Cognition and inference

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

the menu

First a little bit about the chef ...
... and then to the menu, main ingredients:

- Philosophy, emphasis on interpretations, especialy pursuing the theme "Nature versus Observer" (Nature holds the truth, Observer seeks the truth but is confined to belief and may with time acquire knowledge ...).
- Abstraction, no reference to probability.

Ingarden & **Urbanik 1962:** "... 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 ..."
Kolmogorov $\approx$ **1970:** "Information theory must preceed probability theory and not be based on it"

two examples to have in mind

All our models are based on a function $\Phi = \Phi(x, y)$ of two variables, description effort; x represents truth, y belief.

Shannon model, discrete case $\Phi(x, y) = \sum_{i \in \mathbb{A}} x_i \ln \frac{1}{y_i}$ where $x = (x_i)_{i \in \mathbb{A}}$ and $y = (y_i)_{i \in \mathbb{A}}$ are probability distributions over an alphabet $\mathbb{A}$.

A Hilbert-space model Fix y_0 and take $\Phi = \Phi_{|y_0}$ to be $\Phi(x, y) = \|x - y\|^2 - \|x - y_0\|^2$. (Note: $\geq \Phi(x, x)$).

Updating, general idea: Construct a new model from an old one, Φ, by defining updating gain from a prior y_0 to a posterior y to be $\Phi(x, y_0) - \Phi(x, y)$. This function taken with the opposite sign can be used as a new description effort: $\Phi_{|y_0}(x, y) = \Phi(x, y) - \Phi(x, y_0)$.

Elements of the meal

Sets: X State space (truth!), $Y \supseteq X$ Belief Reservoir.

Special subsets: Y_{det} to express certain belief. And then various non-empty subsets of X, preparations (more later).

Relations and functions: $X \otimes Y \subseteq X \times Y$: Domination.
Write $y \succ x$ for $(x, y) \in X \otimes Y$ and assume $x \succ x$ for all x.
A situation $(x, y) \in X \otimes Y$ is a perfect match if $y = x$ and a certain belief if $y \in Y_{\text{det}}$.

$\Phi : X \otimes Y \to]-\infty, \infty]$: description effort or description.
Φ must be calibrated: $\Phi(x, y) = 0$ for certain beliefs.
Observer should adapt Φ to the world! But how?

Key principle Φ satisfies the perfect match principle, PMP, (or is proper) if, for fixed x, Φ is minimized under a perfect match and not otherwise (unless $\Phi(x, x) = \infty$).

Elements of information (for a given proper Φ)

Information is information about truth,
e.g. full information "x" or partial information "$x \in \mathcal{P}$".

Quantitatively, information is saved effort

Thus, $\Phi(x, y)$ = value to Observer of information "x" in a situation with belief y. The unit of description effort is then also a unit of information. (Information is physical!)

Introduce:

Entropy $\mathrm{H}(x)$ = minimal effort required ;
Divergence $\mathrm{D}(x, y)$ = excess description effort.

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

$(\Phi, \mathrm{H}, \mathrm{D})$ is an information triple. Basic axioms:

$\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y)$ (linking identity),

$\mathrm{D} \geq 0$ with equality iff there is a perfect match (fundamental inequality of information theory, FI).

A good meal needs ... preparations

They tell us what *can* be known, and thus provide *limits to knowledge*. They are closely related to exponential families.

Basic preparations (preparations of genus 1) are preparations of the form $\mathcal{P}^y(h) = \{x|\Phi(x,y) = h\}$. They are of strict type. The corresponding slack type preparations are: $\mathcal{P}^y(h^{\leq}) = \{x|\Phi(x,y) \leq h\}$.

With $\mathbf{b} = (b_1, \cdots, b_n)$ and $\mathbf{h} = (h_1, \cdots, h_n)$, we put

$\mathcal{P}^{\mathbf{b}}(\mathbf{h}) = \bigcap_{\nu \leq n} \mathcal{P}^{b_\nu}(h_\nu)$ (if non-empty).

Given $\mathbf{b}$, we denote by $\mathbb{P}^{\mathbf{b}}$ the preparation family of all preparations of the form $\mathcal{P}^{\mathbf{b}}(\mathbf{h})$ for some level values $\mathbf{h} = (h_1, \cdots, h_n)$.

Instructive to look at this for updating in Hilbert space...

... and more preparations

$y \in X$ is robust for a preparation $\mathcal{P}$ if $\Phi(x,y)$ is constant over $\mathcal{P}$, i.e. if, for some h, the level of robustness, $\mathcal{P} \subseteq \mathcal{P}^y(h)$.

The set of y which are robust for $\mathcal{P}$ is the core of $\mathcal{P}$: $\mathrm{core}(\mathcal{P}) = \{y \in X | \exists h : \mathcal{P} \subseteq \mathcal{P}^y(h)\}$.

If $\mathbb{P}$ is a preparation family, we define the core of $\mathbb{P}$ by

$$\mathrm{core}(\mathbb{P}) = \bigcap_{\mathcal{P} \in \mathbb{P}} \mathrm{core}(\mathcal{P}) \quad \text{or} \quad \mathrm{core}(\mathbb{P}) = \{y \in X | \mathbb{P} \prec \mathbb{P}^y\}\,.$$

If $\mathbb{P}$ is the family of *all* preparations, then $\mathrm{core}(\mathbb{P}) = \mathrm{core}(X)$ and this set is either empty or a singleton. In the latter case, say $\mathrm{core}(X) = \{u\}$, u is the uniform state over X.

the scene is set for fight: Nature$\leftrightarrow$Observer

The game $\gamma(\mathcal{P}) = \gamma(\Phi, \mathcal{P})$: Φ is the objective function, Nature maximizer, Observer minimizer. Nature strategies: x's in $\mathcal{P}$. Observer strategies: beliefs $y \succ \mathcal{P}$ ($\forall x \in \mathcal{P} : y \succ x$).

MaxEnt is value for Nature, MinRisk value for Observer:
$\mathrm{H}_{\max}(\mathcal{P}) = \sup_{x\in\mathcal{P}} \mathrm{H}(x) = \sup_{x\in\mathcal{P}} \inf_{y\succ x} \Phi(x, y)$.
$\mathrm{Ri}_{\min}(\mathcal{P}) = \inf_{y\succ\mathcal{P}} \mathrm{Ri}(y) = \inf_{y\succ\mathcal{P}} \sup_{x\in\mathcal{P}} \Phi(x, y)$.
Note: $\mathrm{Ri}(y) = \mathrm{Ri}(y|\mathcal{P})$.

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

If $\mathrm{H}_{\max}(\mathcal{P}) = \mathrm{Ri}_{\min}(\mathcal{P})$ is finite, $\gamma(\mathcal{P})$ is in equilibrium.

The best we can hope for: To deal with a game in equilibrium which has a bioptimal strategy x^* which we can easily identify (thus x^* optimal for both players is sought).

first main course: Pythagoras!

The Pythagorean theorem, direct and dual form.
Assume that $x^* \in \mathcal{P} \subseteq \mathcal{P}^{x^*}(h^{\leq})$ with $h = \mathrm{H}(x^*)$ finite.

Then $\gamma(\mathcal{P})$ is in equilibrium with $\mathrm{H}_{\max}(\mathcal{P}) = \mathrm{Ri}_{\min}(\mathcal{P}) = h$, and x^* is the unique bioptimal strategy. Furthermore,

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

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

If $\mathcal{P} \subseteq \mathcal{P}^{x^*}(h)$, equality holds in the Pythagorean inequality.

Corollary Let $\mathbf{b} = (b_1, \cdots, b_n)$ and consider the family $\mathbb{P}^{\mathbf{b}}$. If $x^* \in \mathrm{core}(\mathbf{b})$, then there is a preparation $\mathcal{P}$ in the family for which $\gamma(\mathcal{P})$ is in equilibrium with x^* as bioptimal strategy.
In fact, with $h_\nu = \Phi(x^*, b_\nu)$ for $\nu \leq n$, $\mathcal{P} = \mathcal{P}^{\mathbf{b}}(\mathbf{h})$ is the one.

more delicate probabilistic models

We now allow Φ of the form: $\Phi(x,y) = \sum_{i\in\mathbb{A}} \pi(x_i,y_i)\kappa(y_i)$.

Instead of x_i you find $\pi(x_i,y_i)$, the interactor π operating on pairs of probabilities, one true, the other believed. We assume that π is sound, i.e. $\pi(s,t) = s$ for a perfect match ($t = s$).

Interpretation: $\pi(s,t)$ is the force you *perceive* as attached to an event with true probability s and believed probability t, e.g.: $\pi_q(s,t) = qs + (1-q)t$. Determines the world $\mathcal{W}_q$. $\mathcal{W}_1$: the classical or Shannon world. $\mathcal{W}_0$: a black hole.

... and instead of $\ln\frac{1}{y_i}$ you find the descriptor κ operating on a believed probability.

Interpretation: κ determines the *cost of information*. It must satisfy $\kappa(1) = 0$, $\kappa'(1) = -1$ (normalization).

Problem: Given π, choose κ such that Φ determined by π and κ is proper. In other words: adapt κ to the world!

Tsallis entropy in special dressing, 2.nd main dish

Theorem. Recall required form: $\Phi(x,y) = \sum_{i\in\mathbb{A}} \pi(x_i,y_i)\kappa(y_i)$.

- Given π, at most one descriptor κ is proper;
- No descriptor is proper for $\mathcal{W}_q$ if $q \le 0$; however, $q = 0$ is a singular case (with H=degr.freedom, $\mathrm{D} \equiv 0$, $\kappa(t) = t^{-1} - 1$);
- For $q > 0$, the ideal descriptor κ_q exists. It is in the power hierarchy and given by $\kappa_q(t) = \ln_q \frac{1}{t}$, the q-logarithm of $\frac{1}{t}$ $\left(= \frac{1}{1-q}\left(t^{q-1} - 1\right)\right)$. The associated entropy function is Havrda&Charvát-Lindhard&Nielsen-Tsallis$\cdots$ entropy;
- Again for $q > 0$, other mean values (e.g. geometric and harmonic) determine the same ideal descriptor;
- To prove FI, simply prove PFI, the pointwise fundamental inequality, $\delta \ge 0$, where the divergence generator δ is defined by $\delta(s,t) = \big(\pi(s,t)\kappa(t) + t\big) - \big(s\kappa(s) + s\big)$ (so that $\mathrm{D}(x,y) = \sum \delta(x_i,y_i)$).

Controls for $\Phi(x,y)=\sum_{i\in\mathbb{A}}\pi(x_i,y_i)\kappa(y_i)$

Rewrite Φ as $\Phi(x,y)=\sum_{i\in\mathbb{A}}\pi(x_i,y_i)w_i$ with $w_i=\kappa(y_i)$. Then w, the control adapted to y points more directly than y to action by Observer (design of experiments...).

Recall: Good, 1952: belief is a tendency to act!

The inverse function to κ is denoted ρ and termed the probability checker: $\rho(a)$ tells you how rare an event you can *control* or *describe* with κ if you have a units (nats) at your disposal (one defines $\rho(a)=0$ if $\kappa(0)\le a$).

Krafts inequality checks if, given $(w_i)_{i\in\mathbb{A}}$, you can hope to use these numbers as efforts (allocated nats, classically corresponding to code lengths). It states: $\sum_{i\in\mathbb{A}}\rho(w_i)\le 1$.

By the one-to-one correspondance $y\leftrightarrow w$ we can choose to express findings in terms of beliefs or controls (or a mixture!).

the desert: crème de la crème

Given: Model $\mathcal{P}$ from a family $\mathbb{P}^{\mathbf{b}}$.
Wanted: 1) MaxEnt distribution 2) I-projection of prior y_0 on $\mathcal{P}$ or, equivalently, $\operatorname{argmin}_{x \in \mathcal{P}} \mathrm{D}(x, y_0)$.
Observation: 2) is reduced to 1) by switching to $\Phi_{|y_0}$.
Strategy for 1): Determine $\mathrm{core}(\mathbb{P}^{\mathbf{b}})$, choose $y \in \mathrm{core}(\mathbb{P}^{\mathbf{b}}) \cap \mathcal{P}$ - and you are done!
Limitation: We only consider the worlds $\mathcal{W}_q$.
Special for these worlds: With $y \leftrightarrow w$, sets of the form $\{\Phi(x,y) = \text{const.}\}$ are of the form $\{\sum_{i \in \mathbb{A}} x_i w_i = \text{const.}\}$.

Analysis: Let $\mathcal{P} = \bigcap_1^n \mathcal{P}^{b_\nu}(h_\nu) \in \mathbb{P}^{\mathbf{b}}$ be of genus n. Then $\mathcal{P} = \bigcap_1^n \{\sum_{i \in \mathbb{A}} x_i w_{\nu,i} = h'_\nu\}$ which is $\subseteq$ some $\{\mathcal{P}^y(h)\}$ if (with $y \leftrightarrow w$) $\subseteq$ some set $\{\sum x_i w_i = h'\}$ and this is OK if $\exists \alpha, \beta = (\beta_1, \cdots, \beta_n)$ s.t. $w = \alpha + (\beta_1 w_1 + \cdots + \beta_n w_n)$.

Theorem ...and only then!

...more of the desert

So the sought $y \leftrightarrow w$ must satisfy $w = \alpha + \sum_1^n \beta_\nu w_\nu$ for suitably chosen α and $\beta = (\beta_1, \cdots, \beta_n)$. Requirements to these constants: $\sum_{i \in \mathbb{A}} \rho_q\big(\alpha + \sum_1^n \beta_\nu w_{\nu,i}\big) = 1$ (Kraft's (in)equality!); this determines α.
And then the β's are determined from the requirement $y \in \mathcal{P}$.

Classically ($q = 1$): Then $\rho_1 : a \mapsto \exp(-a)$ and one obtains $\alpha = \ln Z(\beta)$ with Z the partition function :
$Z(\beta) = \sum_{i \in \mathbb{A}} \exp \sum_1^n -\beta_\nu w_{\nu,i}$.
Thus the possible y are from the exponential family associated with the problem, i.e. distributions of the form $y_i = \exp(-\alpha - \sum_1^n \beta_\nu w_{\nu,i})$ with $\alpha = \ln Z(\beta)$.

Thus the core coincides with the exponential family.

The analysis for 2) leads to the exponential family given by $y_i = y_{0,i} \exp(-\alpha - \sum_1^n \beta_\nu w_{\nu,i})$.

end of meal

A theory of information freed from a tie to probability *is* possible – and useful. Probabilistic models appear as important examples.

Velbekom'!