Game Theoretical Optimization Techniques inspired by Information Theory

Flemming Topsøe [*]
University of Copenhagen, Institute of Mathematical Sciences
Universitetsparken 5, 2100 Copenhagen, Denmark

21.10.07, 8 am

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. Among the two players, one – "Observer" – is imagined to have a "mind", the other – "Nature" – not. Expressing such ideas leads to un-symmetric modeling as the two players are treated quite differently. We demonstrate that the theory can be used as a common framework for diverse applications, which apart from information theory includes problems of geometry.

keywords Entropy, divergence, complexity, two-person zero-sum games, MaxEnt, MinXent. Kuhn-Tucker theorem

[*] This work was supported by the Danish Natural Science Research Council

1 Introduction

Modern information theory with precisely defined notions to worry about was founded by Shannon in 1948, cf. [19]. The theory led in itself to interesting optimization problems, typically centered 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. We point to Kullbacks principle of *minimum information discrimination*, [14], within theoretical statistics, and to Jaynes' principle of *maximum entropy*, [12], designed for statistical physics, but of much wider applicability as witnessed by [13].

A key to Shannons theory is his famous formula

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

here expressed in *natural units*. Another key to what is now known as *Shannon theory* is the concept of a *code*. This notion, in a simplified form, will also be of importance to us. Before entering into that, we note that for Kullback's applications to statistics it was essential to broaden the concept of entropy to a concept, now mainly called *divergence* and, for the case of discrete distributions, 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 important *information inequality* states that $\mathrm{D}(P,Q) \geq 0$ with equality if and only if $P = Q$, cf. [2] or [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. Instead of giving comprehensive references, we refer to the more specialized references [23], [25], [11] and [7]. Common to these is a basic game theoretical approach which is taken 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 of the author. The quantities studied in these references are those given by (1) and (2) and closely related quantities, all based on special features of probability distributions.

In [3], [15], [26] and [21] it became clear that for key conclusions to hold, one need not stick to more habitual concepts of *Shannon theory*, such as (1) and (2). Still, the modeling was basically probabilistic or at least measure theoretic. Work in mathematical finance, cf. [17], is also entering into this picture.

Here, we shall free ourselves from the probabilistic basis and allow a completely abstract set-up, tied together by suitable axioms. To test the good sense in this approach we consider an old problem of Sylvester who wrote "*It is required to find the least circle which shall contain a given system of points in the plane*" – in fact this is the full text of [20]. In addition to demonstrating how Sylvesters problem can be tackled within our general framework, we also show how to derive a version of the *separation theorem* for convex sets in suitable spaces.

The aim of the present research is to develop key elements of an abstract axiomatic theory as indicated above and to demonstrate the soundness of the theory by pointing to interesting applications in diverse fields.

2 Information triples based on complexity

We shall primarily take *complexity* (Φ), *entropy* (H) and *divergence* (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.* Examples from information theory proper as well as from convex analysis and geometry will fit into the theory developed.

Consider two abstract sets X and Y, conceived as *strategy sets* for two "players", *Player I* who "holds the truth" and is also referred to as *Nature*, and then *Player II* who "seeks the truth" and is also referred to as *Observer.*

We imagine that Nature chooses a strategy from X, Observer a strategy from Y. The further game theoretical aspects this points to are taken up in Section 5.

The terminology is chosen only for convenience, 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 modelling 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. Usually, this map, called the *connection*, is given in some natural way and we have decided not to reserve a special symbol as a label for this map. 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 connection is often injective, it need not be so. If $\hat{x_1} = \hat{x_2}$ we say that the two strategies are *equivalent* and write $x_1 \equiv x_2$.

Let us embark on more precise explanations. An *information triple* $\mathcal{I} = (\Phi, \mathrm{H}, \mathrm{D})$ is a set of maps: $\Phi : X \times Y \rightarrow]-\infty, \infty]$, referred to as *complexity*, $\mathrm{H} : X \rightarrow]-\infty, \infty]$, referred to as *entropy*, and $\mathrm{D} : X \times Y \rightarrow [0, \infty]$ referred to as *divergence*. The value $\Phi(x, y)$ is interpreted as the complexity, seen from the point of view of Observer, when he (yes, let Observer be male, Nature female) is using the strategy y and the truth (chosen strategy by Nature) is x. It is technically convenient to allow complexity and entropy to be negative, but for the more natural examples, these quantities are non-negative. The interplay between Φ, H and D and the connection $x \curvearrowright \hat{x}$ is what we set out to axiomatize.

As guiding principle for the first axiom, we hold that *entropy is minimal complexity*, *divergence 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.*

Guided by these principles, we can state the first axiom:

Axiom 1 (linking). For any $(x, y) \in X \times Y$, the *linking identity*

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

holds, as does the biimplication

$$\mathrm{D}(x, y) = 0 \Leftrightarrow y = \hat{x} \,\square \tag{4}$$

So 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. For reasons indicated, we often call $\hat{x}$ the strategy *adapted to x*.

Example 1 (classical information theory). Let $\mathbb{A}$, the *alphabet*, be a discrete set (either finite or countably infinite), put $X = M_+^1(\mathbb{A})$, the set of probability distributions over $\mathbb{A}$, and $Y = K(\mathbb{A})$, the set of *code length functions* over $\mathbb{A}$, i.e. the set of $\kappa : \mathbb{A} \to [0,\infty]$ such that *Krafts equality*

$$\sum_{i \in \mathbb{A}} \exp(-\kappa_i) = 1 \tag{5}$$

holds. Denote by $P \leftrightarrow \kappa$ the bijection between $M_+^1(\mathbb{A})$ and $K(\mathbb{A})$ given by

$$\kappa_i = \ln \frac{1}{p_i} \, ; \, i \in \mathbb{A} \tag{6}$$

and, in the other direction,

$$p_i = \exp(-\kappa_i) \, ; \, i \in \mathbb{A} \, . \tag{7}$$

With Φ defined as *average code length*, i.e.

$$\Phi(P,\kappa) = \langle \kappa, P \rangle = \sum_{i \in \mathbb{A}} p_i \kappa_i \tag{8}$$

(thus $\langle \cdot, P \rangle$ is used for mean values w.r.t. P) and with H and D as the classical quantities given by (1) and (2), Axiom 1 is satisfied (apply the information inequality). Suitable interpretations associated with this example are indicated e.g. in [25]. □

In the last section we shall indicate how this example may be extended to cover many other information triples.

Example 2. Let $X = Y$ be a Hilbert space, e.g. an Euclidean space and define Φ by

$$\Phi(x,y) = \|x - y\|^2 \, . \tag{9}$$

For this complexity function the associated entropy function is trivial: $\mathrm{H}(x) \equiv 0$ and divergence coincides with complexity. With $x \curvearrowright \hat{x}$ as the identity map, Axiom 1 holds. As the reader will realize, we can use the setting here as point of departure for discussion of Sylvester's problem. However, it may appear unnatural with the exponent 2 in (9) and also, it is not very informative to work with an entropy function which is identically 0. In Example 3 we shall modify this example to arrive at a richer structure which simplifies the further analysis – carried out in Example 10 – and also explains the choice of the exponent in (9). □

The three further axioms we shall consider involve extra structure, either related to convexity or to topology.

Axiom 2 (affinity). The strategy set X is a convex set and Φ is affine in its first variable: For every $y \in Y$, and every convex combination of elements in X, say $\overline{x} = \sum_1^n \alpha_\nu x_\nu$,

$$\Phi(\overline{x}, y) = \Phi\Big(\sum \alpha_\nu x_\nu, y\Big) = \sum \alpha_\nu \Phi(x_\nu, y) \,\square \tag{10}$$

In (10) the α_ν's are understood to be non-negative with sum 1. Let us introduce a more streamlined notation by considering the set $\mathrm{MOL}(X)$ of *molecular distributions*, distributions in $M_+^1(\mathbb{A})$ with finite *support*, $\mathrm{supp}(\alpha) = \{x|\alpha_x > 0\}$ (we use α_x rather than $\alpha(x)$ for the weights of α). For such a distribution we denote by $\hat{\alpha}$ the *barycenter* of α:

$$\hat{\alpha} = \sum_{x \in X} \alpha_x \cdot x\,. \tag{11}$$

By considering the natural embedding of X in $\mathrm{MOL}(X)$ we realize that $\alpha \curvearrowright \hat{\alpha}$ is a natural extension of the connection given.

Clearly, Axiom 2 holds for Example 1 but not so for Example 2. This can, however, be remedied, either by introducing a *prior* as discussed later on in Example 7 or by *randomization*:

Example 3 (Example 2 continued). As in Example 2 let Y be some Hilbert space. Consider this time the strategy set $X = \mathrm{MOL}(Y)$ and the connection $\alpha \curvearrowright \hat{\alpha}$ defined by (11) above. This connection is not injective. Two strategies in X are equivalent if they have the same barycenter. Introduce complexity, entropy and divergence as follows:

$$\Phi(\alpha, y) = \sum_{x \in X} \alpha_x \cdot \|x - y\|^2\,, \tag{12}$$

$$\mathrm{H}(\alpha) = \sum_{x \in X} \alpha_x \cdot \|x - \hat{\alpha}\|^2\,, \tag{13}$$

$$\mathrm{D}(\alpha, y) = \|\hat{\alpha} - y\|^2\,. \tag{14}$$

Simple checking shows that this defines an information triple which satisfies Axioms 1 and 2. Whereas complexity is here defined from the complexity measure of Example 2 by a natural process of *randomization*, it is the special properties of the inner product which ensures that there are also natural extensions of the entropy- and divergence functions which ensures that Axiom 1 holds in the new setting. □

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 $\overline{x} = \sum_{x \in X} \alpha_x \cdot x$ determined by the molecular measure α. Define the associated *information rate* by

$$\mathrm{I}(\alpha) = \sum_{x \in X} \alpha_x \, \mathrm{D}(x, \hat{\overline{x}}) \,. \tag{15}$$

Clearly, $\mathrm{I}(\alpha) = 0$ if and only if all x's with $\alpha_x > 0$ are equivalent.

The quantity $\mathrm{I}(\alpha)$ is also referred to as *information transmission rate* as it is associated with the idea that the connection $x \curvearrowright \hat{x}$ could represent *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 (15) 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.

Before continuing with the axiomatics, let us derive the concavity- and convexity results hinted at before:

Theorem 1 (concavity- and convexity properties).

(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*

$$\mathrm{H}\Big(\sum_{x \in X} \alpha_x x\Big) = \sum_{x \in X} \alpha_x \, \mathrm{H}(x) + \mathrm{I}(\alpha) \,. \tag{16}$$

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

$$\sum_{x \in X} \alpha_x \, \mathrm{D}(x, y) = \mathrm{D}(\sum_{x \in X} \alpha_x x, y) + \mathrm{I}(\alpha) \,. \tag{17}$$

(iii) *For elements* $\alpha_1, \cdots, \alpha_m$ *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:*

$$\mathrm{I}\Big(\sum_{k=1}^{m} w_k \alpha_k\Big) = \sum_{k=1}^{m} w_k\, \mathrm{I}(\alpha_k) + \sum_{k=1}^{m} w_k\, \mathrm{D}(\overline{x_k}, \overline{x})\,. \tag{18}$$

Proof. (adapted from [24]).

By the linking identity, the right-hand side of (16) 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 (16).

Adding $\sum \alpha_x\, \mathrm{D}(x, y)$ to both sides of (16), 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 $\mathrm{H}(\overline{x})$, (17) follows.

To establish (18), let $y \in Y$ be arbitrary. As all $\mathrm{H}(\overline{x_k})$ are finite by the assumption $\mathrm{H}(\overline{x}) < \infty$ and by (16) (or rather by (20) below), we find that, for each $k = 1, \cdots, m$,

$$\sum_{x \in X} \alpha_k(x)\, \mathrm{D}(x, y) = \mathrm{D}(\overline{x_k}, y) + \mathrm{I}(\alpha_k)\,.$$

Multiplying with w_k and adding terms, we find that

$$\sum_{x \in X} \alpha(x)\, \mathrm{D}(x, y) = \sum_{k=1}^{m} w_k\, \mathrm{D}(\overline{x_k}, y) + \sum_{k=1}^{m} w_k\, \mathrm{I}(\alpha_k)\,.$$

The left-hand side can be transformed further by (17) and we have proved the following identity

$$\mathrm{D}(\overline{x}, y) + \mathrm{I}(\alpha) = \sum_{k=1}^{m} w_k\, \mathrm{D}(\overline{x_k}, y) + \sum_{k=1}^{m} w_k\, \mathrm{I}(\alpha_k)\,, \tag{19}$$

which is an identity of interest in its own right. We obtain (18) as the special case corresponding to $y = \hat{\overline{x}}$. The reader may want to note that the term $\sum w_k\, \mathrm{I}(\alpha_k)$ is itself a kind of information rate but of a "higher order" as it concerns elements in $\mathrm{MOL}(\mathrm{MOL}(X))$. □

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

$$\mathrm{H}\Big(\sum_{x\in X}\alpha_x x\Big) \geq \sum_{x\in X}\alpha_k\,\mathrm{H}(x)\,, \tag{20}$$

$$\mathrm{D}\Big(\sum_{x\in X}\alpha_x x, y\Big) \leq \sum_{x\in X}\alpha_x\,\mathrm{D}(x,y) \tag{21}$$

$$\mathrm{I}\Big(\sum_{k=1}^{m} w_k\alpha_k\Big) \geq \sum_{k=1}^{m} w_k\,\mathrm{I}(\alpha_k)\,. \tag{22}$$

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

We may conceive the left hand side of (17) 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, (17) is referred to as the *compensation identity.*

Now, let us continue with the axiomatization and introduce topology into the picture.

Axiom 3 (semi-continuity). The strategy set $X = (X, \tau)$ is a topological Hausdorff space and, provided X is assumed to be convex, the algebraic operations are continuous. Further, 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 is called the *reference topology.* We shall only need sequential notions. Thus the essential conditions of lower semi-continuity amount to the requirements

$$\mathrm{D}(x, y_0) \leq \liminf_{n\to\infty} \mathrm{D}(x_n, y_0)\,, \tag{23}$$

$$\mathrm{D}(x_0, \hat{x}) \leq \liminf_{n\to\infty} \mathrm{D}(x_0, \widehat{x_n}) \tag{24}$$

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

The reason why we insist on a Hausdorff topology on X is that we do wish to distinguish the various elements in X and though it may, as we shall se by example, be difficult to distinguish equivalent elements in X, we do at least want that such elements can be distinguished topologically. Of course, if the connection $x \curvearrowright \hat{x}$ is injective, the Hausdorff requirement is quite natural anyhow.

In addition to the sequential notion of convergence induced by the topology τ, we shall also consider an intrinsic notion of convergence, again only for sequences. This notion is called *convergence in divergence* or *information theoretical convergence*. For a sequence (x_n) in X and an element $x \in X$, the notion is denoted $x_n \twoheadrightarrow x$ and defined by

$$x_n \twoheadrightarrow x \Leftrightarrow \mathrm{D}(x_n, \hat{x}) \to 0\,. \tag{25}$$

We note that a given sequence (x_n) either does not converge in divergence or converges in divergence to all x's in a certain equivalence class.

In [10] one finds some of the technical intricacies of convergence in divergence and the topology one may associate with it (when the connection is injective).

The final axiom is also topological and assumes that the previous axioms hold.

Axiom 4 (weak completeness). For a sequence (x_n) in X, put $x_{n,m} = \frac{1}{2}x_n + \frac{1}{2}x_m$ and $y_{n,m} = \widehat{x_{n,m}}$. If the "Cauchy-type property"

$$\lim_{n,m\to\infty} \mathrm{D}(x_n, y_{n,m}) = 0\,, \tag{26}$$

holds, then some subsequence of (x_n) converges in the reference topology: For some $x \in X$ and some subsequence $(x_{n_k})_{k\geq 1}$, $x_{n_k} \to x$. $\square$

Lemma 1. *Assume that Axioms 1-4 hold and that the connection is injective. Then convergence in divergence is at least as strong as convergence in the reference topology:* $x_n \twoheadrightarrow x \Rightarrow x_n \to x$.

Proof. Assume that $x_n \twoheadrightarrow x$. By the compensation identity associated with the convex combination $\frac{1}{2}x_n + \frac{1}{2}x_m$ and the strategy $y = \hat{x}$, one finds that the Cauchy-type property of Axiom 4 holds. Therefore, for some $x_0 \in X$, and some subsequence (x_{n_k}), $x_{n_k} \to x_0$. Then, from lower semi-continuity, we conclude that $\mathrm{D}(x_0, \hat{x}) \leq \liminf \mathrm{D}(x_{n_k}, \hat{x}) = 0$. By Axiom 1, $\hat{x} = \hat{x_0}$, hence $x \equiv x_0$ and as the connection is injective, $x_0 = x$ follows. Applying this argument to any subsequence of (x_n), we find that every subsequence of (x_n) contains a further subsequence which converges to x. As the notion of convergence here involved is topological, we conclude that $x_n \to x$. $\square$

Example 4 (Example 1 continued). The information triple of Example 1 satisfies all four axioms with the topology of pointwise convergence as reference topology. By *Scheffé's lemma*, cf. Lemma 3.1 of [24], the reference topology coincides with the topology of convergence in total variation. Axiom 3 holds since divergence can be written as sums of non-negative continuous functions (write $D(P,\kappa)$ as $\sum(q_i - p_i - p_i \ln \frac{q_i}{p_i})$ with κ adapted to Q). For similar reasons, it also follows that entropy is lower semi-continuous. Axiom 4 holds by *Pinsker's inequality*, cf. [2] or [24]. For this example, convergence in divergence is considerably stronger than convergence in the reference topology. □

The consequences that can be drawn from the four axioms introduced mainly concern game theoretical properties. This will be taken up in Section 5 after we have developed another route to information triples.

3 Information triples based on pay-off

When we deal with systems where concepts involved centres around the notion of "information" in one sense or another, it may well be that "pay-off" is more to the point than "complexity" . Mathematically the change is trivial – it just amounts to a change of sign – however, conceptually the change is important as it leads to situations which are quite different in flavour. We therefore fix special notation enabling us to deal with *information triples based on pay-off* in addition to the information triples in the previous section, which we may refer to as *information triples based on complexity.*

Consider now objects $X, Y, x \curvearrowright \hat{x}, \Psi, \Pi, \mathrm{D}$ which we refer to, respectively, as *strategy set for Nature*, *strategy set for Observer*, *connection* (assumed to map X into Y), *pay-off* (mapping $X \times Y$ into $[-\infty, \infty[$) , *maximal pay-off* (mapping X into $[-\infty, \infty[$) and *divergence* or *discrepancy* (mapping $X \times Y$ into $[0, \infty]$.

Axioms 1 and 2 pertaining to this setting are as follows, briefly expressed:

Axiom* 1. $\Pi(x) = \Psi(x,y) + \mathrm{D}(x,y)$[1] and $\mathrm{D}(x,y) = 0 \Leftrightarrow y = \hat{x}$. □

Axiom* 2. X is convex, Ψ affine in its first variable. □

Axiom* 3 and Axiom* 4 are taken to be verbatim the same as Axiom 3 and Axiom 4.

[1] by convention, this is automatically fulfilled if the right-hand side is indeterminate, of the form $-\infty + \infty$. Alternatively, the equation could be written in the form $-\Psi = -\Pi + \mathrm{D}$ but this is less intuitive.

There is then a complete duality $(\Phi, \mathrm{H}, \mathrm{D}) \leftrightarrow (\Psi, \Pi, \mathrm{D})$ between the two types of information triples we have introduced.

4 Information triples generated by divergence, updating

The starting point for the approaches taken in Sections 2 and 3 have been either complexity or pay-off. Other approaches focus either on entropy or on divergence as the basic object to start with. Here we shall focus on the generation of information triples from divergence. Some examples will be complexity-based, others are most naturally expressed for pay-off based triples.

The geometry-oriented Examples 2 and 3 serve as motivation for the abstract modeling to follow. As point of departure we take two abstract sets, X and Y and a *connection* $x \curvearrowright \hat{x}$ from X to Y and then a function $\mathrm{D} : Y \times Y \to [0, \infty]$, called *divergence* (though it may not come from any natural complexity function in the way described in Section 2).

The axioms we shall now consider are the following:

Axiom 5. *For any pair* (y_1, y_2) *of points in* Y*,* $\mathrm{D}(y_1, y_2) = 0 \Leftrightarrow y_1 = y_2$*.*

Axiom 6. *The set* Y *is a convex set and for every* $\beta \in \mathrm{MOL}(Y)$*, and every* $y_0 \in Y$*,*

$$\sum_{y \in Y} \beta_y \, \mathrm{D}(y, y_0) = \mathrm{D}(\overline{y}, y_0) + \sum_{y \in Y} \beta_y \, \mathrm{D}(y, \overline{y}) \tag{27}$$

with $\overline{y} = \sum_{y \in Y} \beta_y \cdot y$*.*

Axiom 7. *The set* $Y = (Y, \tau)$ *is a topological space for which the algebraic operations are continuous and for each* $y_0 \in Y$ *the mappings* $y \curvearrowright \mathrm{D}(y, y_0)$ *and* $y \curvearrowright \mathrm{D}(y_0, y)$ *are lower semicontinuous.*

Axiom 8. *The set* X *is finite.*

Axioms 5-7 correspond quite closely to Axioms 1-3. As for Axiom 8 this is stronger than what one might have expected considering Axiom 4.

Based on Axioms 5 and 6, we define the *information triple obtained from* D *by randomization over* X. It has $\mathrm{MOL}(X)$ as strategy set for Nature, Y

as strategy set for Observer and the connection $\alpha \curvearrowright \hat{\alpha}$ of $\mathrm{MOL}(X)$ into Y is given by

$$\hat{\alpha} = \sum_{x\in X} \alpha_x \cdot \hat{x}\,. \tag{28}$$

Further, complexity, entropy and divergence are defined by

$$\Phi(\alpha, y) = \sum_{x\in X} \alpha_x \, \mathrm{D}(\hat{x}, y)\,, \tag{29}$$

$$\mathrm{H}(\alpha) = \sum_{x\in X} \alpha_x \, \mathrm{D}(\hat{x}, \hat{\alpha})\,, \tag{30}$$

$$\hat{\mathrm{D}}(\alpha, y) = \mathrm{D}(\hat{\alpha}, y)\,. \tag{31}$$

Clearly, Axioms 1 and 2 are satisfied for this triple.

On $\mathrm{MOL}(X)$ we consider as *reference topology* the topology of pointwise convergence, thus $\alpha_n \to \alpha$ means that $\alpha_n(x) \to \alpha(x)$ for each $x \in X$. Under Axiom 8, this is a compact metrizable topology, hence Axiom 4 automatically holds. If also Axiom 7 is satisfied, the map $\alpha \curvearrowright \hat{\alpha}$ is continuous and the semi-continuity requirements of Axiom 3 hold. Thus Axioms 5-8 for the given divergence function are reflected in the validity of Axioms 1-4 for the generated triple $(\Phi, \mathrm{H}, \hat{\mathrm{D}})$. When we later refer to triples generated in this way, we shall use the same letter, D, for the two types of divergence involved in the definition (31) as no misunderstanding seems likely.

Example 3 is clearly an example which fits the construction above. However, the prototype is in information theory as we shall now indicate:

Example 5. We consider two finite, *alphabets*: $\mathbb{A}$ representing the messages to be sent, and $\mathbb{B}$ representing messages that may be received. We take $X = \mathbb{A}$, $Y = M^1_+(\mathbb{B})$ and for divergence we take ordinary Kullback-Leibler divergence on $Y \times Y$. Further, for each $x \in \mathbb{A}$, we assume that a distribution P_x is given which models what goes on at the receiving end in case x is sent. In other words, P_x is the *conditional distribution* over $\mathbb{B}$ under the condition that x is sent. As connection $X \to Y$ we take $x \curvearrowright P_x$. The model obtained is a *discrete memoryless channel.* We note that $\mathrm{H}(\alpha)$ is the transmission rate as defined by (15) when, as one says, the channel is *driven* by the *source* α. A key object to consider is the *capacity* of the channel defined by

$$C = \sup_{\alpha\in \mathrm{MOL}(\mathbb{A})} \mathrm{H}(\alpha)\,. \tag{32}$$

The related notion of an *optimal source* is defined as a source α such that $\mathrm{H}(\alpha) = \mathrm{C}$. □

We shall return to Examples 3 and 5 in Section 6.

We shall now suggest another axiomatic approach which from a divergence function (which need not be generated from any given complexity- or pay-off function as explained in the two previous sections) leads to pay-off based information triples.

Axiom system for divergence with prior: Given is a convex Hausdorff space X, a set Y, a connection $x \curvearrowright \hat{x}$ between these sets, an element $y_0 \in Y$, the *prior*, and a divergence function $\mathrm{D} : X \times Y \to [0, \infty[$ which satisfies the usual condition $\mathrm{D}(x, y) = 0 \Leftrightarrow y = \hat{x}$, the technical conditions of Axioms 3 and 4 as well as the condition specific to what we have in mind, namely that the function $x \curvearrowright \mathrm{D}(x, y_0) - \mathrm{D}(x, y)$ is affine for evert $y \in Y$. Based on these axioms we define pay-off and maximal pay-off by

$$\Psi(x, y) = \mathrm{D}(x, y_0) - \mathrm{D}(x, y) \tag{33}$$

$$\Pi(x) = \mathrm{D}(x, y_0)\,. \tag{34}$$

It is then a trivial matter to check that $\mathcal{I} = (\Psi, \Pi, \mathrm{D})$ is indeed a pay-off based information triple. This triple is *the pay-off based triple for updating generated by* D *and the prior* y_0.

In suggestive terms we say that $\Psi(x, y)$ is the pay-off of Observer resulting from the *updating strategy* to take y as *posterior* (or *update*). If $\Psi(x, y)$ is negative, it is not a good idea for Observer to update the prior y_0 with y. If Observer sticks to y_0, hence does not change strategy (or *opinion*), his pay-off will be 0. But he may do better. We realize that $\Pi(x)$ is the maximal possible pay-off for Observer as his response if he knows x.

Let us indicate two examples where the construction above is applicable:

Example 6. Let $(\Phi, \mathrm{H}, \mathrm{D})$ be a complexity-based information triple satisfying Axioms 1-4 and assume that Φ is finite. Further, let $y_0 \in Y$. Then we may "forget about Φ and H" and base updating solely on D to obtain the pay-off based triple for updating generated by D and y_0 as explained above. Note that the essential condition of affinity is satisfied since

$$\Psi(x, y) = \mathrm{D}(x, y_0) - \mathrm{D}(x, y) = \Phi(x, y_0) - \Phi(x, y) \tag{35}$$

and the affinity assumption for Φ applies. □

Example 7. Let us return to Example 2 where $X = Y$ is a Hilbert space and squared norm difference plays the key role, cf. (9). We use this as divergence: $\mathrm{D}(x, y) = \|x - y\|^2$. The identity is chosen for the connection and as prior we take any element $y_0 \in Y$. As reference topology we use the norm topology. Clearly, convergence in divergence as enters into Axiom 3, reduces to convergence in this topology. It is easily checked that the requirements in the axiom system for divergence with prior are satisfied and we arrive at an information triple (Ψ, Π, D) with pay-off and maximal pay-off given by

$$\Psi(x, y) = \|x - y_0\|^2 - \|x - y\|^2 \tag{36}$$

$$\Pi(x) = \mathrm{D}(x, y_0)\,. \tag{37}$$

Regarding the requirement of affinity, this follows from the fact that second-order terms in x disappear in (36), and may also be seen from any of the following expressions:

$$\Psi(x, y) = 2\langle y_0 - x, y - y_0\rangle - \|y_0 - y\|^2 \tag{38}$$

$$= \|y_0 - y\|^2 - 2\langle y - x, y - y_0\rangle\,. \tag{39}$$

We return to this example in Example11. □

5 Information triples and games

Consider a complexity-based information triple $\mathcal{I} = (\Phi, \mathrm{H}, \mathrm{D})$ related to the strategy sets X and Y and the connection $x \curvearrowright \hat{x}$. Unless explicitly pointed out to the contrary, we assume that Axioms 1–4 are satisfied. We shall study two-person zero-sum games with Φ as objective function and 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 called the *preparation* of the game and strategies in X_0 are called *consistent strategies*. The game considered 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 $\mathrm{H}_{\max}(X_0)$:

$$\mathrm{H}_{\max}(X_0) = \sup_{x \in X_0} \mathrm{H}(x)\,. \tag{40}$$

For brevity, we often write $\mathrm{H}_{\max}$ in place of $\mathrm{H}_{\max}(X_0)$.

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

$$\mathrm{R}_{\min}(X_0) = \inf_{y \in Y} \sup_{x \in X_0} \Phi(x, y) \tag{41}$$

which may be conceived as the *minimal risk*. It is often written as $\mathrm{R}_{\min}$. For a specific Observer strategy y, the associated *risk* is given by

$$\mathrm{R}(y) = \sup_{x \in X_0} \Phi(x, y)\,. \tag{42}$$

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

Important strategies and sequences of strategies associated with $\gamma(X_0)$ are defined as follows: A strategy x is an *optimal strategy for Nature*, or a MaxEnt-*strategy*, if it is consistent and $\mathrm{H}(x) = \mathrm{H}_{\max}$. A sequence (x_n) of consistent strategies is said to be *asymptotically optimal* if $\lim_{n\to\infty} \mathrm{H}(x_n) = \mathrm{H}_{\max}$. A strategy $x \in X$ (obs, 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 each $x \in X_0$ and each $y \in Y$ the following inequalities (stronger than the trivial $\mathrm{H}(x) \leq \mathrm{H}_{\max}(X_0)$ and $\mathrm{R}_{\min}(X_0) \leq \mathrm{R}(y)$) hold:*

$$\mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}_{\max}(X_0) \tag{43}$$

$$\mathrm{R}_{\min}(X_0) + \mathrm{D}(x^*, y) \leq \mathrm{R}(y)\,. \tag{44}$$

Proof. (modeled after [23])

Let (x_n) be an asymptotically optimal sequence. Assume, as we may, that the sequence $(\mathrm{H}(x_n))_{n\geq 1}$ converges “fast” to $\mathrm{H}_{\max}$ in the sense that

$$\lim_{n\to\infty} n\Big(\mathrm{H}_{\max} - \mathrm{H}(x_n)\Big) = 0\,. \tag{45}$$

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 (16) and convexity of X_0 to find that

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

From this we conclude that (x_n) satisfies the Cauchy-property of Axiom 4. Hence there exists a subsequence $(x_{n_k})_{k\geq 1}$ 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,

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

and it follows that

$$\mathrm{H}(x) + \mathrm{D}(x, \eta_k) \leq n_k\Big(\,\mathrm{H}_{\max} - \mathrm{H}(x_{n_k})\Big) + \mathrm{H}(x_{n_k})\,.$$

By (45) and the lower semi-continuity property (24) applied to $\xi_k \to x^*$ we conclude that $\mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}_{\max}$, i.e. $\Phi(x, y^*) \leq \mathrm{H}_{\max}$. As this holds for any $x \in X_0$, we find that $\mathrm{R}(y^*) \leq \mathrm{H}_{\max}$. 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 (43) – equivalent with $\mathrm{R}(y^*) \leq \mathrm{H}_{\max}$ – 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^*.

As our last task, we establish (44). To this end, consider any $y \in Y$ and exploit again the asymptotically optimal sequence (x_n) which we started out with. Now we use the other semi-continuity property, (23), and observe that

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

This establishes (44) and also implies uniqueness of y^*. Indeed, from (44) we conclude that if $\mathrm{R}_{\min} = \mathrm{R}_{\min}$ for some $y \in Y$, then $\mathrm{D}(x^*, y) = 0$ and $y = \widehat{x^*} = y^*$ follows from Axiom 1. □

Corollary 1. *Denote by* $\mathrm{co}(X_0)$ *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*

$$\mathrm{H}_{\max}(\mathrm{co}(X_0)) = \mathrm{H}_{\max}(X_0)\,. \tag{46}$$

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

$$\mathrm{H}(\mathrm{co}(X_0)) \leq \mathrm{R}_{\min}(\mathrm{co}(X_0)) = \mathrm{R}_{\min}(X_0) = \mathrm{H}_{\max}(X_0) \tag{47}$$

and (47) follows. □

The MaxEnt-strategy need not exist. It is unique if the connection is injective, but otherwise it need not be so. But, using the reasoning from the proof of Theorem 2 we realize that the following holds:

Corollary 2. *A* MaxEnt-*strategy of a convex preparation with* $\mathrm{H}_{\max} < \infty$ *is also an* $\mathrm{H}_{\max}$-*attractor.*

Further results which can be used to show the existence of a (or *the*) MaxEnt-strategy may be obtained from the standard theory of games. Most important is the notion of *Nash equilibrium.* For the game $\gamma(X_0)$, this requires the existence of a pair of strategies $(x^*, y^*) \in (X_0, Y)$ such that the *saddle value inequalities*

$$\Phi(x, y^*) \leq \Phi(x^*, y^*) \leq \Phi(x^*, y) \text{ for all } (x, y) \in X_0 \times Y \tag{48}$$

hold. By standard considerations, $\gamma(X_0)$ is in equilibrium with (x^*, y^*) as (MaxEnt,$\mathrm{R}_{\min}$)-pair, if and only if $\Phi(x^*, y^*) < \infty$ and (48) holds. With our special assumptions, we see that from (48) and finiteness of $\Phi(x^*, y^*)$ it follows that y^* is adapted to x^* (use (48) with $y = \widehat{x^*}$) and then, the right-hand inequality of (48) is automatic. This points to the essential importance of the first half of (48), here called *Nash's inequality*:

$$\Phi(x, y^*) \leq \Phi(x^*, y^*) \text{ for all } x \in X_0\,. \tag{49}$$

By the above discussion we have proved 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 an* (MaxEnt,R$_{\min}$)*-pair is that* $\mathrm{H}(x^*) < \infty$ *and that Nash's inequality holds.*

The reader should note that Nash's inequality is nothing but the inequality (43), 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 two important corollaries which often lead to the determination of MaxEnt (and R$_{\min}$-) distributions. The first one depends on the concept of a *robust Observer strategy.* 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. We put

$$\mathcal{E} = \mathcal{E}(X_0) = \{y | \ y \text{ is robust for } X_0\} \tag{50}$$

and call $\mathcal{E}$ the *exponential family associated with* X_0. A useful extension depends on a *preparation family* which is nothing but a family of preparations, here typically denoted by $\mathcal{X}$. For such a family, we say that $y \in X$ is *robust for the family*, or simply *robust* if the family is understood, if y is robust for every preparation $X_0 \in \mathcal{X}$. The family of all such Observer strategies constitutes the *exponential family associated with* $\mathcal{X}$. Notation and defining formula is given by

$$\mathcal{E} = \mathcal{E}(\mathcal{X}) = \bigcap_{X_0 \in \mathcal{X}} \mathcal{E}(X_0)\,. \tag{51}$$

We can now state the first result hinted at above. It is often applied for a preparation family, however, the essence only involves one preparation:

Corollary 3. *Let X_0 be a preparation, assume that x^* is consistent and that $y^* = \widehat{x^*}$ is robust. Then $(x^*, \widehat{x^*})$ is a* (MaxEnt,R$_{\min}$)*-pair for the game $\gamma(X_0)$.*

We leave the trivial verification to the reader.

The second corollary is more special as it builds on the models constructed in Section 4.

Corollary 4 (Kuhn-Tucker criterion). *Consider an information triple* $(\Phi, \mathrm{H}, \mathrm{D})$ *generated by a divergence which satisfies Axioms 5-8. Consider the game γ related to this triple which has* $\mathrm{MOL}(X)$ *as strategy set for Nature.*

Then, if $\alpha^* \in \mathrm{MOL}(X)$, if $y^* = \widehat{\alpha^*}$ and if, for a finite constant R,

$$\mathrm{D}(\hat{x}, y^*) \leq R \textit{ for all } x \in X \textit{ and} \tag{52}$$

$$\mathrm{D}(\hat{x}, y^*) = R \textit{ for all } x \in X \textit{ with } \alpha_x > 0\,, \tag{53}$$

then (α^*, y^*) is a (MaxEnt,R$_{\min}$)*-pair and R the equilibrium value of γ.*

Proof. For any $\alpha \in \mathrm{MOL}(X)$ we find that

$$\Phi(\alpha, y^*) = \sum_{x \in X} \alpha_x \, \mathrm{D}(\hat{x}, y^*) \leq R = \sum_{x \in X} \alpha^*_x \, \mathrm{D}(\hat{x}, y^*) = \Phi(\alpha^*, y^*)\,,$$

hence Theorem 3 applies and the result follows. □

We mention that the necessity of the conditions related to (52) and (53) can also be proved.

6 Some applications

Example 8 (MaxEnt). Theorems 2 and 3 point directly to a general maximum entropy principle and when specified to the situation in Example 1 is in line with and elaborates on Jaynes classical maximum entropy principle. We refrain from a longer discussion and refer to references already cited.

Example 9 (MinXent). In order to demonstrate the relevance of the theory developed for Kullback's minimum discrimination principle (MinXent), let us return to Example 6. In additions to assumptions made there let us for simplicity assume that the connection $x \curvearrowright \hat{x}$ is injective. Also, let a convex preparation X_0 be given. Consider the two-person zero-sum game γ with pay-off Ψ given by (35).

For $y \in Y$, let

$$\Gamma(y) = \inf_{x \in X_0} \Psi(x, y)\,, \tag{54}$$

the *guaranteed updating gain* associated with the strategy y and let

$$\Gamma_{\max} = \sup_{y \in Y} \Gamma(y) \tag{55}$$

.

Put $\mathrm{D}_{\min} = \inf_{x \in X_0} \mathrm{D}(x, y_0)$. A strategy $x^* \in X$ is the *generalized I-projection of y_0 on X_0* if $x_n \twoheadrightarrow x^*$ for every sequence (x_n) in X_0 which is *asymptotically optimal* in the sense that $\lim_{n\to\infty} \mathrm{D}(x, y_0) = \mathrm{D}_{\min}$. With assumptions made we can now prove:

Theorem 4. *The game γ is in equilibrium, hence $\Gamma_{\max} = \mathrm{D}_{\min}$. There is a unique generalized I-projection x^* of y_0 on X_0 and $y^* = \widehat{x^*}$ is the unique optimal strategy for Observer: $\Gamma(y^*) = \Gamma_{\max}$. Furthermore, for $(x,y) \in X_0 \times Y$,*

$$\mathrm{D}(x, y_0) \geq \mathrm{D}_{\min} + \mathrm{D}(x, y^*)\,, \tag{56}$$

$$\Gamma(y) + \mathrm{D}(x^*, y) \leq \Gamma_{\max}\,. \tag{57}$$

Proof. Consider the triple of maps

$$\begin{aligned} &(x,y) \curvearrowright -\Psi(x,y) = \mathrm{D}(x,y) - \mathrm{D}(x,y_0) \\ &x \curvearrowright \mathrm{H}(x) - \Phi(x,y_0) = -\,\mathrm{D}(x,y_0) \\ &(x,y) \curvearrowright \Phi(x,y) - \mathrm{H}(x) = \mathrm{D}(x,y)\,. \end{aligned}$$

This triple is a complexity-based information triple which, just as $(\Phi, \mathrm{H}, \mathrm{D})$ satisfies Axioms 1-4. Theorem 2 applies and the result follows. □

Example 10 (Sylvester's problem). Let us return to Sylvester's problem and consider a finite subset X_0 of a Euclidean space Y. Points in Y are now referred to as *locations*. As a general reference to *location theory* we mention [5].

Theorem 5. *There exists a unique location y^* for which the maximal distance to points in X_0 is minimal. There also exists $\alpha^* \in \mathrm{MOL}(X_0)$ for which the sum of the weighted squared distances from the barycenter of α^* to the points in X_0 is maximal. Any such molecular distribution has y^* as barycenter. The following relation holds:*

$$\max_{x \in X_0} \|x - y^*\|^2 = \sum_{x \in X_0} \alpha^*_x \|x - y^*\|^2\,, \tag{58}$$

With preparations done in Example 3 and Section 4 an application of Theorem 2 gives the result. Details are left to the reader, except for pointing to the relevance of the relation

$$\begin{aligned} \sup_{\alpha \in \mathrm{MOL}(X_0)} \Phi(\alpha, y) &= \sup_{\alpha \in \mathrm{MOL}(X_0)} \sum_{x \in X_0} \alpha_x \cdot \|x - y\|^2 \\ &= \sup_{x \in X_0} \|x - y\|^2 = \mathrm{R}(y)\,. \end{aligned}$$

We add to the result that the Kuhn-Tucker Theorem of the previous section provides a useful means of identifying the optimal location.

Simple examples illustrate that the molecular distribution α^* may not be unique. Consider for instance as X_0 a set consisting of three points of which one is the midpoint of the two other. Such examples also illustrate that not all molecular distributions in the appropriate equivalence class can be used in the place of $\alpha*$. In other situations – here we can point to the set of 4 corners of a square in the plane – all equivalent α's with the right barycenter can be used. Of course, if $\mathrm{co}(X_0)$ is a simplex, α^* is unique.

The Sylvester problem and the natural problems related to discrete memoryless channels as discussed in Example 5 have much in common, in fact can be treated together. We leave this to the reader to work out.

Example 11 (separation). Let us return to Example 7 and prove a classical result:

Theorem 6. *Assume that X_0 is a closed convex subset of the Hilbert space Y and consider a point $y_0 \notin X_0$. Then there exists a hyperplane which separates y_0 from X_0.*

Proof. Consider the game γ with pay-off given by (36) and X_0 as strategy set for Nature. Conclude from Theorem 2 (applied to the dual game with $-\Psi$ as complexity) that the game is in equilibrium. The value of the game for Nature is $\inf_{x \in X_0} \|x - y_0\|^2$ which is positive by assumption. Therefore, the value of the game for Observer must also be positive, i.e.

$$\sup_{y \in Y} \inf_{x \in X_0} \Big(\|x - y_0\|^2 - \|x - y\|^2 \Big) > 0 \,.$$

We conclude that there exists $y \in Y$ such that

$$\|x - y_0\| > \|x - y\| \text{ for all } x \in X_0 \,.$$

This shows that $y \neq y_0$ and that the hyperplane π of all x with $\|x - y_o\| = \|x - y\|$ separates y_0 from X_0. □

We remark that by translating the hyperplane π found, we can obtain a separating hyperplane through y_0.

Also note that by standard considerations one can use the point separation result above to prove separation for disjoint closed convex sets, one of which is assumed compact.

7 Discussion

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, [18] and also watch out for Harremoës [8]. For the cited works game theory also has a dominant role.

As an indication of the flavour of Kolmogorov's attitude, when discussed from the standpoint of the approach here taken, consider the following example pointed out to me by Harremoës: In case you have accepted the notion of probability distributions but not yet the notion of conditional distributions, you need only observe that if in a standard set-up you consider a (measurable) set A and express information available to you by the preparation of all distributions supported by A, then the I-projection of a distribution on that preparation is nothing but the conditional distribution.

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 for a wide readership. The key testing ground for the choice was that a result like Theorem 2, our main result, should be easy to state and prove, and yet general enough to open up for applications in diverse directions.

It has not been a main aim here to develop new applications, rather we aimed at providing easy access to known results.

It seems that Axiom 1 is natural and central. As to Axiom 2 it turned out that key applications required affinity, hence a natural weakening to concavity has not been taken up. Regarding Axiom **??**, it came as a surprise that apparently one does not need to assume lower semi-continuity of entropy – the requirement of *marginal lower semi-continuity* of divergence appears to be quite sufficient. In this connection one may ask if there are any worth while applications where lower semi-continuety of entropy does *not* hold. Another surprise concerns Axiom 4 where the weak form of completeness is all that is needed, hence a natural strengthening to *strong completeness* (avoiding the passing to a subsequence) was not necessary. As it stands, the weak form of completeness is perhaps more of a weak form of sequential compactness.

Mixtures

Only standard finite mixtures appear throughout the manuscript. More general structures – either countably infinite mixtures or mixtures defined by integration – are known to be of importance in traditional information theory and may be included in further work on the axiomatics.

Notions of convergence or topological notions?

As indicated in the text, a sequential notion of convergence for the reference structure would suffice. However, this refinement is hardly needed. More interesting is it that recently, it was discovered in Harremoës [9], cf. also [10] that for the classical information theoretical situation, cf. Example 1, convergence in divergence is in fact a topological notion. This is true quite generally for injective connections $x \curvearrowright \hat{x}$ as one can easily verify the conditions for a sequential notion of convergence to be topological. However, the resulting topology is quite inticrate (as are related topologies, cf. [10]), but may come to play an important role for the more subtle points of certain optimization problems. In this connection, we mention that essential technical difficulties which do not turn up for the applications we have chosen do turn up, e.g. for MinXent-problems in the continuous case.

The compensation identity

As we have seen this identity (17) plays a significant role. Apparently, it first appeared in [22]. It is also of significance for quantum information theory and is there called *Donalds identity*, cf. [4].

Taking entropy as the basis

For the main approach, complexity is the most dominant notion from which the notions of entropy and divergence can be derived. We showed in Section 4 how focus can be shifted to divergence. In the authors recent work [21] one way of generating information triples based on socalled *Bregman divergence* was discussed. This appears to be of great interest especially for statistical physics perhaps, as witnessed by recent activity in this area. We refer the reader to [21] where one will find expedient methods for MaxEnt- (and MinXent-) calculations. The approach there also has a bearing on more abstract notions of exponential families.

The Gallager-Ryabko theorem

In connection with the discussion of Example 10 we indicated possibilities for a common treatment considering also well known problems of information theory, actually problems indicated in Example 5. For the relevant information theoretical result, known as the *redundancy/ capacity theorem* and due, independently, to Gallager and Ryabko, the reader should consult [6] and [16].

Further work

There are several possibilities for expanding on the axiomatics. Some are indicated in [11], others concern games in networks, covering the non-commutative case (the quantum case), introducing geometry etc. However, rather than working out generalizations, consolidation by pointing to other fruitful applications within the framework here presented may be at least as fruitful. Obvious possibilities concern standard optimization- and duality results from mathematical analysis which we did not find room for here.

References

[1] Aubin, J. P.: *Optima and equilibria. An introduction to nonlinear analysis.* Springer, Berlin (1993)

[2] Cover, T., Thomas, J. A.: *Elements of Information Theory.* Wiley, New York (1991)

[3] Csiszár, I.: Generalized projections for non-negative functions. *Acta Math. Hungar.*, 68, 161–185 (1995)

[4] Donald, M. J.: On the relative entropy. *Commun. Math. Phys.*, 105, 13–34 (1985)

[5] Drezner, Z., Hamacher, H.: *Facility location. Applications and Theory.* Springer, Berlin (2002)

[6] Gallager, R. G.: Source coding with side information and universal coding. Unpublished manuscript.

[7] Grünwald, P. D., Dawid, A. P.: Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Ann. Statist*, 32, 1367–1433 (2004)

[8] Harremoës, P.: Probability and information – Occam's razor in action. In preparation.

[9] Harremoës, P.: The Information Topology. In *Proceedings IEEE International Symposium on Information Theory*, 431, IEEE (2002)

[10] Harremoës, P.: *Information Topologies with Applications.*, In: *Bolyai Society Mathematical Studies, vol. 16*, 113–150 Springer (2007)

[11] Harremoës, P., Topsøe , F.: Unified approach to Optimization Techniques in Shannon Theory. In: *Proceedings, 2002 IEEE International Symposium on Information Theory*, 238, IEEE (2002)

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

[13] Kapur, J. N.: *Maximum Entropy Models in Science and Engineering.* Wiley, New York (1993)

[14] Kullback, S.: *Informaton Theory and Statistics.* Wiley, New York (1959)

[15] Khechinashvili, Z, Glonti, O., Harremoës, P., Topsøe, F.: Nash equilibrium in a game of calibration. *Teoriya Veroyatnostei i ee Primeneniya*, 3, 537–551 (2006)

[16] Ryabko, B. Ya.: Comments on "a source matching approach to finding minimax codes". *IEEE Trans. Inform. Theory*, 27, 780–781 (1981). Including also the ensuing Editor's Note.

[17] Samperi, D. J.: Model Selection Using Entropy and Geometry: Complements to the Six-Author Paper. Available at SSRN: http://ssrn.com/abstract=870072 (2006)

[18] Shafer, G., Vovk V.: *Probability and finance. It's only a game!* Wiley, Chichester (2001)

[19] Shannon, C. E.: A mathematical theory of communication. *Bell Syst. Tech. J.*, 27, 379–423, 623–656 (1948)

[20] Sylvester, J. J.: A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics*, 1, 79 (1857)

[21] Topsøe, F.: Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics. Submitted September 2007.

[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, 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, 2368–2376 (2002)

[26] Topsøe, F.: Entropy and Equilibrium via Games of Complexity. *Physica A*, 340, 11–31 (2004)