Elements of the cognitive universe

Flemming Topsøe
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
Email: topsoe@math.ku.dk

***Abstract*— Based on philosophical considerations, interpretations which connect concepts from the cognitive universe are presented. The resulting abstract theory is downward compatible with basic elements of Shannon Theory. Though philosophically based, the theory is quantitative and besides pragmatic with a multitude of possible applications, some of which are indicated.**

Major attention is payed to the concept of *inference*. Technically, this is based on two-person zero-sum games and notions of *equilibrium* play a central role. This will lead us to abstract versions of the *Pythagorean theorems*, thereby uniting classical geometric results rooted in antiquity and probabilistic results on *information projections* which are well known to the information theory community. Concepts of equilibrium also lead to a notion of *core* which is closely related to that of *exponential families* as known from the the theory of statistical inference.

I. Introduction

We present an abstract quantitative theory addressing cognitive elements such as *truth, belief, knowledge, information, perception* and *inference*. Emphasis is placed on *philosophically based interpretations*. Certain interrelated aspects such as *description, methods of observation* and a notion of *control* are resting somewhat hidden in the background. They are not given concrete form but, somewhat speculatively, postulated as important elements in any concrete and more complete instance of the abstract theory. For classical Shannon Theory, initiated in [15], *coding* provides substance to these concepts.

Much of the extensive research related to cognition is "only philosophical" and not accessible to quantitative analysis. Other parts are hard to apply and mainly of theoretical interest. The aim here is to develop the beginnings of a quantitative theory which is easy to apply. A more complete account than given here will be submitted for publication shortly.

Prompted by the guidelines issued by the ISIT 2011 organizers, let us end the introduction by arguing about the merits of the present contribution.

(i): *"what are the papers main contributions?"*

The main insight gained from the present contribution is that several key elements of Shannon Theory can be developed in an abstract setting. Accordingly, there is no reference to the notion of probability and yet, it is possible to give abstract formulations of several concepts and results from standard probabilistic information theory.

As an important instance of this, we mention *Pythagorean theorems* (inequalities as well as equalities) which contain, on the one hand, classical geometric results rooted in antiquity and, on the other, well known results from Shannon Theory connected with *information projections*. The fact that unifying results like this can be formulated and proved without much difficulty appears to the author as a convincing argument that the abstract theory suggested is "on the right track" and worth pursuing.

(ii): *"why are the contributions of interest?"*

In recent years, the information theory community has shown an interest to "go beyond Shannon" as witnessed by several initiatives (e.g. workshops such as "Facets of Entropy" , Copenhagen, October 2007, and "Information Beyond Shannon " , Venice, December 2008). Many directions are pursued, e.g. related to networking, brain research, bio-informatics, information geometry, statistical physics, quantitative linguistics and certain areas of psychology. If present findings will be of a more lasting value is hard to predict, but the mere possibility of this is a strong reason why the research should be of interest, also for research groups outside information theory proper.

Some of the results developed or indicated are promising in relation to future research, e.g. regarding duality theory and mathematical programming (inference for feasible preparations, see Section IV, appears relevant in this connection). In general, there seems to be a wide basis for future research in a number of directions for which the present research could serve as a source of inspiration.

When successful, abstraction gives double benefit to the researcher: Previous concrete results – here within Shannon Theory – are cast in a new light and given perspective and, secondly, new applications are enabled. With abstraction you focus on the esentials whereas before, it might have been difficult to "see the wood for the trees" .

(iii): *"how does the new contribution relate to prior work?"*

There are many relations and we outline some major ones.

The idea that it could be meaningful to develop an abstract quantitative theory of information without reference to probability originated, as far as the author is aware, in works by Ingarden and Urbanik, cf. [10] where the authors write *"... information seems intuitively a much simpler and more elementary notion than that of probability ... [it] represents a more primary step of knowledge than that of cognition of probability ..."*. We also point to Kolmogorov, who in [13] (but going back at least to 1970 it seems) stated that *"Information*

theory must preceed probability theory and not be based on it". Those who followed these guidelines either based their research on logic or on the the theory of computing. Here we follow a different and more pragmatic approach which is closely related to works of the statisticians associated with the notion of *score functions*, cf. Good [7] and other sources of this concept pointed to in [5].

The role of *description* is emphasized in the present approach. This is done in a way independent of the *minimum description length principle*, but the reader may want to compare with this well known principle, see e.g. Barron, Rissanen and Yu, [1] or Grünwald [8]. The latter reference discusses many issues central to the present research, including the *maximum entropy principle* due to Jaynes, cf. [11].

Game theory (two-person zero-sum games) will be our main technical tool. The relevance of game theory to problems of inference as here studied was first pointed out by Pfaffelhuber [14] and, independently, by the author [16]. The author has, alone or in collaboration with others, continued this research in several papers, see e.g. [9]. Also note the study [8] by Grünwald and Dawid.

The Pythagorean theorem(s) of information theory originated with Čencov and Csiszár, cf. [2] and [3]. The notion of information projection goes back to these sources and have received much attention since then. Suffice it here to mention [4] and the more recent study [6] by Csiszár and Matús. Note that work initiated by Csiszár on information projections (including work in the pipeline, the author understands) uses convex analysis as the main technical tool. It is of interest to compare this with the game theoretical approach which is pursued here. In this connection, Theorem 2 is indicative of the role of convexity assumptions. Also note [18] where, as in [6], a generalized notion of projection, not studied here, is taken into consideration.

Applications of the abstract theory, mainly with a view towards statistical physics, have recently been published by the author, cf. [17] and [19].

II. Philosophy of cognition, a sketch

Nature and *Observer interact* in a *world* whenever Observer is exposed to *situations* from the world. *Nature* does not have a mind and cannot act but is the holder of *"truth"*. *Observer* seeks the truth but is restricted to *belief*. Observer is guided by a creative mind which is exploited to obtain *knowledge* as effortlessly as possible through *experiments* and associated *observations*. Knowledge often comes in the more loose form of *perception* of situations from the world.

"Belief is a tendency to act"[1]. Thus one should be aware of possibilities to transform belief to more action-oriented objects. Such objects we call *controls*. *Description* is the key to control through the design of experiments. An experiment involves a *preparation* which entails a limitation of the *states* – possible *truth instances* – available to Nature. Theoretically possible but unrealistic preparations should be distinguished from *feasible preparations*. Feasible preparations determine the *knowable*, thus provide limitations to what can be known, hence to obtainable *information*.

Description entails an *effort* which depends on the state as well as on Observers belief. This is the key to quantitative considerations. Insight into the knowable also comes from description. Indeed, it appears that the knowable reflects possible beliefs by Observer and limitations on the associated description effort.

To be operational, description effort should satisfy the *perfect match principle*, viz. that effort, given the state, is the least under a *perfect match*, i.e. when belief equals truth. The minimal effort, given the state, is called *entropy*, and the excess effort, taking also belief into consideration, is called *divergence*.

Interaction in any situation takes place as if Nature and Observer are players in a *two-person zero-sum game* with description effort as objective function, Nature as maximizer and Observer as minimizer. Ideally, one should not only aim at *equilibrium* but also at *bi-optimality*, i.e. the identification of optimal strategies which provides Observer not only with insight about *what* can be *infered* but also on *how*.

III. Technical modelling and basic notions

Given are sets X, the *state space*, and $Y \supseteq X$, the *belief reservoir*, as well as a relation $X \otimes Y \subseteq X \times Y$, *domination*. Notation: $y \succ x$ for $(x, y) \in X \otimes Y$. If $y = x$, (x, y) is a *perfect match*. The relation $X \otimes Y$ is assumed to contain every perfect match. Also convenient is the assumption that there exists y which dominates every state x. Often, $X \otimes Y = X \times Y$.

Pairs in $X \otimes Y$ represent *atomic situations*. More complicated situations involve non-empty subsets of X, *preparations*.

Among the two players indicated in Section II, Observer is assumed to have some initial and very limited information, implying that in a situation with preparation $\mathcal{P}$, Observer always chooses a belief instance y which dominates every state in $\mathcal{P}$ (notation: $y \succ \mathcal{P}$).

A non-empty set $Y_{det} \subseteq Y$ determines *certain beliefs*.

Quantitative considerations are enabled through a function $\Phi : X \otimes Y \to]-\infty, \infty]$, the *description*. This function determines the necessary *effort* by Observer in any atomic situation. We assume that $\Phi(x, y) = 0$ if $y \in Y_{det}$ and – the central assumption of our modelling – that Φ satisfies the *perfect match principle*, that $\Phi(x, y) \geq \Phi(x, x)$. More precisely, we assume that there are functions $\mathrm{H} : X \to]-\infty, \infty]$, called *entropy*, and $\mathrm{D} : X \otimes Y \to [0, \infty]$, called *divergence*, such that, for all $(x, y) \in X \otimes Y$, firstly, $\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y)$ (the *linking identity*) and, secondly, D satisfies the *fundamental inequality (of information theory)*, which dictates that $\mathrm{D}(x, y) \geq 0$ with equality if and only if $y = x$. The assumptions made are also expressed by saying that $(\Phi, \mathrm{H}, \mathrm{D})$ is an *(effort-based) information triple*. A triple $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ for which $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$ is an information triple after this definition is a *utility-based information triple* with U as *utility function* (or *pay-off*) and M as *maximal utility* (as before, D is the *divergence*).

[1] a quotation from Good [7].

Sometimes, Φ is considered to determine the *net effort*. Then the *gross effort* is obtained by adding the *overhead*, typically representing the cost of implementing the *method of description*.

Two descriptions which differ only by a positive scalar are *equivalent*. The choice among equivalent descriptions amounts to a choice of *unit*. In probabilistic models, this can be done by setting the overhead to unity.

Often, effort is measured relative to some standard. A particularly important instance of this occurs when Observer has fixed a *prior* $y_0 \in Y$ and wants to *update* his belief by replacing y_0 with a *posterior* y. For example, if Observer obtains the information "$x \in \mathcal{P}$" for some preparation $\mathcal{P}$, he may want to replace the prior by a posterior $y \in \mathcal{P}$. The associated *updating gain* is defined, when not of the form $\infty - \infty$, by

$$\mathrm{U}_{|y_0}(x,y) = \mathrm{D}(x,y_0) - \mathrm{D}(x,y)\,, \tag{1}$$

an expression more likely to be well defined than the otherwise natural expression $\Phi(x,y_0) - \Phi(x,y)$ in which prior effort Φ^{y_0} is compared to posterior effort Φ^{y}. [2]

If $\mathrm{D}^{y_0} < \infty$, then $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ is a utility-based information triple. Note that triples which occur in this way do not require the full description Φ. It suffices to start out with a divergence function D which satisfies the fundamantal inequality in order for the construction to make sense.

Next, let us identify those preparations which can represent realistic information. Given $y \in Y$ and a *level* $h < \infty$, we first define the *level set* $\mathcal{P}^y(h)$ and the *sublevel set* $\mathcal{P}^y(h^\downarrow)$ by

$$\mathcal{P}^y(h) = \{\Phi^y = h\}\,;\ \mathcal{P}^y(h^\downarrow) = \{\Phi^y \leq h\}\,. \tag{2}$$

By a *basic strict*, respectively *basic slack preparation* we understand, respectively a level set and a sublevel set (when non-empty). A *general strict*, respectively *general slack preparation* is a finite non-empty intersection of basic strict, respectively basic slack preparations. The preparations thus identified are those we consider as *feasible*.

If $\mathbf{y}= (y_1, \cdots, y_n)$ are elements of Y and $\mathbf{h}= (h_1, \cdots, h_n)$ are real numbers, we denote by $\mathcal{P}^{\mathbf{y}}(\mathbf{h})$, respectively $\mathcal{P}^{\mathbf{y}}(\mathbf{h}^\downarrow)$, the sets

$$\mathcal{P}^{\mathbf{y}}(\mathbf{h}) = \bigcap_{i\leq n} \mathcal{P}^{y_i}(h_i)\,;\ \mathcal{P}^{\mathbf{y}}(\mathbf{h}^\downarrow) = \bigcap_{i\leq n} \mathcal{P}^{y_i}(h_i^\downarrow)\,. \tag{3}$$

By $\mathbb{P}^{\mathbf{y}}$ we denote the *preparation family* consisting of all preparations of the form $\mathcal{P}^{\mathbf{y}}(\mathbf{h})$ for some choice of $\mathbf{h}$.

Somewhat speculatively we assume that there is a bijective correspondance between Y and a set, W, the *action space*. Elements in W are called *controls*. They point in a more direct way than belief instances to possible actions by Observer. We find that "describtion " and "control" are related concepts, the one often leading to the other. For the models related to Tsallis entropy which are considered in [19], it is evident how controls (there termed "descriptors") are derived from description. For the present study, we have chosen to focus on belief instances, except for just a few remarks on controls.

[2] for a bivariate function $f = f(x,y)$, we denote by f^y the marginal function $x \mapsto f(x,y)$ and by f_x the marginal function $y \mapsto f(x,y)$.

IV. Inference

Consider *partial information* "$x \in \mathcal{P}$". In practice, $\mathcal{P}$ will be a feasible preparation, but we need not assume so for the basic results.

The process of *inference* concerns the identification of "sensible" states in $\mathcal{P}$ – ideally only one such state, the *infered state*. This will be achieved by game theoretical methods involving the previously indicated two-person zero-sum game with Φ as objective function. As it turns out, this will result in "double inference" where also belief instances will be identified. The game considered is denoted $\gamma(\mathcal{P})$ (or $\gamma(\Phi, \mathcal{P})$).

An infered state, say x^*, tells Observer how "close" in some sense he can get to the truth. On the other hand, an infered belief instance y^* is more of an instruction to Observer on how best to act regarding the set-up of experiments. Accordingly, y^* is really best thought of as the corresponding control. In short, double inference gives Observer information both about *what* can be infered about truth and *how*.

The choice of strategy for Observer may be a a real choice, whereas, for Nature, it is more appropriate to have a fictive choice in mind, reflecting Observers thoughts about what the truth could be.

Following standard notions of game theory, the *value* of $\gamma(\mathcal{P})$ for Nature is

$$\sup_{x\in\mathcal{P}} \inf_{y\succ x} \Phi(x,y) = \sup_{x\in\mathcal{P}} \mathrm{H}(x)\,, \tag{4}$$

the MaxEnt-*value*, $\mathrm{H}_{\max}(\mathcal{P})$. Defining *risk* by

$$\mathrm{Ri}(y|\mathcal{P}) = \sup_{x\in\mathcal{P}} \Phi(x,y)\,,$$

the *value* for Observer is the MinRisk-*value* of the game:

$$\mathrm{Ri}_{\min}(\mathcal{P}) = \inf_{y\succ\mathcal{P}} \mathrm{Ri}(y|\mathcal{P})\,. \tag{5}$$

An *optimal strategy for Nature* is a strategy $x^* \in \mathcal{P}$ with $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$. An *optimal strategy for Observer* is a strategy $y^* \succ \mathcal{P}$ with $\mathrm{Ri}(y^*|\mathcal{P}) = \mathrm{Ri}_{\min}(\mathcal{P})$.

The game is in *equilibrium* if $\mathrm{H}_{\max}(\mathcal{P}) = \mathrm{Ri}_{\min}(\mathcal{P}) < \infty$. By $\mathrm{ctr}(\mathcal{P})$, the *centre of* $\mathcal{P}$, we denote the set of $x \in \mathcal{P}$ which dominate $\mathcal{P}$.

Lemma 1: If $\gamma(\mathcal{P})$ is in equilibrium and both players have optimal strategies, then these strategies are unique, coincide and belong to the centre of $\mathcal{P}$.

Proof: Let $x^* \in \mathcal{P}$ be any optimal strategy for Nature and $y^* \succ \mathcal{P}$ any optimal strategy for Observer. By assumption, such strategies exist. Then $\Phi(x^*,y^*) \geq \mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P}) = \mathrm{Ri}_{\min}(\mathcal{P}) = \mathrm{Ri}(y^*|\mathcal{P}) \geq \Phi(x^*,y^*)$, hence $\Phi(x^*,y^*) = \mathrm{H}(x^*)$ and we conclude that $y^* = x^*$ as desired. ∎

For a game in equilibrium with optimal strategies for both players, the common unique strategy secured by Lemma 1 is the *bi-optimal strategy*. In spite of the identity of the optimal strategies in such cases, we often use different notation,

typically with x^* when we focus on optimality for Nature and with y^* when we focus on optimality for Observer.

For results about existence of the bi-optimal strategy, see [18]. Here we focus on *identification* of the bi-optimal strategy.

Theorem 1: [Identification] Let $y^* = x^* \in \mathrm{ctr}(\mathcal{P})$ with $\mathrm{H}(x^*) < \infty$. Then the following conditions are equivalent:
(i) $\gamma(\mathcal{P})$ is in equilibrium and has x^* as bi-optimal strategy;
(ii) For all $x \in \mathcal{P}$, $\Phi(x, y^*) \leq \mathrm{H}(x^*)$;
(iii) $\mathcal{P} \subseteq \mathcal{P}^{y^*}(h^{\downarrow})$ with $h = \mathrm{H}(x^*)$.
When these conditions are satisfied, the *Pythagorean inequality* as well as the *dual Pythagorean inequility* holds, i.e.

$$\forall x \in \mathcal{P} : \mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}_{\max}(\mathcal{P}) \,, \tag{6}$$

$$\forall y \succ \mathcal{P} : \mathrm{Ri}_{\min}(\mathcal{P}) + \mathrm{D}(x^*, y) \leq \mathrm{Ri}(y|\mathcal{P}) \,. \tag{7}$$

Proof: We present an outline. The equivalence of (i) and (ii) follows as one of the saddle-value inequalities of game theory [3] holds by the perfect match principle, and the other is the inequality of (ii). The reformulation of (ii) given in (iii) is evident.

Now assume that (i)-(iii) hold. The Pythagorean inequality (6) is another reformulation of (ii) and the dual Pythagorean inequality (7) holds since, for $y \succ \mathcal{P}$, $\mathrm{Ri}_{\min}(\mathcal{P}) + \mathrm{D}(x^*, y) = \mathrm{H}(x^*) + \mathrm{D}(x^*, y) = \Phi(x^*, y) \leq \mathrm{Ri}(y|\mathcal{P})$. ∎

The role of the sublevel sets is contained in Theorem 1. This is best illuminated by changing the setting, asking which preparations can have a given state as bi-optimal strategy:

Corollary 1: Let x^* be a state with finite entropy $h = \mathrm{H}(x^*)$. Then $\gamma(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy if and only if $\{x^*\} \in \mathcal{P} \subseteq \mathcal{P}^{x^*}(h^{\downarrow})$. In particular, the largest such preparation is the sublevel set $\mathcal{P}^{x^*}(h^{\downarrow})$.

One may think of a bi-optimal strategy as a kind of "uniform" state over $\mathcal{P}$. In case $\mathcal{P} = X$, we are led to the notion of the *uniform state (over X)*. This is a state x^*, necessarily unique, with $h = \mathrm{H}(x^*) < \infty$ and $\mathcal{P}^{x^*}(h^{\downarrow}) = X$.

When, under the condition of equilibrium, we combine the direct and the dual Pythagorean inequality we find that

$$\mathrm{H}(x) + \mathrm{D}(x, x^*) + \mathrm{D}(x^*, y) \leq \mathrm{Ri}(y|\mathcal{P}) \,, \tag{8}$$

whenever $x \in \mathcal{P}$, $y \succ \mathcal{P}$. This is the *combined Pythagorean inequality*. It contains both Pythagorean inequalities. The inequality applied with $y = x$ appears especially attractive (it involves an abstract version of *Jeffrey's divergence*).

The Pythagorean inequalities have implications of a geometric/topological nature. Indeed, for strategies approaching the optimum, respectively $\mathrm{H}(x_n) \to \mathrm{H}_{\max}(\mathcal{P})$ (with x_n's in $\mathcal{P}$) and $\mathrm{Ri}(y_n|\mathcal{P}) \to \mathrm{Ri}_{\min}(\mathcal{P})$ (with the y_n's dominating $\mathcal{P}$), the relevant divergences converge to zero, respectively $\mathrm{D}(x_n, y^*) \to 0$ and $\mathrm{D}(x^*, y_n) \to 0$. This behaviour can be extended to cases when optimal strategies do not exist, cf. [16], [9] and [18].

[3] typically associated with Nash' name but in this setting – two persons and zero sum – going back to von Neumann, see [21] or the account in [12].

Under special circumstances, the Pythagorean inequalities hold more generally than stated in Theorem 1:

Theorem 2: [Inference under convexity] Assume that X is a convex topological space, that the marginal functions Φ^y with $y \in X$ are concave and that the marginal functions D_x with $x \in X$ are lower semi-continuous on X. Let $\mathcal{P}$ be a convex preparation and assume that $y^* = x^* \in \mathrm{ctr}(\mathcal{P})$ has finite entropy. Then, the condition $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$ is not only necessary, but also sufficient for $\gamma(\mathcal{P})$ to be in equilibrium with x^* as bi-optimal strategy, hence also for (6) and (7) to hold.

Proof: Assume that $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$. Then, for any convex combination of states, say $y = \sum \alpha_i x_i$,

$$\begin{aligned}\mathrm{H}(x^*) \geq \mathrm{H}\Big(\sum \alpha_i x_i\Big) = \Phi\Big(\sum \alpha_i x_i, y\Big) \geq \sum \alpha_i \Phi(x_i, y) \\ = \sum \alpha_i \,\mathrm{H}(x_i) + \sum \alpha_i \,\mathrm{D}(x_i, y) \,.\end{aligned}$$

To establish (ii) of Theorem 1, consider $x \in \mathcal{P}$ and apply the inequality above to $y_\varepsilon = (1-\varepsilon)x^* + \varepsilon x$. You find that $\mathrm{H}(x^*) \geq (1-\varepsilon)\,\mathrm{H}(x^*) + \varepsilon\,\mathrm{H}(x) + \varepsilon\,\mathrm{D}(x, y_\varepsilon)$, hence $\mathrm{H}(x) + \mathrm{D}(x, y_\varepsilon) \leq \mathrm{H}(x^*)$. Letting ε tend to 0, $\mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}(x^*)$ follows. As $x \in \mathcal{P}$ was arbitrary, the desired result follows. ∎

One may criticise the result as you cannot apply it to feasible preparations in case the marginals Φ^y are strictly concave, since then the feasible preparations will, typically, not be convex. Rather than reacting negatively towards this observation, we take it as a strong indication that really useful modelling requires that the marginals Φ^y are in fact affine.

Finally, we turn to situations when the Pythagorean inequality holds with equality.

Theorem 3: [Robustness] If $x^* \in \mathcal{P} \subseteq \mathcal{P}^{x^*}(h)$ and $h < \infty$, then $h = \mathrm{H}(x^*)$ and $\gamma(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy. Furthermore, for $x \in \mathcal{P}$,

$$\mathrm{H}(x) + \mathrm{D}(x, y^*) = \mathrm{H}_{\max}(\mathcal{P}) \,. \tag{9}$$

This follows directly from Theorem 1 and the linking identity. The equality (9) is the *Pythagorean equality*, here in a quite abstract version.

The crucial property $\mathcal{P} \subseteq \mathcal{P}^{y^*}(h)$ we refer to as *robustness* and say, provided $h < \infty$, that y^* is *robust* for $\gamma(\mathcal{P})$ with h as *level of robustness*. The set of all robust strategies for $\gamma(\mathcal{P})$ is called the *core* of $\gamma(\mathcal{P})$ and denoted $\mathcal{C}(\mathcal{P})$. For a family $\mathbb{P}$ of preparations the *core* of the family is the intersection of the individual cores:

$$\mathcal{C}(\mathbb{P}) = \bigcap_{\mathcal{P} \in \mathbb{P}} \mathcal{C}(\mathcal{P}) \,. \tag{10}$$

The notion of core is closely related to that of *exponential families* as indicated in [17] (and further discussed in a forthcoming publication). The notion is particularly useful for preparation families of the form $\mathbb{P}^{\mathbf{y}}$. The core for such a family is denoted $\mathcal{C}^{\mathbf{y}}$.

Theorem 4: [Core and inference] Consider a preparation family $\mathbb{P}^{\mathbf{y}}$ with $\mathbf{y} = (y_1, \cdots, y_n)$. Let x^* be a state, put $y^* = x^*$ and assume that $y^* \in \mathcal{C}^{\mathbf{y}}$. Further, put $\mathbf{h} = (h_1, \cdots, h_n)$ with $h_i = \Phi(x^*, y_i)$ for $i = 1, \cdots, n$ and assume that these constants are finite. Then $\gamma\big(\mathcal{P}^{\mathbf{y}}(\mathbf{h})\big)$ is in equilibrium and has x^* as bi-optimal strategy. In particular, x^* is the MaxEnt strategy for $\mathcal{P}^{\mathbf{y}}(\mathbf{h})$.

At times it is advantageous to translate this result to one dealing with controls rather than truth instances.

The notion of robustness has not received much attention in a game theoretical setting. It is implicit in [3] and in [16]. Apparently, the existence of suitable robust strategies is a strong assumption. However, for typical models appearing in applications, the assumption is often fulfilled when optimal strategies exist. Results from[9] point in this direction.

V. Probabilistic and geometric applications

Important discrete probabilistic models may lead either to the theory of Shannon or to that explored in particular by Tsallis, cf. [20]. Very briefly, one takes states and belief instances to be discrete probability distributions over some *alphabet* $\mathbb{A}$: $x = (x_i)_{i\in\mathbb{A}}$, $y = (y_i)_{i\in\mathbb{A}}$, possibly allowing y's to be incomplete ($\sum y_i \leq 1$). Domination $y \succ x$ means that $y_i > 0$ if $x_i > 0$. Certain beliefs are modeled by deterministic distributions. For Shannon theory one takes *Kerridge inaccuracy* $\sum x_i \ln \frac{1}{y_i}$ as description. Controls are then of the form $w = (w_i)_{i\in\mathbb{A}}$ with $w_i = \ln \frac{1}{y_i}$. Standard entropy and divergence emerges. To see how Tsallis-type quantities arise, we ask the reader to consult [19].

For further applications, first consider general abstract problems of updating. Given is a divergence function D, a preparation $\mathcal{P}$ and a prior y_0 for which $\mathrm{D}^{y_0} < \infty$. The game $\gamma(\mathrm{U}_{|y_0}, \mathcal{P})$ is analogous to the previously considered type of games. It has $\mathrm{U}_{|y_0}$ as objective function, Nature as minimizer and Observer as maximizer. A state $x^* \in \mathcal{P}$ with $\mathrm{D}(x^*, y_0) = \inf_{x\in\mathcal{P}} \mathrm{D}(x, y_0)$ is the I-*projection of* y_0 *on* $\mathcal{P}$ if x^* is unique with these properties.

From previous results applied to the game $\gamma(-\mathrm{U}_{|y_0}, \mathcal{P})$ we deduce the following result:

Theorem 5: [I-projection] With assumptions as above, let $x^* \in \mathcal{P}$. Then a necessary and sufficient condition that $\gamma(\mathrm{U}_{|y_0}, \mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy is that, for every $x \in \mathcal{P}$, the Pythagorean inequality holds, i.e.

$$\mathrm{D}(x, y_0) \geq \mathrm{D}(x, x^*) + \mathrm{D}(x^*, y_0)\,. \tag{11}$$

In case X is a convex topological space, $\mathcal{P}$ convex, the y-marginals of $\mathrm{U}_{|y_0}$ affine when $y \in X$ and the x-marginals of D lower semi-continuous on X for $x \in X$, then the above condition is satisfied if only x^* is the I-projection of y_0 on $\mathcal{P}$.

The classical Pythagorean inequality of Shannon theory follows from Theorem 5. For this, one applies Theorem 5 with *Kullback-Leibler divergence* for D.

The classical geometric inequalities associated with Pythagoras name and the connection with standard geometric projections can be obtained, say in an Euclidean space, by considering D given by $\mathrm{D}(x, y) = \|x - y\|^2$.

In the Euclidean setting or even in the abstract setting above one may embark on a closer geometric study of the games $\gamma(\mathrm{U}_{|y_0}, \mathcal{P})$ – also those not in equilibrium. To give an indication, note that the largest *divergence ball* with centre y_0, i.e. set of the form $\{\mathrm{D}^{y_0} < r\}$, which can be placed in the complement of $\mathcal{P}$ determines Natures value in the game. Similarly, if one looks for "large" *halfspaces*, sets of the form $\{\mathrm{D}^{y_0} - \mathrm{D}^{y} < a\}$, which you can place in the complement of $\mathcal{P}$, you are led to visual geometric expressions for Observers value in the game.

References

[1] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, 1998.

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

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

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

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

[6] I. Csiszár and F. Matús. Information projections revisited. *IEEE Trans. Inform. Theory*, 49(6):1474–1490, June 2003.

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

[8] P. D. Grünwald. *the Minimum Description Length principle*. MIT Press, Cambridge, Massachusetts, 2007.

[9] Peter Harremoës and Flemming Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, Sept. 2001.

[10] R. S. Ingarden and K. Urbanik. Information without probability. *Colloq. Math.*, 9:131–150, 1962.

[11] E. T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, 2003.

[12] T. H. Kjeldsen. John von Neumann's Conception of the Minimax Theorem: A Journey Through Different Mathematical Contexts. *Arch. Hist. Exact Sci.*, pages 39–68, 2001.

[13] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983. (from text prepared for the International Congress of Mathematicians, 1970, Nice).

[14] E. Pfaffelhuber. Minimax information gain and minimum discrimination principle. In I. Csiszár and P. Elias, editors, *Topics in Information Theory*, volume 16 of *Colloquia Mathematica Societatis János Bolyai*, pages 493–519. János Bolyai Mathematical Society and North-Holland, 1977.

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

[16] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8 – 27, 1979.

[17] F. Topsøe. Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics. In Qurati Rapisarda Tsallis Abe, Hermann, editor, *Complexity, Metastability, and Non-Extensivity, CTNEXT07*, volume 965 of *AIP Conference Proceedings*, pages 104–113, 2007.

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

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

[20] Constantino Tsallis. *Introduction to Nonextensive Statistical Mechanics*. Springer, Berlin Heidelberg, 2009.

[21] J. von Neumann. Zur Theorie der Gesellschaftsspiele. *Math. Ann.*, 100:295–320, 1928.