplied with a kind of insurance scheme according to which Expert must pay a penalty in the amount of $\Phi(x, y)$ as soon as the true nature of Experts insight, in the form of x, will be revealed to Observer. Clearly, when Φ is a proper score function, it is in Experts own interest to be honest, i.e. to give the advice $y = x$, as this will minimize the penalty to be payed to Observer. Considerations of this nature emerged in statistical studies as detailed in Section 7.

Essential for our analysis are the two concepts, *interaction* (given by Π) and *score* or *effort* (given by Φ). When we work in a world $\mathcal{V}_\Pi$ and Observer has chosen the score function Φ, the resulting model is denoted $\mathcal{V}(\Pi, \Phi)$. Consider such a model and assume that Φ is a proper score function. We may then introduce the important notions *entropy* and *divergence*, here denoted by the letters S and D. Entropy relates to a possible truth instance x, whereas divergence relates to a situation characterized by a pair (x, y) of associated truth- and belief instances:

$$\mathrm{S}(x) = \Phi(x, x) \tag{5}$$

$$\mathrm{D}(x, y) = \Phi(x, y) - \mathrm{S}(x)\,. \tag{6}$$

Thus, conceiving Φ as effort, *entropy is minimal effort* and *divergence is excess* or *redundant effort*. Though infinite values are conceivable, we ignore this problem and simply assume that D can be defined so that $\mathrm{D}(x, y) = 0$ if and only if $y = x$ and so that the *linking identity*

$$\Phi(x, y) = \mathrm{S}(x) + \mathrm{D}(x, y) \tag{7}$$

always holds.

Note that the notions introduced require a proper score function and that entropy is a derived quantity. In consistency with this observation, the thesis that *entropy should never be considered as an isolated quantity* seems to represent a sound and fruitful guiding principle.

The fact that $\mathrm{D}(x, y) \geq 0$ with equality if and only if there is a perfect match is the *fundamental inequality of information theory* (FI), here with information theory understood in a rather general abstract sense.

3. Probabilistic modelling

For our probabilistic modelling, situations are related to a discrete *alphabet* $\mathbb{A}$ and the set of possible truth instances is taken to be the set of probability distributions over $\mathbb{A}$, in symbols $\mathbb{M} = M_+^1(\mathbb{A})$. A typical truth instance is given by the point probabilities: $x = \big(x_i\big)_{i\in\mathbb{A}}$, thus the x_i's are non-negative numbers adding to 1. Similarly, a belief instance is a probability distribution

$y = \big(y_i\big)_{i \in \mathbb{A}} \in \mathbb{M}$ [2]. For a knowledge instance $z = \big(z_i\big)_{i \in \mathbb{A}}$ we only require that the z_i's are well defined real numbers.

We assume that the interactors to be considered act *locally*, i.e., for some function π defined on $[0,1] \times [0,1]$,

$$\Pi(x,y) = \Big(\pi(x_i.y_i)\Big)_{i \in \mathbb{A}}.$$

The function π is the *local interactor* or just the *interactor*. In order to ensure that the corresponding global interactor is sound, we assume that π is *sound*, i.e. that $\pi(s,s) = s$ for all $0 \le s \le 1$. We also assume that suitable regularity conditions are fulfilled, say continuity and continuous differentiability in the interior domain and, further, that π assumes finite values, except possibly for cases with $\pi(s,t) = \infty$ when $t = 0 < s$.

For the probabilistic models, the world defined by π is denoted $\mathcal{V}_\pi$. The primary example is the *q-interactor* π_q given by

$$\pi_q(s,t) = qs + (1-q)t\,. \tag{8}$$

We put $\mathcal{V}_q = \mathcal{V}_{\pi_q}$. The worlds defined in this way are the *Tsallis worlds* in the probabilistic setting.

The equation (8) expresses a linear (or affine) relationship among probabilities. It is a form of *direct linearity*. It is conceivable that, instead, physical circumstances dictate that a linear relationship only applies to certain function values applied to point probabilities. This points to interactors π_q^ξ of the form

$$\pi_q^\xi(s,t) = \xi^{-1}\Big(\pi_q\big(\xi(s), \xi(t)\big)\Big)\,. \tag{9}$$

For this we require that the function ξ is a smooth strictly increasing function on $[0,1]$. When we work with one of these interactors, we speak about *ξ-linearity* and the associated worlds are denoted $\mathcal{V}_q^\xi$. In mathematical terms, the interactors π_q^ξ are *generalized meanvalues*; for a classical reference, see [4].

Of course, for $q = 1$ or $q = 0$ nothing new is obtained in this way. But for other values of q, interesting interactors emerge. For instance, with $\xi(s) = \ln s$ we find that π_q^ξ determines the *geometric average* $(s,t) \mapsto s^q t^{1-q}$. We write this interactor as π_q^G and the associated world as $\mathcal{V}_q^G$ (G for "geometric"), whereas π_q may be written as π_q^A and $\mathcal{V}_q$ as $\mathcal{V}_q^A$ (A for "arithmetic").

Concerning the value of q, the "cleanest" results are obtained for $0 < q \le 1$, but also values $q > 1$ are of interest. We find negative values of q less interesting for reasons given in Theorem 2.

An interactor π is *weakly consistent* if $\sum_{i \in \mathbb{A}} z_i = 1$ for every pair $(x,y) \in \mathbb{M} \times \mathbb{M}$ with $z = \pi(x,y)$. If even $z \in \mathbb{M}$ can be concluded, the interactor is *strongly consistent*.

Proposition 1. *The only worlds $\mathcal{V}_\pi$ with a weakly consistent interactor π are the Tsallis worlds $\mathcal{V}_q$ for some real value of q, and only for $0 \le q \le 1$ do we obtain a world with a strongly consistent interactor.*

The proof follows a standard pattern and is not given here.

We turn to a closer investigation of worlds $\mathcal{V}_\pi$ with the specification of an associated score function. The first thing to do is to agree on what the sensible score functions for $\mathcal{V}_\pi$ are. Just as for the global interactor, they should act locally. Therefore, the key question is what the "local score" associated with a basic event which has true probability s and believed probability t should be. For this we note that $\pi(s,t)$ is the "force" by which the basic event is perceived

[2] Finer modelling will allow that the set of available belief instances is different from $\mathbb{M}$, or even depends on the actual truth instance; an example of this is to allow that y is an incomplete distribution ($\sum y_i < 1$) and to insist that $y \succ x$ i.e. that $x_i > 0 \Rightarrow y_i > 0$.

by Observer. Then we have to consider what effort Observer should attach to the basic event. As Observer does not know s, this can only depend on Observers own belief, represented by t. Thus, for some function κ, Observer should attach the effort $\kappa(t)$ to a single occurrence of the basic event. This should be multiplied by the perceived force. We conclude that the local contribution to the score function should be $\pi(s,t)\kappa(t)$. Based on such more or less – more, hopefully – convincing intuitive ideas, we agree that every acceptable score function is of the form

$$\Phi(x,y) = \sum_{i\in\mathbb{A}} \pi(x_i,y_i)\kappa(y_i) \tag{10}$$

with κ some function defined on $[0,1]$. The function κ is the *descriptor* and we say that κ *generates* Φ. We may interpret κ as the *cost of information*, viz. $\kappa(t)$ is the price Observer is willing to pay or, in other words, the effort (energy) Observer is willing to allocate in order to obtain the information carried by an event with (believed) probability t. We also refer to $\kappa(t)$ as the *description cost* allocated to a probability-t event, though, for non-classical worlds, we are not able to suggest concrete methods of description – such as those based on *coding* – which could be relevant in this respect. Observe that the score $\Phi(x,y)$ is the *perceived average cost of description* or, as the acquisition of knowledge or *information* is concerned, *the perceived average cost of information*. The corresponding *true average cost*, which is the average allocation of description effort by Observer but seen from the point of view of Nature, is

$$\Phi_{\text{true}}(x,y) = \sum_{i\in\mathbb{A}} x_i\kappa(y_i)\,. \tag{11}$$

On the technical side we impose the following conditions on a descriptor: $\kappa(1) = 0$, κ is continuous on $[0,1]$, finite on $]0,1]$, continuously differentiable on $]0,1]$ and, finally, we impose a condition of *normalization*, viz. that $\kappa'(1) = -1$. The last condition corresponds to a choice of *unit*. The choice made gives *natural units*, nats. Had we, instead, imposed the condition $\kappa'(1) = \ln\frac{1}{2}$, we would have obtained *binary units*, bits. We find that 1 nat $\approx$ 1.4427 bits.

The descriptor only determines the score function when the interactor is known. Thus, in (10), we may write $\Phi(x,y|\pi,\kappa)$ or similar for clarification. Normally, π and κ will be understood from the context and we may drop these letters from the notation. To specify that we work in the world with interactor π and consider the score function generated by π and the descriptor κ, we refer to this setting as $\mathcal{V}(\pi,\kappa)$.

Consider $\mathcal{V}(\pi,\kappa)$ and assume that the associated score function is proper. Entropy and divergence are then obtained from (5) and (6). We find:

$$\mathrm{S}(x) = \sum_{i\in\mathbb{A}} x_i\kappa(x_i)\,, \tag{12}$$

$$\mathrm{D}(x,y) = \sum_{i\in\mathbb{A}} \Big(\pi(x_i,y_i)\kappa(y_i) - x_i\kappa(x_i)\Big) \tag{13}$$

$$= \sum_{i\in\mathbb{A}} \delta(x_i,y_i)\,, \tag{14}$$

where δ denotes a special function, the *divergence generator* which is defined by

$$\delta(s,t) = \Big(\pi(s,t)\kappa(t) + t\Big) - \Big(s\kappa(s) + s\Big)\,. \tag{15}$$

Clearly, entropy as defined by (12) is a well defined, possibly infinite quantity. The fact that also the two formulas for divergence give well defined quantities is not immediately obvious. In fact this follows in a quite fundamental way, alas depending on a as yet unsolved technical problem. Thus we can only formulate as a conjecture the following statement:

Conjecture 1. *For $\mathcal{V}(\pi, \kappa)$, the score function defined by π and κ is a proper score function if and only if $\delta(s,t) \geq 0$ for every $(s,t) \in [0,1] \times [0,1]$ with equality only for $t = s$.*

The inequality $\delta \geq 0$ (with equality only as stated) we refer to as the *pointwise fundamental inequality* (PFI). Clearly, it implies the fundamental inequality (FI) and thus the properness of the associated score function. In practice, PFI is a much simpler inequality to verify than FI attacked directly. The missing proof of the "only if" part of the conjecture has no practical consequences as all concrete positive results we need are proved via PFI.

It remains to investigate if the descriptor can be adapted to the interactor in a way so that a proper score function emerges. To answer this question, we shall need the function χ defined by

$$\chi(t) = \frac{\partial \pi}{\partial t}(t,t) \tag{16}$$

(so χ is the partial derivative of $\pi(s,t)$ in direction of t, evaluated at the diagonal $s = t$).

Theorem 1. *Let π be an interactor and assume that χ defined by* (16) *is bounded in the vicinity of* 1. *Then, there exists at most one descriptor κ which generates a proper score function for the world $\mathcal{V}_\pi$. Indeed, if κ is such a descriptor, κ must be the unique solution to the differential equation*

$$t\kappa'(t) + \chi(t)\kappa(t) + 1 = 0 \tag{17}$$

which satisfies the condition $\kappa(1) = 0$.

Proof. Assume that the score function associated with the descriptor κ is proper. For $0 < t < 1$ put

$$f(t) = t\kappa'(t) + \chi(t)\kappa(t)\,.$$

Consider a probability vector $x = (x_1, x_2, x_3)$ with positive point probabilities. Then the function F given by

$$F(y) = F(y_1, y_2, y_3) = \sum_{i=1}^{3} \pi(x_i, y_i)\kappa(y_i)$$

assumes its minimal value for the interior point $y = x$ when restricted to probability distributions. As standard regularity conditions are fulfilled, there exists a Lagrange multiplier λ such that

$$\frac{\partial}{\partial y_i}\big(F(y) - \lambda \sum_{i=1}^{3} y_i\big) = 0 \text{ for } i = 1,2,3$$

when $y = x$. This shows that $f(x_1) = f(x_2) = f(x_3)$.

Using this, first with $(x_1, x_2, x_3) = (\frac{1}{2}, x, \frac{1}{2} - x)$ for a value of x in $]0, \frac{1}{2}[$, and then with $(x_1, x_2, x_3) = (x, \frac{1}{2}(1-x), \frac{1}{2}(1-x))$, one is soon led to the conclusion that f is constant on $]0,1[$. In view of the boundedness condition on χ which is tacitly assumed (weaker conditions will do), the constant must be -1 and the result follows. □

We refer to the one and only possible descriptor which could generate a proper score function for $\mathcal{V}_\pi$ as the *basic candidate* and denote it by $\kappa[\pi]$. As the boundedness condition imposed on χ in Theorem 1 is very weak (perhaps even superfluous), the theorem may be stated briefly by saying that the basic candidate is uniquely defined, given the interactor. The result supports the conjecture. Indeed, if κ generates a proper score function, we may introduce, for each s, the function f_s defined by $f_s(t) = \pi(s,t)\kappa(t) + t$ and then note that by (17), each f_s has a stationary point at $t = s$. As, for every $(x,y) \in \mathbb{M} \times \mathbb{M}$,

$$\sum_{i \in \mathbb{A}} f_{x_i}(y_i) \geq \sum_{i \in \mathbb{A}} f_{x_i}(x_i)\,,$$

it appears plausible that the stationary points of the functions f_s are all minimal points, and this will imply that the conjecture holds.

The conjecture and its relation to PFI tempts us to define an *adjusted* notion of the total description effort, denoted by $\tilde{\Phi}$: $\tilde{\Phi}(x,y) = \sum_{i\in\mathbb{A}}\big(\pi(x_i,y_i)\kappa(y_i)+y_i\big)$. The added terms in the summands, the y_i's, are interpreted as the contributions to the *total overhead* stemming from the respective basic events. Here, "overhead" is related to whatever action, typically observations of an experiment, is involved for Observer in order to acquire the knowledge $z = \Pi(x,y)$. The total overhead in any situation is $\sum y_i = 1$. In other words, the normalization condition $\kappa'(1) = -1$ corresponds to choosing the overhead cost as the unit to work with. Adjusting also the entropy function, one finds that *adjusted entropy* is always bounded below by the overhead cost, 1 nat.

One should of course respect the fact that Theorem 1 only provides a *candidate* for a well behaved descriptor. In each case one has to check if this candidate really provides a descriptor which generates a proper score function. When this is the case, $\kappa = \kappa[\pi]$ is the *ideal descriptor* associated with π and we will typically assume that we work with this descriptor when the world $\mathcal{V}_\pi$ is considered. Notationally we shall write $\mathcal{V}(\pi)$ rather than $\mathcal{V}(\pi,\kappa)$ to indicate this.

In case of the Tsallis worlds, the determination of the basic candidate as well as the necessary checking whether or not it is ideal is straight forward and leads us to a main result. Before stating it, we find it convenient, following Tsallis [5], to introduce the *deformed logarithms* $\ln_q$. For $q \in \mathbb{R}$, they are defined by

$$\ln_q t = \begin{cases} \ln t \text{ if } q = 1 \\ \frac{1}{1-q}\big(t^{1-q}-1\big) \text{ otherwise}. \end{cases} \tag{18}$$

Theorem 2. *Consider one of the worlds $\mathcal{V}_q$. If $q < 0$, no descriptor defines a proper score function for $\mathcal{V}_q$. If $q \geq 0$, there exists a unique descriptor which defines a proper score function for $\mathcal{V}_q$, viz. the descriptor κ_q given by*

$$\kappa_q(y) = \ln_q \frac{1}{y}. \tag{19}$$

All worlds $\mathcal{V}(\pi_q,\kappa_q)$ with $q > 0$ are regular, whereas the world $\mathcal{V}(\pi_0,\kappa_0)$, a black hole, is singular, in fact divergence vanishes identically in this case.

Proof. The function χ from Theorem 1 is the constant function $1-q$. Solving (17), you find that $\kappa[\pi_q]$ is given by (19). From simple examples, one finds that for $q < 0$, the perfect match principle does not hold with this descriptor.

It remains to consider values $q \geq 0$. The cases $q = 0$ and $q = 1$ are left to the reader. For the remaining cases, consider the divergence generator

$$\delta_q(s,t) = \frac{q}{1-q}st^{q-1} + t^q - \frac{1}{1-q}s^q. \tag{20}$$

An application of the geometric-arithmetic mean inequality shows that PFI holds (consider the cases $0 < q < 1$ and $q > 1$ separately and collect the two positive terms). $\square$

Appropriate explicit formulas for score or effort (Φ_q), entropy (S_q), and divergence (D_q) in the Tsallis worlds are easily derived from (10), (12)-(14) and (18)-(20). By Theorem 2, only cases with $q > 0$ yields ideal, non-singular descriptors, hence also proper score functions.

By Theorem 1, you are led in a unique way from an interactor to a descriptor. A natural question arises, that is, if the converse is also true. It is not. Thus, one cannot in general know which world you operate in, i.e. know the interactor, if one only knows the descriptor. Even if the descriptor is ideal, the world cannot be determined. Concrete illustrations of this fact are provided by the following result:

Proposition 2. *For $q > 0$, $\mathcal{V}_q^A$ and $\mathcal{V}_q^G$ both have κ_q given by* (19) *as ideal descriptor.*

We leave the proof to the reader, only noting that for $q > 1$, the fact that κ_q is the ideal descriptor for $\mathcal{V}_q^G$ is a weaker statement than the similar statement – contained in Theorem 2 – for $\mathcal{V}_q^A$ as then $\pi_q^G \geq \pi_q^A$. For $0 < q < 1$, the statement is stronger as then $\pi_q^G \leq \pi_q^A$. Therefore, the statement requires a proof which is in fact easy to accomplish.

As a consequence of our observation note the following: If we only focus on entropy S_q, and this only needs the descriptor κ_q for its specification, we cannot know the score- or the divergence functions. Both Φ_q^A and D_q^A and Φ_q^G and D_q^G are possible sets of associated functions, the one set corresponding to the interactor π_q^A, the other to π_q^G.

4. Preparations and controls

For this section, as well as for Sections 5 and 6, we work in a world $\mathcal{V} = \mathcal{V}(\pi) = \mathcal{V}(\pi, \kappa)$ with κ the ideal descriptor associated with π. We assume that κ is strictly decreasing on $[0, 1]$. The key example is $\pi = \pi_q$ for a $q > 0$. The situations we have in mind all involve distributions over the same alphabet $\mathbb{A}$.

The main issue we shall now discuss is the philosophical question "what *can* Observer know?". What we mean by this is that in any concrete situation, the set of truth instances chosen by Nature will normally be restricted to some subset of $\mathbb{M}$. Such a subset we call a *preparation*. Typically, we denote a preparation by the letter $\mathcal{P}$. But not every subset of $\mathbb{M}$ can, realisticly, occur. The preparations which can actually be realized, we call the *feasible preparations*. They tell us what Observer can know.

In order to arrive at a meaningful, concrete and operational definition of what a feasible preparation is, we put forward the rough idea that what Observer *can* do to enable the acquisition of knowledge is to fix allocations of effort (or energy) to individual events and also to fix an overall threshold and then enforce Nature to choose only truth instances which respect the indicated constraints.

First, regarding the allocation of efforts, we adopt the view that *belief is a tendency to act*. Therefore, we start from a belief instance ξ and transform it to a more suitable object when having possible actions by Observer in mind. The transformation is denoted $\xi \curvearrowright \hat{\xi}$ with $\hat{\xi}$ defined as the family $\hat{\xi} = \big(\kappa(\xi_i)\big)_{i \in \mathbb{A}}$. We find it convenient to define a *control* as a family $w = (w_i)_{i \in \mathbb{A}}$ for which there exists $\xi \in \mathbb{M}$ such that $w = \hat{\xi}$, i.e. such that $w_i = \kappa(\xi_i)$ for each $i \in \mathbb{A}$.

Secondly, we imagine that, corresponding to a chosen control $w = \hat{\xi}$, Observer can realize a certain *experimental set-up* which consists of various parts such as machinery, instruments and so on, including a special handle which he uses to fix the *level of total effort*, h. The idea then is that this results in a restriction of the truth instances chosen by Nature to the set of $x \in \mathbb{M}$ for which the average effort as perceived by Observer is h, i.e. $\Phi(x, \xi) = h$. The preparation obtained by restricting the truth instances in this way is denoted $\mathcal{P}(w, h)$, i.e. (still with $w = \hat{\xi}$)

$$\mathcal{P}(w, h) = \{x \in \mathbb{M} | \Phi(x, \check{w}) = h\} . \tag{21}$$

Note that these preparations are *level sets* obtained by fixing the belief instance in the score function. Preparations of this form are called the *basic preparations*.

Having chosen a control, the scene is set and observations can begin with the reading of measuring instruments etc. Observer may want to use the same experimental set-up for several experiments by adjusting the level of total effort. Of course, the level should always be set so that $\mathcal{P}(w, h)$ is non-empty.

We can now give a precise definition of a feasible preparation: It is a non-empty finite intersection of basic preparations. The *genus* of a feasible preparation is the smallest number of basic preparations needed to define the preparation. Thus, a *feasible preparation of genus* 1 is the same as a basic preparation.

Regarding the previously introduced concept of "situations" , we emphasize that they should always be related to feasible preparations.

By $\mathbb{K}$ we denote the set of controls, i.e. $\mathbb{K} = \{\hat{\xi}|\xi \in \mathbb{M}\}$. The choice of the letter "K" reflects that for the classical world, a control is strongly related to coding. For this reason we may also use the term *coder* instead of control. The transformation $\xi \curvearrowright \hat{\xi}$ is a bijection between $\mathbb{M}$ and $\mathbb{K}$. In order to determine the inverse transformation, denoted $w \curvearrowright \check{w}$, we need to know the inverse function of κ. This function we call the *probability checker* and denote by $\rho = \rho_\kappa$. We insist that ρ is defined on all of $[0, \infty]$ even though there may be a *cut-off* in κ, as the range $[0, \kappa(0)]$ may be a proper subset of $[0, \infty]$. The formal definition is as follows:

$$\rho(a) = \begin{cases} \kappa^{-1}(a) \text{ if } a \leq \kappa(0) \\ 0 \text{ otherwise.} \end{cases} \tag{22}$$

The probability checker provides a tool to determine how "complex" events you can describe with access to a given number of nats. The lower the probability, the more complex the event. With access to a nats, you can describe any event with a probability as low as $\rho(a)$.

For a Tsallis world with $0 < q \leq 1$, however large your resources to nats, there are events so complex that you cannot describe them, whereas, if $q > 1$, you can describe *any* event if you have access to K nats if only K is large enough ($K \geq \frac{1}{q-1}$).

Another illustration: If, for sample point $i \in \mathbb{A}$ you have decided how many nats you are willing to allocate to i, say a_i nats, this will only be feasible if the *generalized Kraft inequality* holds:

$$\sum_{i \in \mathbb{A}} \rho_\kappa(a_i) \leq 1\,. \tag{23}$$

5. Preparations and games

Consider some preparation $\mathcal{P}$. For the time being, we need not assume that $\mathcal{P}$ is feasible. Observer cannot know which truth instance Nature has chosen, except that it is in $\mathcal{P}$. It lies nearby to speculate if Observer can identify some "typical" element of $\mathcal{P}$, perhaps in order to use that as an appropriate belief instance. The key to actually do so is to view the interplay between Nature and Observer as a *game*. The game should be a two-person zero-sum game with Nature and Observer fighting over description effort. Nature is a maximizer and Observer a minimizer.

The available *strategies* for Nature are limited to truth instances in the preparation $\mathcal{P}$. As for Observer, we have chosen to take the set $\mathbb{K}$ of controls, rather than the set $\mathbb{M}$ of belief instances as strategies. We find this most natural in view of the interpretations. If one wishes, a transformation of definitions and results to involve only belief instances is possible, though at times a bit awkward. In view of our focus on controls, we introduce a variant $\Psi : \mathbb{M} \times \mathbb{K} \to [0, \infty]$ of the score function by agreeing that

$$\Psi(x, w) = \Phi(x, \check{w})\,. \tag{24}$$

A further consequence of the change of focus concerns divergence which is better conceived as a *redundancy* of effort corresponding to the given control. Denoting redundancy with R, we define

$$\mathrm{R}(x, w) = \mathrm{D}(x, \check{w})\,. \tag{25}$$

We use the notation $\gamma(\mathcal{P})$ for the game defined above. The two *values* of the game are defined as usual, cf. [6] or [7], for instance. The value seen from the point of view of Nature is

$$\sup_{x \in \mathcal{P}} \inf_{w \in \mathbb{K}} \Psi(x, w)\,. \tag{26}$$

As the infimum here is nothing but the entropy $\mathrm{S}(x)$, we find that the value given by (26) is the *maximum entropy value* associated with $\mathcal{P}$:

$$\mathrm{S}_{\max}(\mathcal{P}) = \sup_{x\in\mathcal{P}} \mathrm{S}(x)\,. \tag{27}$$

As for the other value of the game, it is given by

$$\inf_{w\in\mathbb{K}} \sup_{x\in\mathcal{P}} \Psi(x,w)\,. \tag{28}$$

For every $w \in \mathbb{K}$, the associated *individual risk* is defined by

$$\mathrm{Ri}(w) = \mathrm{Ri}(w|\mathcal{P}) = \sup_{x\in\mathcal{P}} \Psi(x,w)\,. \tag{29}$$

The *minimal risk value* associated with $\mathcal{P}$ is then defined as

$$\mathrm{Ri}_{\min}(\mathcal{P}) = \inf_{w\in\mathbb{K}} \mathrm{Ri}(w)\,, \tag{30}$$

i.e. as the value given by (28).

Clearly, $\mathrm{S}_{\max}(\mathcal{P}) \leq \mathrm{Ri}_{\min}(\mathcal{P})$. This is the *minimax inequality*. If equality holds and defines a finite quantity, the game is in *game theoretical equilibrium* or just in *equilibrium*. An *optimal strategy for Nature* is the same as a truth instance in $\mathcal{P}$ with maximal entropy. An *optimal strategy for Observer* is the same as a control w with $\mathrm{Ri}(w) = \mathrm{Ri}_{\min}$.

Another concept of equilibrium is related to *robustness*: A control w^* is *robust* if, for some $h \in \mathbb{R}$, $\Psi(x,w^*) = h$ for all $x \in \mathcal{P}$. The number h is the *level of robustness*. If Observer chooses a robust strategy, the description effort will be independent of which strategy Nature has chosen. There is an important connection between the two concepts of equilibria:

Theorem 3 (robustness and pythagorean inequalities). *Assume that $x^* \in \mathcal{P}$ and that $w^* = \hat{x^*}$ is robust with robustness level h. Then $\gamma(\mathcal{P})$ is in equilibrium with h as value. Furthermore, for any $x \in \mathcal{P}$,*

$$\mathrm{S}(x) + \mathrm{R}(x,w^*) \leq \mathrm{S}_{\max}(\mathcal{P}) \tag{31}$$

and, for every control w,

$$\mathrm{Ri}(w) \geq \mathrm{S}_{\max}(\mathcal{P}) + \mathrm{R}(x^*,w)\,. \tag{32}$$

Proof. Though the argument is known and easy, cf. Theorem 6.2 of [8], we present it here too. Firstly, by assumption, $\mathrm{Ri}(w^*) = h = \Psi(x^*,w^*) = \mathrm{S}(x^*)$, hence the game is in equilibrium with x^* as optimal strategy for Nature and w^* as optimal strategy for Observer.

For any $x \in \mathcal{P}$ we find that $\mathrm{S}(x) \leq \mathrm{S}(x) + \mathrm{R}(x,w^*) = \Psi(x,w^*) = h$, i.e. (31) holds. And for any control w, $\mathrm{Ri}(w) \geq \Psi(x^*,w)$, hence (32) follows from the linking identity (7). □

Equation (31), in the form $\mathrm{S}(x) + \mathrm{D}(x,x^*) \leq \mathrm{S}_{\max}(\mathcal{P})$, is the *pythagorean inequality* going back to Čencov [9] and Csiszár [10], and (32) is the *reverse pythagorean inequality*. From a game theoretical point of view, these inequalities are trivial consequences of Nash's *saddle value inequalities*.

6. Exponential families

We shall develop a simple criterion which facilitates the identification of situations of equilibrium and also facilitates the search for optimal strategies for both Nature and Observer. The first result will help us understand why the level sets play a central role. Actually, we first run into *sub-level sets*, defined by

$$\mathcal{P}_{\leq}(w,h) = \{x|\Psi(x,w) \leq h\} \tag{33}$$

Proposition 3. *Consider a truth instance x^* and a control w^*. Then a necessary and sufficient condition that there exists a preparation $\mathcal{P}$ for which the game $\gamma(\mathcal{P})$ is in equilibrium and has x^* and w^* as optimal strategies, is that $h^* = \Psi(x^*,w^*) < \infty$ and that $w^* = \hat{x^*}$. And when these conditions are fulfilled, the largest such set is the sub-level set $\mathcal{P}_{\leq}(w^*,h^*)$.*

Proof. (Compare with [8], Theorem 8.2). If, for some $\mathcal{P}$, $\gamma(\mathcal{P})$ is in equilibrium with x^* and w^* as optimal strategies then, by *Nash's saddle value inequalities*,

$$\Psi(x,w^*) \leq \Psi(x^*,w^*) \leq \Psi(x^*,w) \tag{34}$$

for every $x \in \mathcal{P}$ and every control w. Furthermore, $\Psi(x^*,w^*)$ is finite. Let $y^* = \check{w^*}$. Then, by the right hand inequality of (34), and as Φ is a proper score function, $\Phi(x^*,x^*) \geq \Phi(x^*,y^*) \geq \Phi(x^*,x^*)$, hence $\Phi(x^*,y^*) = \Phi(x^*,x^*)$ and we conclude that $y^* = x^*$. It follows that $w^* = \hat{x^*}$. It only remains to remark that the left hand inequality of (34) implies directly that $\mathcal{P} \subseteq \mathcal{P}_{\leq}(w^*,h^*)$ and the first part of the proof is complete.

For the second half, assume that $h^* = \Psi(x^*,w^*) < \infty$ and that $w^* = \hat{x^*}$. From the latter condition, the saddle value inequalities (34) are verified for the preparation $\mathcal{P}_{\leq}(w^*,h^*)$ and then, using also the former condition, equilibrium follows as well as optimality of the strategies x^* and w^*. □

The result points to a unique role for the sub-level sets. However, we hold the view that it is the highest level which is of relevance, and hence we stick to (intersections of) level sets, rather than sub-level sets. Also, this will allow us to exploit results on robustness.

These considerations lead to an expedient approach to the most common type of problems within statistical physics as well as many other branches of science where questions of entropy optimization and equilibrium come up. First, for a finite set $\mathbf{w} = (w_1,\cdots,w_n)$ of controls and a corresponding set of levels, $\mathbf{h} = (h_1,\cdots,h_n)$, we put $\mathcal{P}(\mathbf{w},\mathbf{h}) = \cap_{\nu=1}^{n}\mathcal{P}(w_\nu,h_\nu)$ (if non-empty, this is a feasible preparation of genus at most n) and for $\mathbf{w}$ fixed, we denote by $\mathbb{P}(\mathbf{w})$ the family of all feasible preparations of this form. The corresponding *exponential family*, denoted by $\hat{\mathcal{E}}(\mathbf{w})$, is the set of controls ε which are robust for all preparations in the family $\mathbb{P}(\mathbf{w})$. In terms of belief instances, the exponential family is the family $\mathcal{E}(\mathbf{w})$ of all belief instances ξ for which $\xi = \check{\varepsilon}$ for some $\varepsilon \in \mathcal{E}(\mathbf{w})$.

From the definitions introduced and from Theorem 3 on robustness you find the following simple, but useful result:

Theorem 4. *Consider a family $\mathbf{w} = (w_1,\cdots,w_n)$ of controls and the associated family $\mathbb{P}(\mathbf{w})$ of preparations. Let x^* be a truth instance, ε^* the corresponding control ($\varepsilon^* = \hat{x^*}$) and assume that $\varepsilon^* \in \hat{\mathcal{E}}(\mathbf{w})$. Put $\mathbf{h} = (h_1,\cdots,h_n)$ with $h_\nu = \Psi(x^*,w_\nu)$ for $\nu = 1,\cdots,n$. Then $\gamma\big(\mathcal{P}(\mathbf{w},\mathbf{h})\big)$ is in equilibrium and has x^* and w^* as optimal strategies. In particular, x^* is the maximum entropy distribution for the preparation $\mathcal{P}(\mathbf{w},\mathbf{h})$.*

A simplification occurs in case the world we consider is one of the Tsallis worlds $\mathcal{V}_q = \mathcal{V}(\pi_q,\kappa_q)$ for a $q > 0$ as then one can identify controls in the exponential family. The reason is that in view of the linear character of π_q, we realize that every control ε of the form $\alpha + \sum_1^n \beta_\nu w_\nu$ for suitable constants α and $\beta_1, \cdots, \beta_n$ is a member of $\check{\mathcal{E}}(\mathbf{w})$. Typically, for given $\beta_1, \cdots, \beta_n$ you adapt α to these numbers so that $\alpha + \sum_1^n \beta_\nu w_\nu$ is a genuine control. This requires that

$$\sum_{i \in \mathbb{A}} \rho_q\big(\alpha + \beta_1 w_{1,i} + \cdots + \beta_n w_{n,i}\big) = 1\,, \tag{35}$$

where ρ_q is the probability checker for κ_q, cf. (23). Note that for the classical case $q = 1$, an extra simplification is that, given the β's, α can be solved from this equation.

7. Discussion

The present study has been preceeded by [11] and [12], and will, according to plan, be followed by a more detailed publication. However, it may well be that the present exposition is the ideal starting point for any reader who will embark on research in the direction taken. In particular, we have made certain simplifying assumptions and also, we have toned down the very abstract considerations and mainly stayed within more standard probabilistic modelling.

The main result on the identification of Tsallis entropy, Theorem 2, is derived based on the assumption that, firstly, one should allow for an interaction between truth, belief and knowledge and, secondly, one should accept a rather innocent variational principle, that optimal performance is obtained when there is a perfect match between truth and belief. It should be emphasized that though these principles could, by some adversary, be viewed as axioms, they are intended as key elements of an operational interpretation. Further studies may justify this view, possibly based on a deeper insight into physics than here displayed. It is a bit disturbing to the author that, according to previous studies (also reported in the recent monograph [2]), Tsallis entropy with q negative is of importance for the discussion of certain physical phenomena, whereas our approach does not really accept negative values of q.

It is noted that our study is devoid of a dynamical dimension. The only implicit dynamical element is the natural succession in time of truth, belief and knowledge (via experiments and observations). Again, further research is needed on this point.

Another direction of research, which should fit well into the pronounced philosophical basis, would be an extension to also cover the quantum setting.

The author has put great emphasis on the choice of terminology. However, especially concerning one issue, the proper naming of what here appears as score functions, there may well be other attractive options. The term *inaccuracy*, following Kerridge [13], did not appeal to the author, *effort function* hits better the intended interpretation and is kept "in reserve" whereas *complexity* is overloaded and *description function* or similar is fine for the classical setting, but with no clarification of an eventual role of more physical description (or coding) outside the classical case, it would appear a bit premature. Finally, *score function* and *proper score function* were chosen as these notions appear in statistical decision theory and has a long history which can be traced from Csiszár, [14]. There, and in works referred to, one will also find what we have termed the Perfect Match Principle. One disadvantage with the term "score" is that it gives the impression of something attractive seen with Observers eyes, quite opposite to what is the case. Of course, one could say that a high score is attractive when seen with the eyes of Nature if you follow the game theoretical line of thought. Or one could say that when Observer *has* acquired knowledge, the score can be taken as a quantitative measure of the information gained. The situation is similar to the one concerning the appearance of both entropy and "negentropy" in certain texts. Without citing extensively from the literature on score functions, we point to [15] by Good, mainly because there we found the view that *belief is a tendency to act*, a view which fits nicely into the philosophy here adopted. As a final comment, note that when restricted to the classical setting, the consideration of proper scoring functions gives an elegant introduction to Shannon entropy which is more descriptive than axiomatic, we find, and which provides an attractive supplement to the more common approaches via coding.

As already indicated, further research on the fundamental nature of the quantities characterized is much desired. This also concerns a more complete interpretation of descriptors outside the classical case. In this connection, [16], [17] and references there as well as [18] may be relevant.

Regarding the, admittedly, rather brief treatment of maximum entropy optimization in Section 6, note that this does not rely on the introduction of Lagrange multipliers, but uses a more direct approach in line with [19]. Also note that the usual introduction of *partition functions* does not, apparently, generalize in an adequate way when we study generalized entropies. In contrast, the *log-partition function* does generalize in a natural way, viz. as the solution α to (35).

Further concrete results on proper scoring functions may be developed by going outside the probabilistic setting, say by considering more general real valued functions than distributions and other concrete generalized meanvalues than the two chosen here, the arithmetic- and the geometric meanvalues.

The connection (duality) between feasible preparations and exponential families might fit well into geometric ideas as developed by Amari and his school, cf. [20]. Other possibilities to apply duality considerations are rather obvious, but not developed here. In this connection, a somewhat different and purely axiomatic approach developed in a preliminary setting in [21] should also be of relevance.

The untraditional approach in Section 6 to the much studied notion of exponential families is emphasized. It appears to be natural with reference to reasonably sound interpretations (via controls etc.) and also, the approach applies in other contexts than probabilistic ones, e.g. related to geometry (an application to Sylvesters problem of location theory is briefly indicated in [22]). The key role of game theory is evident. To the author, it is surprising that general and rather simple results as developed by Nash are so powerful – beyond what many researchers have presently realized, it seems. For the classical case, the game theoretical approach was studied in [8]. From that reference, we point to the discussion after Corollary 4.2 and to the important Theorem 6.9 which shows that no opportunities are missed when focusing on robustness as is done in our definition of exponential families.

For other works which emphasize the role of exponential families for statistical physics, see Naudts, [23] and [24]. These works have some similarity with the present approach.

Finally, some comments on the appearance of Tsallis entropy. The first publications are Havrda and Charvát [25] and, independently, Daróczy [26]. The latter author emphasized a characterization via functional equations, cf. also [27] and the more recent reference work [28]. The first appearance in the physical literature is due to Lindhard and Nielsen [29]. Subsequently, Lindhard gave a careful treatment of aspects of the measuring process, cf. [30], which may also be of relevance to our treatment. The trend-setting publication [1] from 1988 by Tsallis marks the efficient introduction of the generalized entropies within the physical community. At the time of publication, Tsallis was unaware of the earlier research. Regarding [29] and [30], these papers were largely ignored, though there is a casual reference to Lindhard's work in [31].

The success of Tsallis in launching the entropy measures which now bear his name is due to the direct approach and the fact that when combined with Jaynes *Maximum Entropy Principle*, cf. [32], main problems of statistical physics lead to *power laws*, a popular class of distributions when heavy-tailed distributions are needed.

The present approach is in line with earlier game theoretical considerations, cf. [19]. Because of a relation to Bregman divergences, we also point the reader to [23] and works referred to there.

Acknowledgments

Bjarne Andresen introduced me to Tsallis entropy, Imre Csiszár provided hints to the literature on score functions and Stig Stenstrup pointed me to the works by Lindhard and Nielsen. Interest and discussions with Jan Naudts, Hiroki Suyari and Constantino Tsallis before and during the workshop in Kyoto on generalized entropies is gratefully acknowledged, as are comments, mainly on terminology and philosophy, from Peter Harremoës and Nisheeth Srivastava.

References

[1] Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Physics* **1988**, *52*, 479–487.

[2] Tsallis, C. *Introduction to Nonextensive Statistical Mechanics.* Springer, Berlin Heidelberg, 2009.

[3] Wikipedia contributers. Bayesian probability **2009**.

[4] Hardy, G. H.; Littlewood, J. E.; Pólya, G. *Inequalities.* Cambridge University Press, second edition Ed., 1959.

[5] Tsallis, C. What are the numbers that experiments provide? *Quimica Nova* **1994**, *17*, 468.

[6] Aubin, J. P. *Optima and equilibria. An introduction to nonlinear analysis.* Springer, Berlin, 1993.

[7] Cesa-Bianchi, N.; Lugosi, G. *Prediction, learning and games.* Cambridge University Press, Cambridge, 2006.

[8] Harremoës, P.; Topsøe, F. Maximum entropy fundamentals. *Entropy*(Sept. 2001), *3*, 191–226.

[9] Čencov, N. N. *Statistical Decision Rules and Optimal Inference.* Nauka, Moscow, 1972. In russian, translation in "Translations of Mathematical Monographs", 53. AmericanMathematical Society, 1982.

[10] Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.

[11] Topsøe, F. Truth, Belief and Experience – a route to Information. *Journal of Contemporary Mathematical Analysis – Armen. Acad. Scienc.* **2009**, *44*, 105–110.

[12] Topsøe, F. On truth, belief and knowledge. In *2009 IEEE International Symposium on Information Theory.* IEEE, Washington, (June 2009), pp. 139–143.

[13] Kerridge, D. F. Inaccuracy and inference. *J. Roy. Stat. Soc. B.* **1961**, *23*, 184–194.

[14] Csiszár, I. Axiomatic Characterizations of Information Measures. *Entropy* **2008**, *10*, 261–273.

[15] Good, I. J. Rationel decisions. *J. Royal Statist. Soc., Series B* **1952**, *14*, 107–114.

[16] van der Lubbe, J. C. A. On certain coding theorems for the information of order α and oftypeβ. In *Trans. Eighth Prague Conf. on Inform. Theory, Statist. Decision Functions,RandomProcesses.* Czech. Acad. Science, Academia Publ., Prague, 1978. Prague, 1979.

[17] Ahlswede, R. Identification Entropy, In *General Theory of Information Transfer and Combinatorics*; *Lecture Notes in Computer Science*, *4123*, 595–613. Springer, Berlin, 2006.

[18] Suyari, H. Tsallis entropy as a lower bound of average description length for the q-generalized code tree. In *Proceedings ISIT 2007.* IEEE, 2007, pp. 901–905.

[19] Topsøe, F. Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics. In *Complexity, Metastability, and Non-Extensivity, CTNEXT07*; Abe, Hermann, Q. R. T., Ed., 2007, Vol. 965, *AIP Conference Proceedings*, pp. 104–113.

[20] Amari, S.; Nagaoka, H. *Methods of Information Geometry*, Vol. 191, *Translations of Mathematical monographs.* Oxford University Press, 2000.

[21] Topsøe, F. Game Theoretical Optimization inspired by Information Theory. *J. Global Optim.* **2009**, pp. 553–564.

[22] Topsøe, F. Intrinsic Methods for Optimization Problems. In *2008 IEEE International Symposium on Information Theory.* IEEE, Washington, (July 2008), pp. 1627–1630.

[23] Naudts, J. Generalised exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.

[24] Naudts, J. The q-exponential family in statistical physics. *Cent. Eur. J. Phys.* **2009**, *7*, 405–413.

[25] Havrda, J.; Charvát, F. Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika* **1967**, *3*, 30–35.

[26] Daróczy, Z. Generalized Information Functions. *Information and Control* **1970**, *16*, 36–51.

[27] Aczél, J.; Daróczy, Z. *On measures of information and their characterizations.* Academic Press, New York, 1975.

[28] B. Ebanks, P. S.; Sander, W. *Characterizations of Information Measures.* World Scientific, Singapore, 1998.

[29] Lindhard, J.; Nielsen, V. Studies in Statistical Dynamics. *Mat. Fys. Medd. Dan. Vid. Selsk.* **1971**, *38*, 1–42.

[30] Lindhard, J. On the Theory of Measurement and its Consequences in Statistical Dynamics. *Mat. Fys. Medd. Dan. Vid. Selsk.* **1974**, *39*, 1–39.

[31] Jaynes, E. T. Where do we Stand on Maximum Entropy?, In *The Maximum Entrropy Formalism*; Levine, R.; Tribus, M., Eds., pp. 1–104. M.I.T. Press, Cambridge, MA, 1979.

[32] Jaynes, E. T. Information theory and statistical mechanics, I and II. *Physical Reviews* **1957**, *106 and 108*, 620–630 and 171–190.