COGNITION AND INFERENCE IN AN ABSTRACT SETTING

Flemming Topsøe

University of Copenhagen
Department of Mathematical Sciences
Universitetsparken 5, dk-2100 Copenhagen, Denmark, topsoe@math.ku.dk

ABSTRACT

We continue the development of an abstract, though quantitative, theory of cognition which is rooted in philosophical considerations. Applications include classical Shannon theory and results from geometry. Special attention is payed to *inference* which is treated as the outcome of a *situation of conflict* between *Nature* and *Observer*, “you”.

1. INTRODUCTION

Last year, at WITMSE 2010, the author presented basic elements of an abstract theory of cognition, cf. [1].

Topics which we will focus on this year include those discussed in Section 3.

Emphasis will be on concrete results, especially concerning *identification*. From the point of view of applications, these are the most useful ones and also the technically simplest to establish.

Two key characteristica of the theory presented is that it is not tied to probabilistic notions and that it invokes game theoretical considerations. The desirability of a non-probabilistic approach has been advanced before, cf. [2], [3], [4], [5] and also the recent survey [6]. The relevance of games goes back to [7] and [8], cf. also [9], [10] and, as an example of a utility-based work, [11].

2. ELEMENTS OF COGNITION

In this section we outline parts of the abstract theory under development. Some passages are taken from [12].

2.1. Philosophical background

Observer is placed in a *world* and *interacts* with *Nature* when confronted with *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 form of *perception* of phenomena 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: *“what you can know depends on what you can describe”*.

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* [2].

Interaction between Nature and Observer takes place as if they 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 *inferred* but also on *how*.

2.2. Truth, belief and description

Given are sets X, the *state space*, and $Y \supseteq X$, the *belief reservoir*, as well as a relation $X \otimes Y \subseteq X \times Y$, called *visibility*. A non-empty set $Y_{det} \subseteq Y$ determines *certain beliefs*. We write $y \succ x$ for $(x, y) \in X \otimes Y$ and say, either that “y can see x”, that “x is visible from y” or similar. By $]y[$ we denote the *outlook* from y, the set of x which are visible from y, and by $[x]$ we denote the *watchout* for x, the set of y from which x can be seen. We asume that $x \succ x$ for all x and that there exists y from which all of X is visible, i.e. $]y[= X$.

A *preparation* is a non-empty subset $\mathcal{P}$ of X. A pair $(x, y) \in X \otimes Y$ is an *atomic situation*. The *watchout* for

[1] a quotation from Good [13].

[2] the term “divergence” appears justified as the quantity typically stands for the discrepancy, counted non-negative, between the “actual” and the “best possible” performance. Regarding “entropy”, terminology is less convincing and some other terminology for the abstract setting may be preferable.

a preparation $\mathcal{P}$ is the set $[\mathcal{P}] = \bigcap_{x \in \mathcal{P}} [x]$, i.e. the set of y from which all of $\mathcal{P}$ is visible. By assumption this set is non-empty. We may write $y \succ \mathcal{P}$ in place of $y \in [\mathcal{P}]$. For many applications, $X \otimes Y = X \times Y$.

Quantitative considerations are enabled through a function $\Phi : X \otimes Y \to]-\infty, \infty]$, the *description*, also called the *effort function*. 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* or is *proper*, essentially 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,

$$\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y)\,, \tag{1}$$

the *linking identity*, and, secondly, the *fundamental inequality* holds for D, i.e. $\mathrm{D}(x, y) \geq 0$ 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* and M as *maximal utility* (as before, D is the *divergence*).

Two descriptions which differ only by a positive scalar are *equivalent*. The choice among equivalent descriptions amounts to a choice of *unit*.

With a proper description Φ, we define a (strict) *feasible preparation* as one of the form $\{\Phi^y = h\}$ or a finite intersection of such sets. Here, Φ^y denotes the marginal function $x \to \Phi(x, y)$ defined on $]y[$. This definition is sound on philosophical grounds. Further, it goes well with a definition of *core*, really an abstract notion of *exponential families*: For a family $\mathbb{P}$ of preparations (typically feasible ones), $\mathrm{core}(\mathbb{P})$ is the set of y such that, for each $\mathcal{P} \in \mathbb{P}$, there exists h such that $\Phi^y = h$ on $\mathcal{P}$. See [14].

The choice of a proper effort function in concrete cases of interest is essential for the theory to render useful results. As examples of appropriate choices, we refer to [1], where cases of probabilistic modelling which lead to *Tsallis entropy* are discussed.

2.3. Inference

Consider *partial information* "$x \in \mathcal{P}$".

The standard process of *inference* concerns the identification of a state in $\mathcal{P}$, the *inferred* state. This will be achieved by game theoretical methods involving the previously indicated game, $\gamma = \gamma(\mathcal{P}|\Phi)$, with Φ as objective function. For γ, also belief instances will be identified. An inferred belief instance y^* is, via the associated control, more of an instruction to Observer on how best to act regarding the set-up of experiments. Double inference gives Observer information both about *what* can be inferred about truth and *how*.

The *value* of $\gamma(\mathcal{P})$ for Nature is

$$\sup_{x \in \mathcal{P}} \inf_{y \succ x} \Phi(x, y) = \sup_{x \in \mathcal{P}} \mathrm{H}(x)\,, \tag{2}$$

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:

$$\mathrm{Ri}_{\min}(\mathcal{P}) = \inf_{y \succ \mathcal{P}} \mathrm{Ri}(y|\mathcal{P})\,. \tag{3}$$

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 $\mathcal{P} \cap [\mathcal{P}]$.

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 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.

Theorem 1 *Let $y^* = x^* \in \mathrm{ctr}(\mathcal{P})$ with $\mathrm{H}(x^*) < \infty$. Then $\gamma(\mathcal{P})$ is in equilibrium and has x^* as bi-optimal strategy if and only if, for all $x \in \mathcal{P}$, $\Phi(x, y^*) \leq \mathrm{H}(x^*)$. When this condition is satisfied, the Pythagorean inequality as well as the dual Pythagorean inequility holds, i.e.*

$$\forall x \in \mathcal{P} : \mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}(x^*)\,, \tag{4}$$

$$\forall y \succ \mathcal{P} : \mathrm{Ri}_{\min}(\mathcal{P}) + \mathrm{D}(x^*, y) \leq \mathrm{Ri}(y|\mathcal{P})\,. \tag{5}$$

Proof In brief: In view of the assumptions imposed, the condition stated is one of the famous *saddle-value inequalities* often ascribed to Nash (but in the present simple case due to von Neumann), and the other saddle-value inequality is automatically fulfilled due to the perfect match principle. The result follows from these observations.

The Pythagorean inequality is a simple reformulation of the inequality $\Phi(x, y^*) \leq \mathrm{H}(x^*)$ and the dual Pythagorean inequality 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 results above are developed for an effort-based information triple. Similar, or rather dual results apply to utility-based information triples. Then Nature is a minimizer, Observer a maximizer. We leave it to the reader to formulate appropriate concepts and results.

3. SPECIAL FEATURES

3.1. Adding a geometric flavour

Let us look specifically at models of *updating*. For this, D is a *divergence function* on $X \otimes Y$, i.e. it satisfies the fundamental inequality, y_0 is a suitable *prior* and $\mathcal{P}$ a preparation such that $\mathrm{D}^{y_0} < \infty$ on $\mathcal{P}$. We consider the utility-based information triple $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ with $\mathrm{U}_{|y_0}(x,y) = \mathrm{D}(x,y_0) - \mathrm{D}(x,y)$, representing *updating gain*. The associated game is denoted $\gamma = \gamma(\mathcal{P}|\,\mathrm{U}_{|y_0})$. An optimal strategy x^* for Nature, if unique, is the D*-projection of* y_0 *on* $\mathcal{P}$, i.e. the unique element in $\mathcal{P}$ such that $\mathrm{D}(x^*, y_0) = \mathrm{D}^{y_0}_{\min}(\mathcal{P})$, the infimum of $\mathrm{D}(x,y_0)$ with $x \in \mathcal{P}$. Given $y \succ \mathcal{P}$, the *guaranteed updating gain* for Observer associated with the *posterior* y and the *maximum guaranteed updating gain* are given by

$$\mathrm{Gtu}(y|\mathcal{P}, y_0) = \inf_{x\in\mathcal{P}} \mathrm{U}_{|y_0}(x,y) \tag{6}$$

$$\mathrm{Gtu}_{\max}(\mathcal{P}, y_0) = \sup_{y\succ\mathcal{P}} \mathrm{Gtu}(y|\mathcal{P}, y_0)\,. \tag{7}$$

Before introducing geometry-like elements, note the following result which follows directly from Theorem 1:

Theorem 2 *A necessaary and sufficient condition that* γ *is in equilibrium with* $x^* \in \mathrm{ctr}(\mathcal{P})$ *as bi-optimal strategy is that the Pythagorean inequality holds which, in this case means that, for* $x \in \mathcal{P}$,

$$\mathrm{D}(x,y_0) \geq \mathrm{D}(x,x^*) + \mathrm{D}(x^*,y_0)\,. \tag{8}$$

If so, x^* *is the* D*-projection of* y_0 *on* $\mathcal{P}$.

Next, consider the *open divergence ball with centre* y_0 *and radius* r, defined as the set

$$\mathrm{B}(y_0, r) = \{\mathrm{D}^{y_0} < r\}\,. \tag{9}$$

Also consider *open half-spaces of size* a,

$$\sigma^+(y, a|y_0) = \{x|\,\mathrm{U}_{|y_0} < a\}\,, \tag{10}$$

and, in particular, the open half-space

$$\sigma^+(y|y_0) = \{x|\,\mathrm{U}_{|y_0} < \mathrm{D}(y,y_0)\}\,. \tag{11}$$

We say that a set is *external to* $\mathcal{P}$ if it is contained in the complement of $\mathcal{P}$. The following result characterizes the values for the players in γ in geometrically flavoured terms, also in cases where γ is not in equilibrium:

Proposition 1 *(i) The value* $\mathrm{D}^{y_0}_{\min}(\mathcal{P})$ *is the size of the largest ball* $\mathrm{B}(y_0, r)$ *which is external to* $\mathcal{P}$*, and the maximal guaranteed updating gain* $\mathrm{Gtu}_{\max}(\mathcal{P}, y_0)$ *is the supremum of a for which there exists* $y \succ \mathcal{P}$ *such that the half-space* $\sigma^+(y, a|y_0)$ *is external to* $\mathcal{P}$.

(ii) The updating game $\gamma(\mathcal{P}|\,\mathrm{U}_{|y_0})$ *is in equilibrium and has a bi-optimal strategy if and only if, for some* $y \in \mathcal{P}$, $\sigma^+(y|y_0)$ *is external to* $\mathcal{P}$ *. When this condition holds, y is the bi-optimal strategy, in particular, y is the* D*-projection of* y_0 *on* $\mathcal{P}$.

If you consider the case where divergence is squared Euclidean distance, the geometric significance of this result becomes clear.

3.2. Adding convexity

For this subsection, X is a convex topological space, the marginals Φ^y are affine and the marginals $\mathrm{D}_x : y \to \mathrm{D}(x,y)$ are lower semi-continuous on X.

Then, for every convex combination $\overline{x} = \sum \alpha_i x_i$,

$$\mathrm{H}(\overline{x}) = \sum \alpha_i \,\mathrm{H}(x_i) + \sum \alpha_i \,\mathrm{D}(x_i, \overline{x})\,. \tag{12}$$

In particular, H is strictly concave on X.

Further, if $\mathrm{H}(\overline{x}) < \infty$, then, for every $y \in Y$, the *compensation identity* holds:

$$\sum \alpha_i \,\mathrm{D}(x_i, y) = \mathrm{D}(\overline{x}, y) + \sum \alpha_i \,\mathrm{D}(x_i, \overline{x})\,. \tag{13}$$

In particular, for $y \in Y$, the restriction of D^y to convex preparations $\mathcal{P}$ with $\mathrm{H}_{\max}(\mathcal{P}) < \infty$ is strictly convex.

Let us look at a game $\gamma(\mathcal{P})$. From Theorem 1 we realize the importance of the condition

$$\forall x \in \mathcal{P} : \Phi(x, y^*) \leq \mathrm{H}(x^*)\,. \tag{14}$$

with $y^* = x^* \in \mathcal{P}$. It leads to equilibrium of $\gamma(\mathcal{P})$ and bi-optimality of x^*. In particular, it implies that $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$. Under the extra assumptions imposed, (14) actually follows from the formally weaker condition $\mathrm{H}(x^*) = \mathrm{H}_{\max}$ as we shall now see:

Theorem 3 *If* $\mathcal{P}$ *is convex and* $x^* \in \mathrm{ctr}(\mathcal{P})$ *has finite entropy, then the condition* $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$ *is not only necessary, but also sufficient for* (14) *to hold, hence for* $\gamma(\mathcal{P})$ *to be in equilibrium with* x^* *as bi-optimal strategy.*

Proof Consider an element $x \in \mathcal{P}$ and apply (12) to a convex combination of the form $y_n = (1 - \frac{1}{n})x^* + \frac{1}{n}x$. We find that $\mathrm{H}(x^*) \geq \mathrm{H}(y_n) \geq (1-\frac{1}{n})\,\mathrm{H}(x^*) + \frac{1}{n}\,\mathrm{H}(x) + \frac{1}{n}\,\mathrm{D}(x, y_n)$ from which we conclude that $\mathrm{H}(x) + \mathrm{D}(x, y_n) \leq \mathrm{H}(x^*)$. Exploiting the assumed lower semi-continuity, $\mathrm{H}(x) + \mathrm{D}(x, x^*) \leq \mathrm{H}(x^*)$ follows. As $x \in \mathcal{P}$ was arbitrary, (14) holds. Then apply Theorem 1. □

We find it important that Theorem 3 also applies to the updating models of Theorem 2. Analyzing this it appears that this is indeed the case, provided you assume that the divergence function which Theorem 2 depends on satisfies the compensation identity. In this way one derives abstract versions of by now classical results of Shannon theory related to *information projections* and Pythagorean inequalities. These results go back to Čencov and Csiszár, cf. [15] and [16]. Also of relevance are [17] and [18]

3.3. Axiomatization

The key object which appears to be worth while axiomatizing is the informastion triples. Basic conditions are centred around the linking identity, the fundamental identity, convexity of X and affinity of the marginals Φ^y. This may be supplied with topological conditions. Details may be found in [19]. One may start from *atomic triples* for which X and Y are the reals or the non-negative reals. A proces of integration leads to more complicated triples, often related to *Bregman divergencies*. Other processes involve *relativization* and *randomization*. A systematic study as indicated also helps in defining concrete triples of interest.

4. REFERENCES

[1] Topsøe, F. Cognition beyond Shannon. Proceedings of the 2010 Workshop on Information Theoretic Methods in Science and Engineering, Tampere, 2010.

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

[3] Kolmogorov, A.N. Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory* **1968**, *14*, 662–664.

[4] de Fériet, K. La theorie génerélisée de l'information et la mesure subjective de l'information. In *Théories de l'information (Colloq. Iiformation et Questionnaires, Marseille-Luminy, 1973*; Springer: Berlin, 1974; pp. 1–35.

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

[6] Rathmanner, S.; Hutter, M. A Philosophical Treatise of Universal Induction. *Entropy* **2011**, *13*, 1076–1136.

[7] Pfaffelhuber, E. Minimax Information Gain and Minimum Discrimination Principle. Topics in Information Theory; Csiszár, I.; Elias, P., Eds. János Bolyai Mathematical Society and North-Holland, 1977, Vol. 16, *Colloquia Mathematica Societatis János Bolyai*, pp. 493–519.

[8] Topsøe, F. Information Theoretical Optimization Techniques. *Kybernetika* **1979**, *15*, 8 – 27.

[9] Harremoës, P.; Topsøe, F. Maximum Entropy Fundamentals. *Entropy* **2001**, *3*, 191–226.

[10] Grünwald, P.D.; Dawid, A.P. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Annals of Mathematical Statistics* **2004**, *32*, 1367–1433.

[11] C. Friedman, J.H.; Sandow, S. A Utility-Based Approach to Some Information Measures. *Entropy* **2007**, *9(1)*, 1–26.

[12] Topsøe, F. Paradigms of Cognition. manuscript in preparation, 2011.

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

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

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

[16] Csiszár, I. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.* **1975**, *3*, 146–158.

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

[18] Csiszár, I.; Matús, F. Information projections revisited. *IEEE Trans. Inform. Theory* **2003**, *49*, 1474–1490.

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