COGNITION BEYOND SHANNON

Flemming Topsøe

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

ABSTRACT

We focus on an abstract philosophy related to the process of cognition, reflecting, what you may call "post-Shannon thoughts". Details on concrete modelling in a probabilistic setting are also given.

1. INTRODUCTION

Originally, the driving force behind the research was to overcome difficulties in extending interpretations associated with classical information theory as developed by Shannon [1] and followers, to the theory promoted by Tsallis for statistical physics and thermodynamics, cf. [2], [3]. In this connection, note that many do not recognize Tsallis' theory as physically sound, cf. [4] and [5].

Our findings have resulted in an extension of the scope by developing highly abstract elements of a "philosophy of cognition". Related endeavours include Ingarden and Urbanik [6] who wrote *"... 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 ..."*. Also note that Kolmogorov around 1970, published in [7], stated that *"Information theory must preceed probability theory and not be based on it"*. Game theoretical considerations appear important for some of the "post-Shannon" work, see e.g. [8].

Formally, our treatment is independent of previous research. However, unconsciously it depends no doubt on tradition developed over time, especially regarding the interface between statistics and information theory, see for instance works on "score functions", as in Good [9] and Fischer [10] and works by Rissanen, Barron and others related to the "minimum description length principle", see [11] and [12].

2. PHILOSOPHICAL CONSIDERATIONS

In this section we introduce a number of concepts related to the general process of cognition.

2.1. Truth, Belief and Knowledge

Consider two "players", *Nature* and *Observer*. Nature has no mind and is the holder of *truth*. Observer seeks the truth but is relegated to *belief*. However, Observer possesses a conscious and creative mind which can be exploited through the design of observation strategies to obtain *knowledge* as effortlessly as possible.

The interplay between Nature and Observer takes place in a certain *world*, $\mathcal{W}$, understood to be one out of many possible worlds. *Situations* from $\mathcal{W}$ are characterized by a *truth instance* (or a *state*), $x \in X$, a *belief instance*, $y \in Y$, and a *knowledge instance*, $z \in Z$. Assume, for simplicity, that $X = Y \subseteq Z$. We think of *knowledge* as the *synthesis of extensive experience*[1].

Knowledge may also be conceived as *the way the situation is perceived by Observer*. We assume that knowledge depends on the situation through the truth- and belief instances – and not on other factors. Formally, we operate with a function $\Pi : X \times Y \to Z$, and interpret any value $z = \Pi(x, y)$ as the knowledge instance in a situation with x as truth- and y as belief instance. Π is the *interactor*. Two worlds with the same interactor are identified and we denote by $\mathcal{W}_\Pi$ the world with interactor Π.

2.2. Examples of Worlds

All worlds considered will be *sound* in the sense that $\Pi(x, y) = x$ provided *belief matches truth*, i.e. provided $y = x$.

The *classical world*, $\mathcal{W}_1$, is characterized by the interactor Π_1 given by $\Pi_1(x, y) = x$. This is the *world of observable truth*. As another extreme, consider $\mathcal{W}_0$, a *black hole*, characterized by the interactor Π_0 given by $\Pi_0(x, y) = y$. In this world, no matter what Observer does, he will only see a mirror image of himself: *what you see is what you believe!* By contrast, in a classical world, *what you see is what is true!*

Mixtures of the two worlds make sense when Z is embedded in a linear space. Then, to each $q \in \mathbf{R}$, we may consider the world $\mathcal{W}_q$ with interactor Π_q given by $\Pi_q(x, y) = qx + (1 - q)y$. These worlds turn out to be closely related to Tsallis' theory.

2.3. Description

The further development depends on the introduction of concepts related to *description*. We maintain that "description requires effort". A *description* or, as we shall also say, an *effort function*, is a function $(x, y) \to \Phi(x, y)$ (often assuming only non-negative values) which to any situation assigns the needed effort by Observer on the way to knowledge. Introducing a notion of *certain beliefs*, we

[1] For concrete probabilistic modelling, the considerations correspond to an ideal situation where, so to speak, "the laws of large numbers have taken over".

"calibrate" description so that $\Phi(x,y) = 0$ in any situation of certain belief. Descriptions which only differ from eachother by a positive scalar factor are considered *equivalent*. Choosing the scalar factor amounts to fixing a *unit of effort* which, as we shall see, is the same as a *unit of information*.

Description provides the *net effort*. Sometimes it is appropriate to add an *overhead* c, say the cost of observation or sampling. This gives you the *gross description*, $\tilde{\Phi}$.

The description Φ is *proper* if it satisfies the *perfect match principle*, i.e. if, for every $(x,y) \in X \times Y$

$$\Phi(x,y) \geq \Phi(x,x)$$

with equality only under a perfect match ($y = x$), or if $\Phi(x,x) = \infty$.

We hold the thesis that *given the world, a proper description is unique modulo equivalence*. Assuming that a choice of unit has been made, we talk about the *ideal description* in case there exists a unique proper description.

2.4. Elements of Information

Assume now that Φ is the ideal description for the world $\mathcal{W}$. Elements of *information* can then be introduced into $\mathcal{W}$. Information always concerns the truth instance. In the case of *full information*, the actual truth instance x is revealed to Observer. We use "x" to denote this piece of information. Perhaps, Observer only speculates about the potential possibility of acquering information or we may imagine the involvement of a third party, an *informer*[2]. As in Shannon theory, specific semantic considerations are not of concern to us. Instead, we focus on semantic-independent measures of information and adopt the view that, quantitatively, *information is saved effort*. Therefore, $\Phi(x,y)$ is the value to Observer of the information "x" in a situation with belief instance y. Information is thus measured by an effort and we may claim that *information is physical!* By *entropy* we understand the *guaranteed saving in effort*. Accordingly, the *entropy of* x, understood as the entropy associated with the information "x", is the minimum over y of $\Phi(x,y)$. Denoting entropy by H, we therefore find, by appeal to the perfect match principle, that

$$\mathrm{H}(x) = \Phi(x,x)\,.$$

We measure the "mismatch", or *divergence* (D), as we shall say, between x and y by the difference between actual and minimal effort, i.e.

$$\mathrm{D}(x,y) = \Phi(x,y) - \mathrm{H}(x)\,,$$

here neglecting complications with the indeterminate form $\infty - \infty$. Rewriting the definition as $\Phi(x,y) = \mathrm{D}(x,y) + \mathrm{H}(x)$, we are faced with the *linking identity* which may be taken as the key to an axiomatic treatment of basic elements of information, cf. [13].

[2]For example, at the airport, you may speculate about the departure time of your flight when you hear the announcement that "the flight to Tampere departs at 4 p.m."

Divergence is non-negative and only vanishes on the diagonal, $y = x$. These facts constitute the *fundamental inequality of information theory*, here understood in a general abstract sense.

2.5. Preparations, Exponential Families

The further abstract development involves, in particular, a study of *partial information*, information of the form "$x \in \mathcal{P}$" with $\mathcal{P}$, the *preparation*, a non-empty subset of X. Game theoretical considerations with description as objective function come into play and leads, via the notion of *Nash-equilibrium*, to abstract versions of well known results involving *Pythagorean inequalities*. We refer to [14], [15], [16], [17] and [18].

But one can go further than this and ask the essential question "What *can* Observer know?". The key to an acceptable answer lies in the view, quoted from [9], that "*Belief is a tendency to act*". This tempts us to transform belief instances to other objects, termed *controls*. They help you to limit the set of all preparations to a much smaller set of *feasible preparations* – and, at the same time leads rather naturally to the consideration of an abstract version of *exponential families*.

3. PROBABILISTIC MODELLING

We leave the very abstract setting and turn to more concrete probabilistic modelling involving discrete distributions. For this, we fix a discrete *alphabet*, $\mathbf{A}$, and take as X and Y the set of probability distributions over $\mathbf{A}$. A world $\mathcal{W} = \mathcal{W}_\pi$ is then identified via a *local interactor*, π, which induces the (global) interactor according to the formula $\Pi(x,y) = \big(\pi(x_i,y_i)\big)_{i\in\mathbf{A}}$. Descriptions are then defined via the choice of a *descriptor* $\kappa : [0,1] \to [0,\infty]$ by the formula

$$\Phi(x,y) = \sum_{i\in\mathbf{A}} \pi(x_i,y_i)\kappa(y_i)\,.$$

We insist that $\kappa(1) = 0$ and that the condition of *normalization*, $\kappa'(1) = -1$ holds.

3.1. Key results

Let us point to the key results in the probabilistic setting: Firstly, given the local interactor, at most one description of the type indicated can be proper. Secondly, for the checking of properness of a candidate description, one need only check the *pointwise fundamental inequality* which is the inequality $\delta(s,t) \geq 0$ with equality only for $t = s$, where δ, the *divergence generator* is defined by

$$\delta(s,t) = \big(\pi(s,t)\kappa(t) + t\big) - \big(s\kappa(s) + s\big)\,.$$

Finally, the third result we wish to point out concerns the mixtures $\mathcal{W}_q$ which are defined by the local interactors π_q given by $\pi_q(s,t) = qs + (1-q)t$. When $q < 0$, no descriptor defines a proper description, whereas when $q > 0$ the ideal description exists and is obtained by choosing

$\kappa_q(t) = \ln_q \frac{1}{t}$, where the *q-logarithm* is defined by

$$\ln_q t = \begin{cases} \ln t \text{ if } q = 1 \\ \frac{1}{1-q}\left(t^{1-q} - 1\right) \text{ otherwise}\,, \end{cases}$$

cf. [19]. The case $q = 0$, a black hole, is a singular case for which the appropriate descriptor is $\kappa(t) = \frac{1}{t} - 1$, leading to a divergence function which vanishes identically.

The resulting entropy in case $q > 0$ is now mainly known as *Tsallis entropy*. Important original papers on this form of entropy are [20], [21], [22], [23] and [2]. In the recent monograph [3] the reader finds applications to statistical physics and thermodynamics as well as a comprehensive list of publications, including also more mathematically directed research.

3.2. The divergence generator

Let us discuss the significance of the divergence generator for a world $\mathcal{W}_\pi$ based on a general local interactor π.

If we accept the conjecture that the pointwise fundamental inequality is necessary for the resulting description to be proper, we easily conclude, as claimed, that only one descriptor κ is possible. Indeed, fixing $s \in]0, 1]$ and expressing that the divergence generator $\delta(s, t)$ must have a stationary point at $t = s$, you are led to the differential equation

$$\frac{\partial \pi}{\partial t}(s, s)\kappa(s) + s\kappa'(s) + 1 = 0\,,$$

which, with $\kappa(1) = 0$, leads to a unique function κ. Whether or not the function identified leads to a proper description has to be tested on a case by case basis.

The considerations also indicate that different worlds may lead to the same descriptor. This is for instance the case related to the mixtures $\mathcal{W}_q$ with a positive q. Then the same descriptor, $\kappa_q = \kappa_q^A$ ("A" for "arithmetic"), gives a proper description, and this holds, also if π_q is replaced by the *geometric average*, $\pi_q^G(s, t) = s^q t^{1-q}$.

Considering the form of the divergence generator, you realize that it is natural to add as overhead cost to Φ as well as to H the constant $c = 1$. This gives you *gross description* $\tilde{\Phi}$ and *gross entropy* $\tilde{\mathrm{H}}$. The interpretation this points to is that the unit of information can be identified as a kind of entrance fee or as an effort associated with a single observation. Thus, *information should be measured against the cost of observation* or, for statistics, *... against the cost of sampling*.

Yet another feature of the divergence generator is that it suggests a generalization to the continuous case of divergence which preserves desirable properties. As we know, this is not so obvious if we turn to entropy or description itself.

3.3. Feasible preparations, exponential families

We shall now indicate how feasible preparations and exponential families come into play for our probabilistic models, only having models with proper descriptions in mind. Fact is that the transformation we need to turn a belief instance into a control are given by the descriptor. Thus, a control w^* is a function on $\mathbf{A}$, $w^* = (w_i^*)_{i\in\mathbf{A}}$ for which there exists y^* with $w_i^* = \kappa(y_i^*)$ for $i \in \mathbf{A}$. This points to a general version of *Kraft's (in)equality* which now reads $\sum_{i\in\mathbf{A}} \rho(w_i^*) = 1$ with ρ the inverse function of κ.

The, admitted somewhat speculative, idea is that Observer can set-up experiments with a preparation which fixes the level of $\Phi(x, y^*)$. Thus, a typical feasible preparation is of the form $\mathcal{P} = \{x|\Phi(x, y^*) = h\}$ and any feasible preparation is a non-empty finite intersection of such sets. The associated *exponential families* enter the picture as the typical belief instances (or control instances if you transform to these objects) which lead to Nash-equilibria for the games loosely indicated before. These are instances y' for which $\Phi(x, y')$ is constant, independent of x as long as x is an instance in a feasible preparation of the type considered. Further details will follow in a forthcoming publication, expanding on [24].

For the worlds $\mathcal{W}_q$, the linear character of $\Phi(x, y)$ in x will play an important role. This relates to the presently popular *Bregman divergencies*, see [25] for the original paper.

4. CONCLUSION

Paradigms of cognition have been presented which go beyond Shannon, regarding level of abstraction and also regarding some of the items dealt with. The approach is philosophical and, though originally intended to provide transparent interpretations of key elements of the theory developed by Tsallis and his followers, it is also illuminating if you restrict attention to classical Shannon theory.

5. REFERENCES

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

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

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

[4] D.H.E. Gross, "Comment on: "nonextensivity: from low-dimensional maps to hamiltonian systems" by tsallis et al.," arXiv:0210448[cond-mat.stat-mech], 2002.

[5] C. R. Shalizi, "Tsallis Statistics, Statistical Mechanics for Non-extensive Systems and Long-Range Interactions," Tech. Rep., 2007, Informal notes from the authors homepage.

[6] R. S. Ingarden and K. Urbanik, "Information without probability," *Colloq. Math.*, vol. 9, pp. 131–150, 1962.

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

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

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

[10] P. Fischer, "On the Inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$," *Metrika*, pp. 199–208, 1972.

[11] A. R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in codinig and modeling," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998, Commemorative issue.

[12] Jorma Rissanen, *Information and Complexity in Statistical Modeling*, Springer, New York, 2007.

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

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

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

[16] F. Topsøe, "Information theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. 8 – 27, 1979.

[17] Peter Harremoës and Flemming Topsøe, "Maximum entropy fundamentals," *Entropy*, vol. 3, no. 3, pp. 191–226, Sept. 2001.

[18] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory," *Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1367–1433, 2004.

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

[20] J. Havrda and F. Charvát, "Quantification method of classification processes. Concept of structural$a$-entropy," *Kybernetika*, vol. 3, pp. 30–35, 1967, Review by I. Csiszár in MR, vol. 34, no.8875.

[21] Z. Daróczy, "Generalized Information Functions," *Information and Control*, vol. 16, pp. 36–51, 1970.

[22] J. Lindhard and V. Nielsen, "Studies in Statistical Dynamics," *Mat. Fys. Medd. Dan. Vid. Selsk.*, vol. 38, no. 9, pp. 1–42, 1971.

[23] J. Lindhard, "On the Theory of Measurement and its Consequences in Statistical Dynamics," *Mat. Fys. Medd. Dan. Vid. Selsk.*, vol. 39, no. 1, pp. 1–39, 1974.

[24] F. Topsøe, "Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics," in *Complexity, Metastability, and Nonextensivity, CTNEXT07*, Qurati Rapisarda Tsallis Abe, Hermann, Ed., 2007, vol. 965 of *AIP Conference Proceedings*, pp. 104–113.

[25] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967, Translated from Russian.