A Quantitative Theory of Cognition with Applications

– slide presentation, intended as an "appetizer" in relation to the full manuscript

Flemming Topsøe, topsoe@math.ku.dk

Department of Mathematical Sciences, University of Copenhagen

My ambition:

- to present a quantitative theory of cognition, involving elements such as TRUTH, BELIEF, KNOWLEDGE, CONTROL, . . . which is abstract (e.g. non-probabilistic), inspired by INFORMATION THEORY and GAME THEORY, building largely on NATURAL INTERPRETATIONS (well, and some speculative considerations!) and with

WIDE APPLICATIONS of interest to the "convexity community", to information theorists, to statisticians etcetera!

The world (Ω), Nature and Observer, situations

NATURE: holder of truth! X state space with elements x, truth instances or states. A preparation is a subset $\mathcal{P} \subseteq X$.

OBSERVER: is concerned about truth but restricted to: belief, action and control! Action and control will here be identified and derived from belief: *"belief is a tendency to act"* (Good 1952). To model these thoughts, introduce:

- Y belief reservoir. $Y \supseteq X$. Elements y are belief instances.

- $\hat{Y}$ action space or control space. You may think of $w \in \hat{Y}$ either as an action or as a control.

- A map, $y \mapsto \hat{y}$, response maps Y into $\hat{Y}$.

Atomic situations: Either certain pairs (x, y) ("Y-domain") or certain pairs (x, w) ("$\hat{Y}$-domain"). Notation: $y \succ x$ (x is visible from y) or $w \succ x$ (x is controlled by w)$\cdots$

Interaction, knowledge, effort

Truth and belief interact and lead to knowledge: $z = \Pi(x, y)$ or, in the $\hat{Y}$-domain, $z = \hat{\Pi}(x, w)$. Note: $\Pi(x, y) = \hat{\Pi}(x, \hat{y})$. Knowledge instances z belong to the knowledge base Z. Interpretation: z represents the way situations from the world are presented to Observer or how situations are perceived by Observer. Π or $\hat{\Pi}$ is the interactor. It characterizes the world: $\Omega = \Omega_{\Pi}$.

Examples: If $Z \supseteq Y \supseteq X$, consider the classical world Ω_1 (fits in with Shannon theory...) with interactor $\Pi_1(x, y) = x$ or a black hole Ω_0 with interactor $\Pi_0(x, y) = y$ (or mixtures if Z is an affine space, fits in with Tsallis theory...).

Perception requires effort!

An effort functionmaps atomic situations, (x, y) or (x, w), into $]-\infty, \infty]$. Convenient to allow negative values as it enables an easy switch from effort- to utility-based concepts by a change of sign. Precise definitions $\cdots$

$\hat{\Phi}$ effort function: $\forall w \succ x : \hat{\Phi}(x, w) \geq \hat{\Phi}(x, \hat{x})$.
$\hat{\Phi}$ proper: "=" only if $w = \hat{x}$ (w adapted to x) or rhs$=\infty$.

Φ effort function: $\forall y \succ x : \Phi(x, y) \geq \Phi(x, x)$.
Φ proper: "=" only if $y = x$ (perfect match) or rhs$=\infty$.

If response is injective, $\Phi(x, y) = \hat{\Phi}(x, \hat{y})$.

Choice among scalarly equivalent effort functions amounts to choice of unit. In a world $\Omega = \Omega_\Pi$ there may, modulo equivalence, only be *one* choice of a proper effort function. This applies to Shannon and to Tsallis theory.

Similar definitions for utility-based concepts: $\hat{U}$ (U) is ... iff $\hat{\Phi} = -U$ ($\Phi = -U$) is so.

Entropy, redundancy, divergence, inf. triples

$\hat{\Phi}$ proper eff.fct., Φ the derived eff.fct. View $\hat{\Phi}(x,w)$ /$\Phi(x,y)$ as information content of "x" in situation (x,w) /(x,y).

Entropy is minimal effort, given x (or guaranteed information or necessary allocation of effort): $\mathrm{H}(x)=\hat{\Phi}(x,\hat{x})=\Phi(x,x)$.

Redundancy is redundant effort: $\hat{\mathrm{D}}(x,w)=\hat{\Phi}(x,w)-\mathrm{H}(x)$.
Rewritten: $\hat{\Phi}(x,w)=\mathrm{H}(x)+\hat{\mathrm{D}}(x,w)$ (linking identity).
Further: $\hat{\mathrm{D}}(x,w)\geq 0$, "$=$" iff $w=\hat{x}$ (fundamental inequality). But: difficulty with $\mathrm{H}=\infty$! Therefore define:

Information triple ($\hat{Y}$-domain): a triple $(\hat{\Phi},\mathrm{H},\hat{\mathrm{D}})$ s.t. linking identity and fundamental inequality hold. For the Y-domain, $(\Phi,\mathrm{H},\mathrm{D})$, we require linking ($\Phi=\mathrm{H}+\mathrm{D}$) and fundamental inequality ($\mathrm{D}(x,y)\geq 0$, "$=$" iff $y=x$). D is divergence.

Utility-based inf. trpl.: $(\hat{\mathrm{U}},\mathrm{M},\hat{\mathrm{D}})$ s.t. $(-\hat{\mathrm{U}},-\mathrm{M},\hat{\mathrm{D}})$ is effort-based trpl. Similarly, $(\mathrm{U},\mathrm{M},\mathrm{D})$ in Y-domain. M: max utility.

Relativization, Updating

Given information triple $(\Phi, \mathrm{H}, \mathrm{D})$. Consider prior y_0 chosen by Observer who seeks an update with posterior y. Updating gain defined by

$\mathrm{U}_{|y_0}(x, y) = \Phi(x, y_0) - \Phi(x, y) = \mathrm{D}(x, y_0) - \mathrm{D}(x, y)$. (Latter expression preferable!).

Assume that the marginal D^{y_0} is finite on some preparation $\mathcal{P}$. Then $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ is a utility-based inf. trpl. on $\mathcal{P} \otimes Y = \{(x, y) | y \succ x, x \in \mathcal{P}\}$. Note: Φ not needed; construction makes sense based only on a general divergence function D.

Control determines what *can* be known!

Feasible preparations are determined from $\hat{\Phi}$ (or from Φ): A feasible preparation is a finite intersection of primitive preparations, and these fall in two types, either strict or slack. Notation and definitions for the primitive preparations are:

$$\mathcal{P}^w(h) = \{\hat{\Phi}^w = h\} = \{x|\hat{\Phi}(x,w) = h\}\,;$$
$$\mathcal{P}^w(h^{\downarrow}) = \{\hat{\Phi}^w \leq h\} = \{x|\hat{\Phi}(x,w) \leq h\}\,.$$

The number h is the level, respectively maximum level of the preparation in question (assuming $\mathcal{P}^w(h) \neq \emptyset$).

Games associated w. $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ or $(\Phi, \mathrm{H}, \mathrm{D})$

$\mathcal{P}$ a preparation. Game $\hat{\gamma}(\mathcal{P}) = \hat{\gamma}(\mathcal{P}|\hat{\Phi})$: $\hat{\Phi}$ objective function, Nature maximizer chooses $x \in \mathcal{P}$, Observer minimizer chooses $w \succ \mathcal{P}$. Values for Nature, resp. for Observer are:

$$\sup_{x\in\mathcal{P}} \inf_{w\succ x} \hat{\Phi}(x,w) = \sup_{x\in\mathcal{P}} \mathrm{H}(x) = \mathrm{H}_{\max}(\mathcal{P})$$
$$\inf_{w\succ\mathcal{P}} \sup_{x\in\mathcal{P}} \hat{\Phi}(x,w) = \inf_{w\succ\mathcal{P}} \hat{\mathrm{Ri}}(w|\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P}).$$

Ri stands for risk. Optimal strategies: for Nature, $x^* \in \mathcal{P}$ s.t. $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$; for Observer, $w \succ \mathcal{P}$ s.t. $\hat{\mathrm{Ri}}(w|\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$.

If "=" holds in minimax inequality $\mathrm{H}_{\max}(\mathcal{P}) \leq \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$ and common value is finite, the game is in equilibrium.

Similar notions apply to the game $\gamma(\mathcal{P})$ for the Y-region.

Basic results for $\hat{\gamma}$ and γ

[Basics] If one of the games $\hat{\gamma}(\mathcal{P})$ and $\gamma(\mathcal{P})$ is in equilibrium and has optimal strategies for both players, so does the other - and if so, optimal strategies are unique and "agree", i.e. if they are (x^*, w^*) and (x^{**}, y^*), then $x^{**} = y^* = x^*$ and $w^* = \hat{x^*}$. [x^* is the bi-optimal strategy. It satisfies: $x^* \succ \mathcal{P}$, $x^* \in \mathcal{P}$, notationally, $x^* \in \mathrm{ctr}(\mathcal{P})$, the centre of $\mathcal{P}$.]

[Identification] With $x^* \in \mathcal{P}$ and $w^* \succ \mathcal{P}$, $\hat{\gamma}(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy if and only if the Nash inequalities hold. If $x^* \in \mathrm{ctr}(\mathcal{P})$ and $w^* = \hat{x^*}$ is already known, it is enough to check one of these: $\forall x \in \mathcal{P} : \hat{\Phi}(x, w^*) \leq \mathrm{H}(x^*)$.

… and more, key results

[Properties] When conditions hold, the direct as well as the dual Pythagorean inequalities hold:
$\forall x \in \mathcal{P} : \mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) \leq \mathrm{H}(x^*)\,,$
$\forall w \succ \mathcal{P} : \hat{\mathrm{Ri}}(w^*|\mathcal{P}) + \hat{\mathrm{D}}(x^*, w) \leq \hat{\mathrm{Ri}}(w|\mathcal{P})\,.$
In particular, x^* is the MaxEnt-attractor, i.e. $x_n \overset{\mathrm{D}}{\to} x^*$ ($\mathrm{D}(x_n, x^*) \to 0$) for any (x_n) in $\mathcal{P}$ with $\mathrm{H}(x_n) \to \mathrm{H}_{\max}(\mathcal{P})$.

[Robustness, core] Let (x^*, w^*) be strategies for $\hat{\gamma}(\mathcal{P})$ with $w^* = \hat{x^*}$. If w^* is robust at the level of robustness h, i.e. if $\hat{\Phi}(x, w^*) = h$ for all $x \in \mathcal{P}$ and h is finite, then $\hat{\gamma}(\mathcal{P})$ is in equilibrium with h as value and with x^* as the bi-optimal strategy. Further, the Pythagorean equality holds:

$$\forall x \in \mathcal{P} :\ \mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) = \mathrm{H}_{\max}(\mathcal{P}).$$

The results have natural counterparts for the game $\gamma(\mathcal{P})$.

updating, given a prior

Analogous results hold for utility-based information triples. Here, we only focus on updating in the Y-domain.

The setting is a general divergence function D (note, not necessarily derived from an effort-function), a preparation $\mathcal{P}$ and a prior $y_0 \in Y$ with $\mathrm{D}^{y_0} < \infty$ on $\mathcal{P}$. The associated updating triple is $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$. An optimal strategy for Nature is here called a D-projection of y_0 on $\mathcal{P}$.

If $x^* \in \mathrm{ctr}(\mathcal{P})$, the game is in equilibrium with x^* as bi-optimal state iff the Pythagorean inequality for updating,

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

holds for every $x \in \mathcal{P}$.

where does convexity come in?

Given information triple $(\Phi, \mathrm{H}, \mathrm{D})$ in Y-domain. Add assumptions:

- X is a convex topological space,
- all "views" $]y[= \{x | y \succ x\}$ are convex,
- $y \succ \overline{x}$ with $\overline{x}$ a convex combination $\overline{x} = \sum \alpha_i x_i$ iff $y \succ x_i$ for all i with $\alpha_i > 0$ ("if" suffices for some applications)
- all marginals Φ^y are affine;
- suitable (!) semi-continuity assumptions.

For every convex combination $\overline{x} = \sum \alpha_i x_i$
$\mathrm{H}\left(\sum \alpha_i x_i\right) = \sum \alpha_i \,\mathrm{H}(x_i) + \sum \alpha_i \,\mathrm{D}(x_i, \overline{x})$ and, if $\mathrm{H}(\overline{x}) < \infty$,
then, for every $y \in Y$, the compensation identity holds:
$\sum \alpha_i \,\mathrm{D}(x_i, y) = \mathrm{D}\left(\sum \alpha_i x_i, y\right) + \sum \alpha_i \,\mathrm{D}(x_i, \overline{x})$.

… continued

The condition $\forall x \in \mathcal{P} : \Phi(x, y^*) \leq \mathrm{H}(x^*)$ is central! From it, and from $x^* \in \mathcal{P}$, you conclude equilibrium of $\gamma(\mathcal{P})$ and bi-optimality of x^*. In particular, $\mathrm{H}(x^*) = \mathrm{H}_{\mathrm{max}}(\mathcal{P})$.

With convexity assumptions, $\mathrm{H}(x^*) = \mathrm{H}_{\mathrm{max}}(\mathcal{P})$ actually suffices for these conclusions!

Further elaborations for updating games with convex preparations as well as analytical existence results, exploiting convexity- and topological assumptions, can be established.

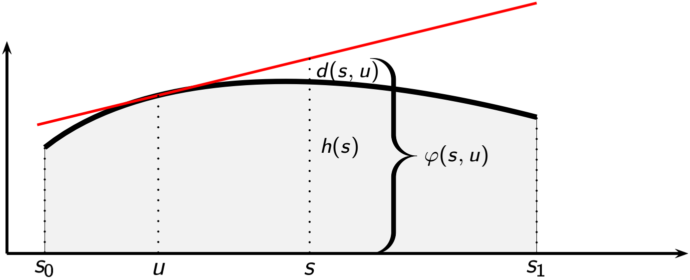
Other interpretations

Applications to Shannon theory is obvious! But a change of interpretation opens for other applications. Three indications:

- Utility can be handled via a simple change of sign as already discussed;
- What if Nature can communicate? Then we speak of Expert and Observer becomes Customer. Customer asks for advice but for despicable reasons Expert may give advice against better knowing. How to keep the expert honest? Via a paying scheme based on a proper effort function! In fact classical (Brier, 1950 on weather forecasting,...);
- Think of states as causes, and response as the transformation into associated consequences. This results in models of cause and effect. An example is problems of capacity in information theory. Then considerations of risk become important (Kuhn-Tucker theorems of inf. theory...).

Atomic triples, Bregman construction

$Y = X$, a subinterval of $]-\infty, \infty[$, $\hat{x} = x$. An information triple $(\phi, \mathrm{h}, \mathrm{d})$ in this simple setting is an atomic information triple over I. The important affinity property holds automatically by a Bregman construction based on a smooth concave entropy function h. Indeed, then $\phi(s, u) = \mathrm{h}(u) + (s - u)\,\mathrm{h}'(u)$:



… large potential…

- generator need not be smooth; this points to the good sense of extended modelling, allowing response to be set-valued;
- a natural process of integration preserves key properties;
- controls may be defined by (sub)regions corresponding to straight lines. This points to basic affine properties of the measuring process. Duality theory appears as a natural application (not worked out);
- is a representation theorem possible?

Integration of atomic triples

Integration of $(\phi, \mathrm{h}, \mathrm{d})$ over set $T = (T, \mu)$. Let $X = Y$ be "appopriate" function space of (measurable) functions $x : T \mapsto I$ for which $\int_T \mathrm{h}(x(t))d\mu(t)$ converges. Define $(\Phi, \mathrm{H}, \mathrm{D})$ by integration, i.e.

$$\begin{aligned}\Phi(x,y) &= \int_T \phi(x(t),y(t))d\mu(t)\,,\\ \mathrm{H}(x) &= \int_T \mathrm{h}(x(t))d\mu(t)\\ \mathrm{D}(x,y) &= \int_T \mathrm{d}(x(t),y(t))d\mu(t)\,.\end{aligned}$$

The basic facts, $\mathrm{d}(s,u) \geq 0$ with equality iff $u = s$ is the pointwise fundamental inequality. Examples: next slide...

Two examples

1.st example: Take $\mathrm{h}(s) = -s^2$ on $]-\infty, \infty[$ as generator. Then $\mathrm{d}(s,u) = (s-u)^2$ and by suitable integration with L^2 (or l^2) as functionspace, you enter into Hilbert space theory with $\mathrm{D}(x,y) = \|x-y\|^2$ ...

2.nd example: Take $\mathrm{h}(s) = s \ln \frac{1}{s}$ on $[0,1]$ (or...). Then $\mathrm{d}(s,u) = u - s + s \ln \frac{s}{u}$. With discrete probability distributions (w.r.t. counting measure) as function space, you enter discrete Shannon theory. With more general integration, continuous information theory with versions of Kullback-Leibler divergence are obtained.

Updating games in first example leads to standard results of projection and the connection to classical Pythagorean theorems. Updating for the second example leads to information projections and to the Pythagorean theorems of information theory.

indication of further applications

Apart from information theoretical applications, we can point to certain problems of location theory, especially Sylvesters problem, to application in statistical physics (explanation of Tsallis entropy ...), applications to statistics, especially to exponential families ...

Concluding remarks

• Key points: A quantitative and abstract theory of cognition.

• A main feature: Interpretations guide the way!

• Technical advantage: Concrete optimization problems are, typically, handled by the robustness theorem. This is in contrast to the most common approaches to optimization, where a technique based on Lagrange multipliers play the main role.

• Challenges: Consolidate! (more applications, more theoretical results, e.g. on the core and the connection to exponential families, expansion of the setting, e.g. can quantum information theory be covered?...)

Thank you for going through this appetizer!