J Glob Optim (2009) 43:553–564
DOI 10.1007/s10898-008-9340-8

Game theoretical optimization inspired by information theory

Flemming Topsøe

Received: 28 June 2008 / Accepted: 12 August 2008 / Published online: 4 September 2008

Abstract Inspired by previous work on information theoretical optimization problems, the basics of an axiomatic theory of certain special two-person zero-sum games is developed. One of the players, "Observer", is imagined to have a "mind", the other, "Nature", not. These ideas lead to un-symmetric modeling as the two players are treated quite differently. Basic concavity- and convexity results as well as a general minimax theorem are derived from the axioms.

Keywords Entropy · Redundancy · Divergence · Complexity · Two-person zero-sum games · MaxEnt

1 Introduction

Modern information theory with its precisely defined notions to worry about was founded by Shannon in 1948, cf. [21]. The theory led in itself to interesting optimization problems, typically centred around the concept of *capacity*. Relatively soon after, it was realized that information theoretical reasoning also leads to principles of scientific *inference* in other disciplines. Especially we point to Jaynes work [14] in statistical physics (see also [15]), and to Kullback's work [17] in statistics (see also [5]).

Central to Shannons theory is his famous formula for the *entropy* of a discrete distribution P, say of finite support, which, in *natural units*, is expressed in terms of the point probabilities p_i by

$$\mathrm{H}(P) = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}. \tag{1}$$

F. Topsøe (✉)
Institute of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
e-mail: topsoe@math.ku.dk

For Kullback's applications to statistics it was essential to broaden the concept of entropy to a concept involving two distributions. This led to a new quantity, *divergence*, which for discrete distributions is defined by

$$\mathrm{D}(P, Q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}. \tag{2}$$

The relation to entropy becomes clear if one takes Q to be a uniform distribution. For this reason, divergence is also called *relative entropy* or *cross entropy*. The *fundamental inequality* of information theory, also called the *information inequality*, states that $\mathrm{D}(P, Q) \geq 0$ with equality if and only if $P = Q$, cf. [3,24].

Jaynes argues that if $\mathcal{P}$ is a set of probability distributions which models our *knowledge* in a given situation, it is sensible to *infer* that distribution in $\mathcal{P}$ which has maximal entropy. This is Jaynes *maximum entropy principle*, MaxEnt. Kullback suggests that if Q—typically, *not* a distribution in the model $\mathcal{P}$—represents *prior knowledge*, one should infer that distribution $P \in \mathcal{P}$ which minimizes $\mathrm{D}(P, Q)$. This is Kullbacks *minimum information discrimination principle*, also referred to as MinXent, the principle of *minimum cross entropy*.

The good sense of the principles pointed to has since been discussed thoroughly in many works. Here we refer to the more specialized references [8,12,23,25] where a game theoretical approach is considered to lie behind MaxEnt as well as MinXent. A principle of *game theoretical equilibrium* (GTE)—the principle to investigate conditions for equilibrium and to search for optimal strategies for both players in the games considered—is promoted as a key principle in the cited works. The quantities studied in these references are all based on special features of probability distributions.

In [4,7,26,27] it became clear that for key conclusions to hold, one need not stick to habitual concepts of Shannon theory, such as (1) and (2). Still, the modeling was basically probabilistic or at least measure theoretic.

Here, we shall free ourselves from the probabilistic basis and allow an abstract set-up, tied together by four axioms. The first two contain the basic algebraic structure. It is shown that they suffice for the derivation of concavity- and convexity properties, parallel to well-known properties of information theory. The two last axioms are topological and of a more technical nature. They aim at enabling abstract existence results related to optimization.

The results as presented here are intended as an appetizer. The full justification of the abstract approach lies in the richness of applications it facilitates. A comprehensive and systematic study of the possibilities will be taken up subsequently. On the one part, this will connect to problems and results easily recognizable from information theory and, on the other, to different areas within optimization and duality theory.

2 Basic axioms of information triples

We shall take *complexity* (Φ), *entropy* (H) and *redundancy* (D) as key objects to work with and axiomatize useful properties pertaining to triples $\mathcal{I} = (\Phi, \mathrm{H}, \mathrm{D})$, here referred to as *information triples*.

Consider two abstract sets X and Y, conceived as *strategy sets* for, respectively, *Player I*—a female say, representing *"Nature"*—who "holds the truth" and then *Player II* who "seeks the truth" and may be taken to be male, acting as *Observer*. Player I chooses a strategy from X, Player II one from Y.

The terminology is chosen so as to arise inspiring associations and to ease the appreciation of concepts involved, and cannot be taken as an argument in favour of existence of "absolute truth"—if anything, truth rather reflects the state of Observer regarding beliefs, knowledge and other features related to a person with a mind. Regarding the other player, our modeling views Nature as a "person" without a mind.

Besides the two strategy sets, we have also given a map $x \curvearrowright \hat{x}$ of X into Y, the *response*. Usually, the response is given in some natural way and we have decided not to reserve a special letter as a label for it. Intuitively, when you see things from the point of view of Nature, you work with x and when you take the point of view of Observer, you rather work with $\hat{x}$ or some $y \in Y$. We stress that though the response is often injective, it need not be so. If x_1 and x_2 have the same response ($\hat{x_1} = \hat{x_2}$), we say that the two strategies are *equivalent* and write $x_1 \equiv x_2$. When $y = \hat{x}$, we express this by saying that y is *adapted to* x.

More precisely, an *information triple* $\mathcal{I} = (\Phi, \mathrm{H}, \mathrm{D})$ related to the given strategy sets and the given response, is a set of three maps with values in $]-\infty, \infty]$: Φ defined on $X \times Y$, referred to as *complexity*, H defined on X, referred to as *entropy*, and D defined on $X \times Y$, referred to as *redundancy*. Often, especially if the strategy sets are equal and the identity is taken as response, redundancy is also called *divergence*. The value $\Phi(x, y)$ is interpreted as the complexity, seen from the point of view of Observer, when he is using the strategy y and the truth (chosen strategy by Nature) is x.

As guiding principle for the first axiom, we hold that *entropy is minimal complexity*, *redundancy represents actual complexity as related to minimal complexity* and, furthermore, *Observer can always react optimally to any strategy chosen by Nature—if only her choice is known to Observer*.

Axiom 1 (*linking, fundamental inequality*) For any $(x, y) \in X \times Y$,

$$\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y), \tag{3}$$

$$\mathrm{D}(x, y) \geq 0 \ \text{ and} \tag{4}$$

$$\mathrm{D}(x, y) = 0 \Leftrightarrow y = \hat{x}. \tag{5}$$

The identity (3) is the *linking identity* and, taken together with (5), (4) is the *fundamental inequality*. As we see, entropy is indeed minimal complexity: $\mathrm{H}(x) = \min_{y \in Y} \Phi(x, y)$ and the minimum is assumed for $y = \hat{x}$ and, if $\mathrm{H}(x) < \infty$, the minimum is not assumed for any other strategy.

It is technically convenient to allow complexity and entropy to take negative values, but for many natural examples, these quantities are non-negative, just as is redundancy.

We shall take a game theoretical view and are thus led to consider the maximin-value based on complexity, hence Axiom 1 points directly to a general maximum entropy principle.

For the next axiom it is convenient to introduce the set $M_+^1(X)$ of probability distributions over X, typically identified by the set $\alpha = (\alpha_x)_{x \in X}$ of point probabilities. The set $\mathrm{MOL}(X)$ of *molecular measures* is the set of $\alpha \in M_+^1(X)$ for which the *support* $\{x | \alpha_x > 0\}$ is finite.

Axiom 2 (*affinity*) The strategy set X is convex and Φ is affine in its first variable: For $y \in Y$, and $\alpha \in \mathrm{MOL}(X)$,

$$\Phi\left(\sum_{x \in X} \alpha_x x, y\right) = \sum_{x \in X} \alpha_x \Phi(x, y). \tag{6}$$

Sometimes we only need a weaker axiom with equality in (6) replaced by the inequality "$\geq$", corresponding to a condition of concavity.

For the present publication, we only point to two representative examples. Further examples will be presented elsewhere.

Example 1 Let $\mathbb{A}$ be a discrete set, either finite or countably infinite, and put $X = M_+^1(\mathbb{A})$. Further, let $Y = K(\mathbb{A})$, the set of (idealized) *code length functions* over $\mathbb{A}$, i.e. the set of functions $y = (y_i)_{i\in\mathbb{A}}$ for which *Kraft's equality*

$$\sum_{i\in\mathbb{A}} e^{-y_i} = 1 \tag{7}$$

holds. Let the response $y = \hat{x}$ be given by $y_i = \ln \frac{1}{x_i}$ for $i \in \mathbb{A}$. With the definitions

$$\begin{aligned}\Phi(x, y) &= \sum_{i\in\mathbb{A}} x_i y_i,\\ \mathrm{H}(x) &= \sum_{i\in\mathbb{A}} x_i \ln \frac{1}{x_i} \quad \text{and}\\ \mathrm{D}(x, y) &= \sum_{i\in\mathbb{A}} x_i (y_i + \ln x_i),\end{aligned}$$

we obtain a key example related to information theory, possibly presented in a slightly unfamiliar form, e.g. with response different from the identity (compare with (1) and (2)). The validity of Axioms 1 and 2 is readily established and well-known; as for the fundamental inequality, it depends on the *log-sum inequality* cf. e.g. [3,24].

Example 2 Let $X = Y$ be a real Hilbert space and take as response the identity map on X. Let y_0 be a point in Y (the "prior") and define $(\Phi, \mathrm{H}, \mathrm{D})$ by

$$\Phi(x, y) = \|x - y\|^2 - \|x - y_0\|^2, \tag{8}$$

$$\mathrm{H}(x) = -\|x - y_0\|^2, \tag{9}$$

$$\mathrm{D}(x, y) = \|x - y\|^2. \tag{10}$$

The easy verification of Axioms 1 and 2 is left to the reader.

Useful concavity- and convexity results can be derived from the two first axioms. The results are connected with yet another key quantity known from information theory. Consider a convex combination $\overline{x} = \sum_{x\in X} \alpha_x x$ determined by the molecular measure α. Define the associated *information transmission rate* by

$$\mathrm{I}(\alpha) = \sum_{x\in X} \alpha_x \, \mathrm{D}(x, \hat{\overline{x}}). \tag{11}$$

Clearly, $\mathrm{I}(\alpha) = 0$ if and only if all x's with $\alpha_x > 0$ are equivalent. The idea behind the definition is that the response $x \curvearrowright \hat{x}$ represents *communication* from Nature to Observer with x as the *message sent* and $\hat{x}$ as the *message received*. If Nature selects the message to be sent according to some distribution determined by the weights α_x and if Observer finds that $\overline{x}$ best represents what he has to be prepared for, he chooses the strategy $\hat{\overline{x}}$. Nature actually sends an $x \in X$ with weight α_x and this represents a kind of "surprisal" to Observer, measured by $\mathrm{D}(x, \hat{\overline{x}})$. The greater the surprisal, the better can Observer distinguish between the possible messages sent by Nature. Taking the average as in (11) we arrive at the (average) *information* per communicated message, hence this can be interpreted as the *information rate* which Observer obtains from his choice of strategy.

Based only on Axioms 1 and 2 we obtain some important identities:

Theorem 1 (basic identities)

(i) *Let* $\overline{x} = \sum_{x \in X} \alpha_x x$ *be a convex combination of elements in* X *corresponding to* $\alpha \in \mathrm{MOL}(X)$. *Then*

$$\mathrm{H}\left(\sum_{x \in X} \alpha_x x\right) = \sum_{x \in X} \alpha_x \,\mathrm{H}(x) + \mathrm{I}(\alpha). \tag{12}$$

(ii) *With notation as in* (i), *assume that* $\mathrm{H}(\overline{x}) < \infty$ *and let* $y \in Y$. *Then*

$$\sum_{x \in X} \alpha_x \,\mathrm{D}(x, y) = \mathrm{D}\left(\sum_{x \in X} \alpha_x x, y\right) + \mathrm{I}(\alpha). \tag{13}$$

(iii) *For elements* $\alpha_1, \ldots, \alpha_m$ *in* $\mathrm{MOL}(X)$ *with barycentres* $\overline{x_1} \;\ldots, \overline{x_m}$, *and for any mixture* $\alpha = \sum_1^m w_k \alpha_k$ *with a barycentre* $\overline{x}$ *of finite entropy, the following identity holds:*

$$\mathrm{I}\left(\sum_{k=1}^m w_k \alpha_k\right) = \sum_{k=1}^m w_k \,\mathrm{I}(\alpha_k) + \sum_{k=1}^m w_k \,\mathrm{D}(\overline{x_k}, \hat{\overline{x}}). \tag{14}$$

Proof (adapted from [24]). By the linking identity, the right-hand side of (12) may be written as $\sum \alpha_x \Phi(x, \hat{\overline{x}})$ which by affinity equals $\Phi(\overline{x}, \hat{\overline{x}})$, the left-hand side of (12).

Adding $\sum \alpha_x \,\mathrm{D}(x, y)$ to both sides of (12), applying linking and subsequently affinity we conclude that

$$\begin{aligned}\mathrm{H}(\overline{x}) + \sum \alpha_x \,\mathrm{D}(x, y) &= \sum \alpha_x \Phi(x, y) + \mathrm{I}(\alpha)\\ &= \Phi(\overline{x}, y) + \mathrm{I}(\alpha)\\ &= \mathrm{H}(\overline{x}) + \mathrm{D}(\overline{x}, y) + \mathrm{I}(\alpha).\end{aligned}$$

Subtracting $\mathrm{H}(\overline{x})$, (13) follows.

As to (14), this identity can be derived by similar manipulations as above. Or else one may note the great similarity with (12), introduce a new information triple by a process of *randomization* and then derive (14) as a corollary to (12). Details are left to the reader. □

We may conceive the left hand side of (13) as an attempt to calculate $\mathrm{I}(\alpha)$. Then the equation says that this is in error, as it underestimates $\mathrm{I}(\alpha)$ by the "compensation term" $\mathrm{D}(\overline{x}, y)$. For this reason, (13) is referred to as the *compensation identity*. Note also that in case D is finite, the compensation identity amounts to the independence of y of the expression $\sum \alpha_x \,\mathrm{D}(x, y) - \mathrm{D}(\sum \alpha_x x, y)$.

From the theorem we obtain the following concavity-and convexity relations:

$$\mathrm{H}\left(\sum_{x \in X} \alpha_x x\right) \geq \sum_{x \in X} \alpha_k \,\mathrm{H}(x), \tag{15}$$

$$\mathrm{D}\left(\sum_{x \in X} \alpha_x x, y\right) \leq \sum_{x \in X} \alpha_x \,\mathrm{D}(x, y), \tag{16}$$

$$\mathrm{I}\left(\sum_{k=1}^m w_k \alpha_k\right) \geq \sum_{k=1}^m w_k \,\mathrm{I}(\alpha_k). \tag{17}$$

For the discussion of equalities in these inequalities, recall that $\mathrm{I}(\alpha) = 0$ requires that $\hat{x}$ be independent of x for all x with $\alpha_x > 0$.

We note that whereas the identities (12), (13) and (14) require the full strength of Axiom 2, the inequality (15) only requires that Φ is concave in its first variable, hence only requires the weak form of Axiom 2.

3 Two axioms involving topology

For the further study of information triples $(\Phi, \mathrm{H}, \mathrm{D})$ we introduce topological axioms which focus on D, redundancy.

Axiom 3 (*semi-continuity*) The strategy set X is a topological space, $X = (X, \tau)$, for which the algebraic operations on X are continuous. Furthermore, for each $(x_0, y_0) \in X \times Y$, the two maps $x \curvearrowright \mathrm{D}(x, y_0)$ and $x \curvearrowright \mathrm{D}(x_0, \hat{x})$ are τ-lower semi-continuous.

The topology τ, called the *reference topology*, will usually be a "nice" topology, especially Hausdorff so that one can distinguish topologically between the strategies in X (if the response is not injective one cannot use this map to distinguish). We shall only need sequential notions of convergence. With $x_n \to x$ as notation for convergence in τ, the semi-continuity requirements in Axiom 3 thus amount to the requirements

$$\mathrm{D}(x, y_0) \leq \liminf_{n\to\infty} \mathrm{D}(x_n, y_0), \tag{18}$$

$$\mathrm{D}(x_0, \hat{x}) \leq \liminf_{n\to\infty} \mathrm{D}(x_0, \widehat{x_n}) \tag{19}$$

whenever $x_n \to x$ and $(x_0, y_0) \in X \times Y$.

Our last axiom depends on the notion of a D-*Cauchy sequence*, a sequence $(x_n)_{n\geq 1}$ of strategies in X such that

$$\lim_{n,m\to\infty} \mathrm{D}\left(x_n, \widehat{x_{n,m}}\right) = 0 \text{ where } x_{n,m} = \frac{1}{2}x_n + \frac{1}{2}x_m. \tag{20}$$

Axiom 4 (*weak completeness*) Every D-Cauchy sequence has a subsequence which converges in the reference topology.

For instance, Axiom 4 is satisfied if (X, τ) is sequentially compact.

In addition to convergence in the reference topology, we shall also consider an intrinsic notion of convergence, again only paying attention to sequences: A sequence (x_n) in X is said to *converge in redundancy to x* or to *converge in divergence to x* (whatever appears most natural in the situation considered) if $\mathrm{D}(x_n, \hat{x}) \to 0$. For this we use the notation $x_n \twoheadrightarrow x$. In [10] one finds some of the technical intricacies of convergence in divergence and the topology one may associate with it (when the response is injective).

Examples 1 and 2 satisfy Axioms 3 and 4 with the topology of pointwise convergence as reference topology for Example 1 and the norm topology as reference topology for Example 2. The fact that Example 1 satisfies Axiom 4, follows from *Pinsker's inequality* (for which see e.g. [3,24]).

Let us collect some simple results on convergence.

Proposition 1

(i) *Every sequence which converges in redundancy is a* D-*Cauchy sequence, hence contains a subsequence which converges in the reference topology.*
(ii) *For a sequence which converges in redundancy as well as in the reference topology, the two limit strategies are equivalent.*

(iii) *If the response is injective, then convergence in divergence implies convergence in the reference topology.*

Proof To prove (i), assume that $x_n \twoheadrightarrow x$. Then (x_n) must be a D-Cauchy sequence as follows from the compensation identity (13) applied to the convex combination $\frac{1}{2}x_n + \frac{1}{2}x_m$ with $y = \hat{x}$. Then apply Axiom 4.

As to (ii), assume that $x_n \twoheadrightarrow x$ and that $x_n \to x'$. Then $\mathrm{D}(x', \hat{x}) \leq \liminf \mathrm{D}(x_n, \hat{x}) = 0$ by Axiom 3 and Axiom 1 then shows that $x' \equiv x$.

Finally, to prove (iii), assume that the response is injective and consider a sequence which converges in divergence, say $x_n \twoheadrightarrow x$. By (i), for some $x_0 \in X$ and some subsequence (x_{n_k}), $x_{n_k} \to x_0$. By (ii) applied to the subsequence (x_{n_k}), it follows that $x_0 \equiv x$, hence $x_0 = x$ by injectivity of the response. Applying this argument to subsequences of (x_n), we find that every such subsequence contains a further subsequence which converges to x. As the notion of convergence here involved is topological, we conclude that $x_n \to x$. □

Regarding the last property, one may note that for the implication $x_n \twoheadrightarrow x \Rightarrow x_n \to x$ to hold generally, the response must be injective if only the reference topology satisfies some weak separation axiom—Hausdorff or even weaker (consider sequences $x', x', \cdots$ with $x' \equiv x$).

4 Information triples and games

Consider an information triple $\mathcal{I} = (\Phi, \mathrm{H}, \mathrm{D})$ related to the strategy sets X and Y and the response $x \curvearrowright \hat{x}$. We shall study two-person zero-sum games with Φ as objective function, Observer as minimizer and Nature as optimizer, but with the important restriction that Nature has to choose a strategy from a certain non-empty subset X_0 of X. This set is the *preparation* of the game and strategies in X_0 are *consistent strategies*. The game is denoted $\gamma(X_0)$.

The *values* of the game are defined as usual, cf. e.g. [1]. For Nature, the value is

$$\sup_{x \in X_0} \inf_{y \in Y} \Phi(x, y)$$

which we recognize as the *maximum entropy value*, denoted by $\mathrm{H}_{\max}(X_0)$ (or just $\mathrm{H}_{\max}$):

$$\mathrm{H}_{\max}(X_0) = \sup_{x \in X_0} \mathrm{H}(x). \tag{21}$$

As for Observer, the value is denoted $\mathrm{R}_{\min}(X_0)$ (or just $\mathrm{R}_{\min}$):

$$\mathrm{R}_{\min}(X_0) = \inf_{y \in Y} \sup_{x \in X_0} \Phi(x, y) \tag{22}$$

which may be conceived as *minimal risk*. For a specific Observer strategy y, the associated *risk* is given by

$$\mathrm{R}(y) = \sup_{x \in X_0} \Phi(x, y). \tag{23}$$

By the *minimax inequality*, $\mathrm{H}_{\max} \leq \mathrm{R}_{\min}$. The game is in *equilibrium* if $\mathrm{H}_{\max} = \mathrm{R}_{\min} < \infty$.

A strategy x is an *optimal strategy for Nature* in the game $\gamma(X_0)$, or a MaxEnt-*strategy*, if it is consistent and $\mathrm{H}(x) = \mathrm{H}_{\max}$. A sequence (x_n) of consistent strategies is *asymptotically optimal* if $\lim_{n\to\infty} \mathrm{H}(x_n) = \mathrm{H}_{\max}$. A strategy $x \in X$ (note, not necessarily consistent) is an $\mathrm{H}_{\max}$-*attractor* if $x_n \twoheadrightarrow x$ for every asymptotically optimal sequence (x_n). A strategy $y \in Y$ is an *optimal strategy for Observer*, or a $\mathrm{R}_{\min}$-*strategy*, if $\mathrm{R}(y) = \mathrm{R}_{\min}(X_0)$. A pair $(x^*, y^*) \in X_0 \times Y$ is an *optimal pair*, or a (MaxEnt,$\mathrm{R}_{\min}$)-*pair*, if x^* is a MaxEnt-strategy and y^* a $\mathrm{R}_{\min}$-strategy.

The main result can now be formulated:

Theorem 2 *If X_0 is convex and $\mathrm{H}_{\max}(X_0) < \infty$, then Observer has a unique optimal strategy y^* and, regarding Nature, an $\mathrm{H}_{\max}$-attractor x^* exists and $y^* = \widehat{x^*}$. All $\mathrm{H}_{\max}$-attractors are equivalent. Furthermore, the game is in equilibrium and for $x \in X_0$ and $y \in Y$ the following inequalities (stronger than the trivial inequalities $\mathrm{H}(x) \le \mathrm{H}_{\max}(X_0)$ and $\mathrm{R}_{\min}(X_0) \le \mathrm{R}(y)$) hold:*

$$\mathrm{H}(x) + \mathrm{D}(x, y^*) \le \mathrm{H}_{\max}(X_0), \tag{24}$$

$$\mathrm{R}_{\min}(X_0) + \mathrm{D}(x^*, y) \le \mathrm{R}(y). \tag{25}$$

Proof (modeled after [23])

Let (x_n) be an asymptotically optimal sequence. Assume, as we may, that the sequence $(\mathrm{H}(x_n))_{n\ge1}$ converges "fast" to $\mathrm{H}_{\max}$ in the sense that

$$\lim_{n\to\infty} n\,(\mathrm{H}_{\max} - \mathrm{H}(x_n)) = 0. \tag{26}$$

This information will be used later. For now, we put $x_{n,m} = \frac{1}{2}x_n + \frac{1}{2}x_m$ and $y_{n,m} = \widehat{x_{n,m}}$ and use (12) and convexity of X_0 to find that

$$\mathrm{H}_{\max} \ge \mathrm{H}(x_{n,m}) = \frac{1}{2}\mathrm{H}(x_n) + \frac{1}{2}\mathrm{H}(x_m) + \frac{1}{2}\mathrm{D}(x_n, y_{n,m}) + \frac{1}{2}\mathrm{D}(x_m, y_{n,m}).$$

From this we conclude that (x_n) is a D-Cauchy sequence. Then, by Axiom 4, there exists a subsequence $(x_{n_k})_{k\ge1}$ and an element $x^* \in X$ such that $x_{n_k} \to x^*$ as $k \to \infty$.

For the next part of the proof we consider any consistent strategy x and put $\xi_k = (1 - \frac{1}{n_k})x_{n_k} + \frac{1}{n_k}x$. By continuity of the algebraic operations, $\xi_k \to x^*$. Put $\eta_k = \widehat{\xi_k}$. For each k,

$$\mathrm{H}_{\max} \ge \mathrm{H}(\xi_k) \ge \left(1 - \frac{1}{n_k}\right)\mathrm{H}(x_{n_k}) + \frac{1}{n_k}\mathrm{H}(x) + \frac{1}{n_k}\mathrm{D}(x, \eta_k)$$

and it follows that

$$\mathrm{H}(x) + \mathrm{D}(x, \eta_k) \le n_k\left(\mathrm{H}_{\max} - \mathrm{H}(x_{n_k})\right) + \mathrm{H}(x_{n_k}).$$

By (26) and (19) applied to $\xi_k \to x^*$ we conclude that $\mathrm{H}(x) + \mathrm{D}(x, y^*) \le \mathrm{H}_{\max}$, i.e. $\Phi(x, y^*) \le \mathrm{H}_{\max}$. As this holds for every consistent strategy, $\mathrm{R}(y^*) \le \mathrm{H}_{\max}$. By the minimax inequality, the opposite inequality also holds. Thus, y^* is an optimal strategy for Observer and the game is in equilibrium. As we also established (24)—equivalent with $\mathrm{R}(y^*) \le \mathrm{H}_{\max}$—it follows that any asymptotically optimal sequence converges in divergence to x^*. Thus x^* is indeed an $\mathrm{H}_{\max}$-attractor. Clearly, any other $\mathrm{H}_{\max}$-attractor must be equivalent to x^*.

In order to establish (25), consider any $y \in Y$ and exploit again the asymptotically optimal sequence (x_n) which we started out with. Now we use (18), and observe that

$$\begin{aligned}\mathrm{R}(y) &= \sup_{x\in X_0} \Phi(x, y) \ge \liminf_{n\to\infty} \Phi(x_n, y) = \liminf_{n\to\infty} (\mathrm{H}(x_n) + \mathrm{D}(x_n, y)) \\ &\ge \mathrm{H}_{\max} + \mathrm{D}(x^*, y) = \mathrm{R}_{\min} + \mathrm{D}(x^*, y).\end{aligned}$$

This establishes (25) and also implies uniqueness of y^*. Indeed, from (25) we conclude that if $\mathrm{R}(y) = \mathrm{R}_{\min}$ for some $y \in Y$, then $\mathrm{D}(x^*, y) = 0$ and $y = \widehat{x^*} = y^*$ follows from Axiom 1. □

Though all four axioms are in play for this result, we note that Axiom 2 is only needed in the weaker form with concavity instead of affinity.

Corollary 1 *Let* $\mathrm{co}(X_0)$ *be the convex hull of* X_0*. Then, a necessary and sufficient condition that* $\gamma(X_0)$ *is in equilibrium, is that* $\mathrm{H}_{\max}(X_0) < \infty$ *and that maximum entropy is not increased by taking mixtures in the sense that*

$$\mathrm{H}_{\max}(\mathrm{co}(X_0)) = \mathrm{H}_{\max}(X_0). \tag{27}$$

Proof Sufficiency follows by Theorem 2 since, by Axiom 2, $\mathrm{R}_{\min}(\mathrm{co}(X_0)) = \mathrm{R}_{\min}(X_0)$. This equation is also behind the proof of necessity. Indeed, if $\gamma(X_0)$ is in equilibrium, then

$$\mathrm{H}(\mathrm{co}(X_0)) \leq \mathrm{R}_{\min}(\mathrm{co}(X_0)) = \mathrm{R}_{\min}(X_0) = \mathrm{H}_{\max}(X_0)$$

and (27) follows. □

Note that this proof uses the full strength of Axiom 2.

The MaxEnt-strategy need not exist, even in very simple models ([13,23]). It is unique if the response is injective, but otherwise need not be so. Using the reasoning from the proof of Theorem 2 we realize that a MaxEnt-strategy of a convex preparation with $\mathrm{H}_{\max} < \infty$ is also an $\mathrm{H}_{\max}$-attractor.

Results which imply the existence of a (or *the*) MaxEnt-strategy may be obtained from game theory, especially from considerations involving the notion of *Nash equilibrium*. Thus, if, for $(x^*, y^*) \in (X_0, Y)$, the *saddle value inequalities* hold, i.e.

$$\Phi(x, y^*) \leq \Phi(x^*, y^*) \leq \Phi(x^*, y) \quad \text{for all } (x, y) \in X_0 \times Y \tag{28}$$

and if $\Phi(x^*, y^*) < \infty$, then $\gamma(X_0)$ is in equilibrium with (x^*, y^*) as (MaxEnt,$\mathrm{R}_{\min}$)-pair. With our special assumptions, we see that from (28) and finiteness of $\Phi(x^*, y^*)$ it follows that y^* is adapted to x^* (use (28) with $y = \widehat{x^*}$) and then, the right-hand inequality of (28) is automatic. This points to the essential importance of the first half of (28):

$$\Phi(x, y^*) \leq \Phi(x^*, y^*) \quad \text{for all } x \in X_0. \tag{29}$$

The discussion above contains the key steps to the following result:

Theorem 3 *Consider the game* $\gamma(X_0)$*. Let* x^* *be consistent and* y^* *adapted to* x^**. Then a necessary and sufficient condition that* $\gamma(X_0)$ *is in equilibrium with* (x^*, y^*) *as a* (MaxEnt,$\mathrm{R}_{\min}$)*-pair is that* $\mathrm{H}(x^*) < \infty$ *and* (29) *hold.*

Note that (29) is nothing but the inequality (24), except that for Theorem 3 we must assume that x^* is consistent, whereas in Theorem 2 a main point is that we deal with strategies x^* which need not be so.

Theorem 3 has important corollaries which often lead to the determination of MaxEnt (and $\mathrm{R}_{\min}$-) distributions. Here we mention one such result. For a single preparation X_0, we say that $y \in Y$ is *robust* if there exists a finite constant $\rho = \rho(X_0)$, the *level of robustness*, such that $\Phi(x, y) = \rho$ for every consistent strategy. The set of robust strategies is the *exponential family associated with* X_0. We then obtain the following corollary to Theorem 3:

Corollary 2 *Let* X_0 *be a preparation, assume that* x^* *is consistent and* $y^* = \widehat{x^*}$ *robust. Then* $(x^*, \widehat{x^*})$ *is a* (MaxEnt, $\mathrm{R}_{\min}$)*-pair for* $\gamma(X_0)$.

5 Discussion

5.1 Information before probability

In 1983 Kolmogorov stated that "*Information theory must precede probability theory and not be based on it"* The present research may be seen as an attempt to go some way in this direction and as such is in line with ongoing tendencies, cf. Shafer and Vovk [20] and also watch out for Harremoës [11]. For the cited works, just as for the present research, game theoretical considerations occupy a central position.

5.2 The choice of axioms

The axioms have been chosen as a balance between generality and a wish to make them acceptable on intuitive grounds and smooth to work with. The key testing ground for the choice was that a result like Theorem 2 should be easy to state and prove, and yet general enough to open up for applications in diverse directions.

5.3 Notions of convergence or topological notions?

As indicated, a sequential notion of convergence for the reference structure would suffice. This refinement may be needed for some continuous models where notions such as convergence almost everywhere are appropriate. Recently, it was discovered in Harremoës [9], cf. also [10], that for classical information theory convergence in divergence is in fact a topological notion. This is true quite generally for injective responses $x \curvearrowright \hat{x}$ as one can easily verify the conditions for a sequential notion of convergence to be topological. However, the resulting topology is quite intricate (as are related topologies, cf. [10]), but may come to play an important role for the more subtle points of certain optimization problems.

5.4 The compensation identity

Further work on the axioms will reveal the significance of the identity (13). Apparently, it first appeared in [22]. It is also of significance for quantum information theory and is there called *Donalds identity*, cf. [6].

5.5 Complexity, entropy or divergence as the basis?

Complexity is the most dominant notion in our axiomatization. From this, notions of entropy and divergence can be derived. It turns out that the presently popular approach with a focus on entropy which goes back to Bregman, cf. [2], can also be handled based on our axioms. Relevant recent works include [4,18,19,27].

As hinted at very briefly, one can also put emphasis on divergence. Thus all three of complexity, entropy and divergence can be put in the central position. We find that conceptually as well as on theoretical grounds, it is advantageous to emphasize complexity as this quantity has the simplest interpretation. However, one should not consider this quantity alone. In [16] we have an early contribution with focus on complexity, there termed *inaccuracy*, and intended mainly for applications to statistics.

5.6 Further work

There are several possibilities for expanding on the axiomatics. Some are indicated in [12], others concern games in networks, the non-commutative case (the quantum case), introducing geometry, investigating representation theorems etc. Other possibilities concern the discussion of optimization- and duality results from mathematical analysis which did not find a place in this study.

Acknowledgements This work was supported by the Danish Natural Science Research Council. Thanks are due to the organizers of the "International Conference on Nonlinear Analysis and Optimization", Isfahan 2007, especially to professor Jafar Zafarani, for a well structured conference which also gave the author the opportunity to present his ideas to a professional audience of experts in optimization theory. Furthermore, constructive criticism by a conscientious referee led to a tight presentation of essentials to the benefit of readers of the present journal.

References

1. Aubin, J.P.: Optima and Equilibria. An Introduction to Nonlinear Analysis. Springer, Berlin (1993)
2. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**, 200–217 (1967)
3. Cover, T., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
4. Csiszár, I.: Generalized projections for non-negative functions. Acta Math. Hungar. **68**, 161–185 (1995)
5. Csiszár, I., Shields, P.: Information Theory and Statistics: A Tutorial Foundations and Trends in Communications and Information Theory. Now Publishers Inc., USA (2004)
6. Donald, M.J.: On the relative entropy. Commun. Math. Phys. **105**, 13–34 (1985)
7. Glonti, O., Harremoës, P., Khechinashvili, Z., Topsøe, F.: Nash equilibrium in a game of calibration. Teoriya Veroyatnostei i Ee Primeneniya **51**(3), 537–551 (2006)
8. Grünwald, P.D., Dawid, A.P.: Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Ann. Math. Stat. **32**(4), 1367–1433 (2004)
9. Harremoës, P.: The information topology. In: Proceedings IEEE International Symposium on Information Theory, Lausanne, June 2002, p. 431. IEEE
10. Harremoës, P.: Information topologies with applications. In: Csiszár, I., Katona, G.O.H., Tardos, G. (eds.) Entropy, Search, Complexity, volume 16 of Bolyai Society Mathematical Studies, pp. 113–150. János Bolyai Mathematical Society and Springer-Verlag (2007)
11. Harremoës, P.: Probability and information—Occam's razor in action. In preparation (2008)
12. Harremoës, P., Topsøe, F.: Unified approach to optimization techniques in Shannon theory. In: Proceedings of the IEEE International Symposium on Information Theory, Lausanne, June 2002, p. 238. IEEE
13. Ingarden, R.S., Urbanik, K.: Quantum informational thermodynamics. Acta Physica Polonica **21**, 281–304 (1962)
14. Jaynes, E.T.: Information theory and statistical mechanics, I and II. Phys. Rev. **106/108**, 620–630 and 171–190 (1957)
15. Kapur, J.N.: Maximum Entropy Models in Science and Engineering. Wiley, New York (1993), first edition 1989
16. Kerridge, D.F.: Inaccuracy and inference. J. Roy. Stat. Soc. B. **23**, 184–194 (1961)
17. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
18. Naudts, J.: Continuity of a class of entropies and relative entropies. Rev. Math. Phys. **16**, 809–822 (2004)
19. Naudts, J.: Estimators, escort probabilities, and ϕ-exponential families in statistical physics. J. Ineq. Pure and Appl. Math. **5**, 1–15 (2004)
20. Shafer, G., Vovk, V.: Probability and finance. It's only a game! Wiley, Chichester (2001)
21. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 and 623–656 (1948)
22. Topsøe, F.: An information theoretical identity and a problem involving capacity. Studia Scientiarum Mathematicarum Hungarica **2**, 291–292 (1967)
23. Topsøe, F.: Information theoretical optimization techniques. Kybernetika **15**(1), 8–27 (1979)

24. Topsøe, F.: Basic concepts, identities and inequalities—the toolkit of information theory. Entropy **3**, 162–190 (2001)
25. Topsøe, F.: Maximum entropy versus minimum risk and applications to some classical discrete distributions. IEEE Trans. Inform. Theory. **48**(8), 2368–2376 (2002)
26. Topsøe, F.: Entropy and equilibrium via games of complexity. Physica A. **340**(1–3), 11–31 (2004)
27. Topsøe, F.: Exponential families and MaxEnt calculations for entropy measures of statistical physics. In: Hermann, Q., Rapisarda, T., Abe, M. (eds.) Complexity, Metastability, and Non-Extensivity, CTNEXT07, volume 965 of AIP Conference Proceedings, pp. 104–113 (2007)