Cognition and Inference in an abstract setting

Flemming Topsøe, topsoe@math.ku.dk
Department of Mathematical Sciences, University of Copenhagen

WITMSE2011, Helsinki, August 7-10, 2011

Two examples

Shannon Theory, MaxEnt:

states: Distributions $x = (x_i)_{i=0,1,\cdots}$;

Kerridge inaccuracy: $\Phi(x,y) = \sum x_i \log \frac{1}{y_i}$;

Entropy: $\mathrm{H}(x) = \sum x_i \log \frac{1}{x_i}$;

Divergence: $\mathrm{D}(x,y) = \sum x_i \log \frac{x_i}{y_i}$;

preparation: A set $\mathcal{P}$ of distributions, say those with given mean "energy": $\mathcal{P} = \{x | \sum_0^\infty x_i E_i = \overline{E}\}$.

Problem: Search for the MaxEnt distribution in $\mathcal{P}$.

Euclidean space, projection:

states: Elements in $X = \mathbb{R}^2$, say;

prior: $y_0 \in X$;

preparation: some (convex) subset $\mathcal{P}$ of X;

Problem: Find the projection of y_0 on $\mathcal{P}$.

Information triples, I

Goal of talk: Indicate to you that information theoretical thinking is useful in a much broader context than that known from Shannon theory; we may even free ourselves from the tie to probability based modelling.

Our start for an abstract theory: information triples:

Either effort-based: $(\Phi, \mathrm{H}, \mathrm{D})$ if $\cdots$ (see next slide)
or utility-based: $(\mathrm{U}, \mathrm{M}, \mathrm{D})$, i.e. $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$ is effort-based.

Examples:

1: $\Phi(x, y) = \sum x_i \log \frac{1}{y_i}$, $\mathrm{H}(x) = \sum x_i \log \frac{1}{x_i}$,
$\mathrm{D}(x, y) = \sum x_i \log \frac{x_i}{y_i}$.

2: $\mathrm{U}(x, y) = \|x - y_0\|^2 - \|x - y\|^2$,
$\mathrm{M}(x) = \|x - y_0\|^2$, $\mathrm{D}(x, y) = \|x - y\|^2$.

State space X; elements are states or truth instances, (x).
Will study preparations, i.e. non-empty subsets $\mathcal{P} \subseteq X$.
Belief reservoir Y; elements are belief instances, (y).
For this talk: $X = Y$.

Information Triples, II, (Φ, H, D) and (U, M, D)

Φ and D (U and D) are defined on $X \times Y$, H (M) on X.

AXIOM 1 (the basics) $\Phi > -\infty$ ($U < +\infty$)
$\Phi(x, y) = H(x) + D(x, y)$, the linking identity ($U = M - D$)
$D(x, y) \geq 0$ with equality iff $y = x$, the fundamental inequality .

Φ is the description or the effort function,
H is min-effort or entropy, D is divergence.
(U the utility M the max-utility)
Add convexity! Use $\overline{x} = \sum \alpha_i x_i$ for a convex combination.

AXIOM 2 (affinity) X is a convex space and, for each y, the marginal function Φ^y (U^y) is affine.

First consequences, convexity properties

Lemma

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

(ii) If $\mathrm{H}(\overline{x}) < \infty$, $y \in Y$, then compensation identity holds:

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

Proof (i): rhs $= \sum \alpha_i \Phi(x_i, \overline{x}) = \Phi(\overline{x}, \overline{x}) = \mathrm{H}(\overline{x})$.

(ii): lhs of (i)+lhs of (ii)

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

Now subtract $\mathrm{H}(\overline{x})$. □

Updating

Problem: Given prior y_0, to define utility function $\mathrm{U}_{|y_0}$ such that $\mathrm{U}_{|y_0}(x,y)$ is a measure of the updating gain when truth is x and your posterior belief is y. Typically, the posterior is sought among y's in a given preparation $\mathcal{P}$.

1. Defined as saved effort: Based on triple $(\Phi, \mathrm{H}, \mathrm{D})$:

$$\mathrm{U}_{|y_0}(x,y) = \Phi(x,y_0) - \Phi(x,y) \tag{1}$$
$$= \mathrm{D}(x,y_0) - \mathrm{D}(x,y)\,. \tag{2}$$

2. Directly via D: Given only D, use (2) as definition. This gives utility-based triple $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$. Technically, assume that $\mathrm{D}^{y_0} < \infty$ on preparations $\mathcal{P}$ you want to consider.

This defines genuine triples satisfying axioms 1 and 2 *iff* D satisfies the fundamental inequality and the compensation identity.

Conclude: Problems of updating can be treated as special cases of inference for information triples.

Games of information: Observer versus Nature

Game $\gamma = \gamma(\mathcal{P}) = \gamma(\mathcal{P}|\Phi)$ has Φ as objective function, Nature as maximizer with strategies $x \in \mathcal{P}$, Observer as minimizer with strategies $y \in Y$.

Values of γ are, for Nature MaxEnt and, for Observer, MinRisk:
$\mathrm{H}_{\max}(\mathcal{P}) = \sup_{x\in\mathcal{P}} \mathrm{H}(x) = \sup_{x\in\mathcal{P}} \inf_y \Phi(x,y)$.
$\mathrm{Ri}_{\min}(\mathcal{P}) = \inf_y \mathrm{Ri}(y) = \inf_y \sup_{x\in\mathcal{P}} \Phi(x,y)$.
$x^* \in \mathcal{P}$ optimal strategy for Nature $\therefore \mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$.
$y^* \in Y$ optimal strategy for Observer $\therefore \mathrm{Ri}(y^*) = \mathrm{Ri}_{\min}(\mathcal{P})$.
Minimax inequality: $\mathrm{H}_{\max} \leq \mathrm{Ri}_{\min}$.
If $\mathrm{H}_{\max} = \mathrm{Ri}_{\min} < \infty$, γ is in equilibrium.

If γ is in equilibrium and both players have optimal strategies, these are unique and coincide. The strategy in question $y^* = x^*$ is the bi-optimal strategy.

Nash and Pythagoras

Theorem [Axiom 1] Given $y^* = x^* \in \mathcal{P}$ with $\mathrm{H}(x^*) < \infty$. Then the following conditions are equivalent:

- $\gamma(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy;
- The Nash-saddle-value inequalities hold;
- For all $x \in \mathcal{P}$, the abstract Pythagorean inequality holds:

$$\mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}(x^*) \tag{3}$$

$$\Big(\mathrm{M}(x) \geq \mathrm{D}(x, y^*) + \mathrm{M}(x^*) \text{ for utility-based model } \Big) \tag{4}$$

$$\Big(\mathrm{D}(x, y_0) \geq \mathrm{D}(x, y^*) + \mathrm{D}(x^*, y_0) \text{ for updating } \Big) . \tag{5}$$

With $\mathrm{D}(x, y) = \|x - y\|^2$, (5) is the classical inequality.

Theorem [Axioms 1+2] The condition that x^* is an optimal strategy for Nature is sufficient to ensure that (3)[(4)/(5)] holds. For the updating model the condition is that x^* is the D-projection of y_0 on $\mathcal{P}$.

Adding a geometric flavour

We only do this for the models of updating. Two type of sets will be involved: open divergence balls and open half spaces:

$$B(y_0, r) = \{D(x, y_0) < r\}$$

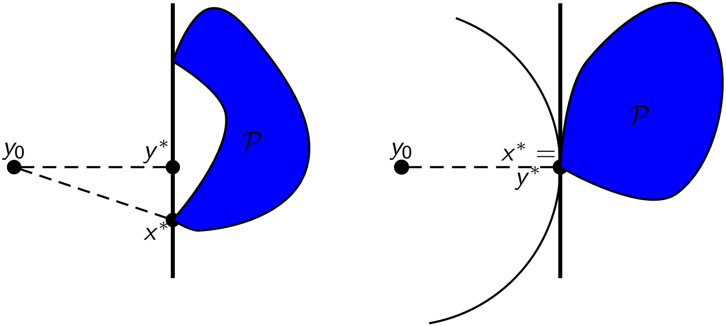
$$\sigma^+(y|y_0) = \{U_{|y_0} < D(y, y_0)\}$$
$$= \{D(x, y_0) < D(x, y) + D(y, y_0)\}\,.$$

The sizes of these sets are, respectively, r and $D(y, y_0)$.

For the updating game $\gamma(\mathcal{P}|\,U_{|y_0})$, the MinDiv-value $D^{y_0}_{\min}(\mathcal{P})$ is the size of the largest ball $B(y_0, r)$ which is external to $\mathcal{P}$ (i.e. contained in the complement of $\mathcal{P}$), and the other value of the game, the maximal guaranteed updating gain is, loosely expressed, the size of the largest half space external to $\mathcal{P}$.

In particular, $\gamma(\mathcal{P}|\,U_{|y_0})$ is in equilibrium and has a bi-optimal strategy if and only if, for some $y \in \mathcal{P}$, the half-space $\sigma^+(y|y_0)$ 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 $\mathcal{P}$.

optimal strategies under no equilibrium/ and under equilibrium



Topics left out

- Tsallis entropy
- Bregman divergencies
- Feasible preparations
- Control and description
- Core, an abstract notion generalizing exponential families

Instead of conclusions

• Should Shannon Theory be taught and learned this way?
• Is the philosophical approach important – and helpful?
• Is the focus on game theory justified?
• Is the abstract approach also the right entrance point to areas of pure mathematics (optimization, duality theory ...)?
• – and to areas of (statistical) physics?
• Is the theory a good "selling argument" which could pave the way for more widespread adoption and recognition of ideas of Information Theory as initiated by Shannon?

My preliminary answers: I believe in a great potential of the theory indicated, but to which extent it is justified as a "stand alone theory" and to which extent it is a supplement to existing theories is of course not clear right now.