Description, Entropy and Divergence
a non-probabilistic approach

Igor Vajda in memoriam – friend and scholar

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

information without probability

Inspiration: Ingarden and Urbanik: "... 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 ..."

Previous work: Kampé de Fériet, Kolmogorov, Jumarie, Shafer and Vovk.

Areas involved: semiotics, philosophy, symbolic linguistics, social information, learning theory, logic ...

Criticism of some of this: quite philosophical, not open to quantitative analysis, impractical, of theoretical interest only, not downward compatible with Shannon theory.

Recent inspiration: letter from Igor...

interpretations first or axiomatics first?

Interpretations and operational definitions, of course!
But we are past that stage. So let us axiomatize!

• Ingredients: A set X, the state space, another set Y, the action space, a map between them. $x \mapsto \hat{x}$, the response. Equivalence defined by $x_1 \equiv x_2 \Leftrightarrow \hat{x}_1 = \hat{x}_2$.

• More ingredients: A function $\Phi : X \times Y \mapsto]-\infty, \infty]$, description effort, a function $\mathrm{H} : X \mapsto]-\infty, \infty]$, entropy, a function $\mathrm{D} : X \times Y \mapsto [0, \infty]$, divergence.

Axiom 1 For all $(x, y) \in X \times Y$:
$\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y)$ (linking identity)
$\mathrm{D} \geq 0$ and $\mathrm{D}(x, y) = 0 \Leftrightarrow y = \hat{x}$ (fundamental inequality)

Classically: Φ Kerridge inaccuracy H Shannon entropy D Kullback-Leibler divergence

given by, respectively, $\sum p_i \ln \frac{1}{q_i}$, $\sum p_i \ln \frac{1}{p_i}$ and $\sum p_i \ln \frac{p_i}{q_i}$.

adding extra structure; convexity, topology

Axiom 2 X convex, Φ affine in x:
$\forall y \in Y$, $\forall \alpha = (\alpha_x)$: $\Phi\Big(\sum_x \alpha_x x, y\Big) = \sum_x \alpha_x \Phi(x, y)$.

$\alpha = (\alpha_x)_x \in \mathrm{MOL}(X)$: α_x's ≥ 0, finite support, $\sum_x \alpha_x = 1$.

Axiom 3 A topology τ on X renders algebraic operations continuous. Further, for $y_0 \in Y$, $x \mapsto \mathrm{D}(x, y_0)$ and for $x_0 \in X$, $x \mapsto \mathrm{D}(x_0, \hat{x})$ are lower semi-continuous.

Axiom 4 If (x_n) is D-Cauchy , i.e. if $\mathrm{D}(x_n, y_{n,m}) \to 0$ for $n, m \to \infty$ with $y_{n,m} = (\frac{1}{2}x_n + \frac{1}{2}x_m)\hat{}$, then it has a τ-convergent subsequence.

Write $x_n \twoheadrightarrow x$ if $\mathrm{D}(x_n, \hat{x}) \to 0$ (strong convergence).

Observations: $x_n \twoheadrightarrow x \Rightarrow$ D-Cauchy;
$x_n \twoheadrightarrow x_1$, $x_n \to x_2 \Rightarrow x_1 \equiv x_2$;
$x_n \twoheadrightarrow x \Rightarrow x_n \to x$ if response is injective.

consequences of Axioms 1 and 2

Put $\sum_x \alpha_x x = \overline{x} = b(\alpha)$, the barycentre of α and define information (transmission) rate as $\mathrm{I}(\alpha) = \sum_x \alpha_x \,\mathrm{D}(x, \overline{x})$.

Theorem α's $\in \mathrm{MOL}(X)$, $\beta \in \mathrm{MOL}(\mathrm{MOL}(X))$.

(1) $\mathrm{H}\left(\sum_x \alpha_x x\right) = \sum_x \alpha_x \,\mathrm{H}(x) + \mathrm{I}(\alpha)$.

(2) $\sum_x \alpha_x \,\mathrm{D}(x, y) = \mathrm{D}(\sum_x \alpha_x x, y) + \mathrm{I}(\alpha)$

(3) $\mathrm{I}\left(\sum_\alpha \beta_\alpha \alpha\right) = \sum_\alpha \beta_\alpha \,\mathrm{I}(\alpha) + \mathrm{J}(\beta)$ where

$$\mathrm{J}(\beta) = \sum_\alpha \beta_\alpha \,\mathrm{D}\left(b(\alpha), b(\alpha_0)\hat{}\right),\ \alpha_0 = b\left(\sum_\alpha \beta_\alpha \alpha\right).$$

Note: (1) OK. re (2): valid for all y if only $\mathrm{H}(\overline{x}) < \infty$.
re (3): OK if $\beta_\alpha > 0 \Rightarrow \mathrm{H}(b(\alpha)) < \infty$.

Proof (1): Look at rhs, use linking, then affinity.
(2): Add $\sum_x \alpha_x \,\mathrm{D}(x, y)$ to both sides of (1), apply linking and affinity to rhs, subtract $\mathrm{H}(\overline{x})$ from both sides.
(3): Special case of (1) applied to new information triple. □

... continued

Note that standard concavity- and convexity results follow directly from (1), (2) and (3).

Also note that, for the classical probabilistic setting, the quantity $\mathrm{I}(\alpha)$ for α of the form $\frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$ is the nowadays quite important Jensen-Shannon divergence. It is the square of a metric – a central property of a type studied by Österreicher and I. Vajda and others. First proof is by Endres and Schindelin 2003.

The identity (2), perhaps best written:

$$\sum_x \alpha_x \, \mathrm{D}(x, y) = \mathrm{D}(\overline{x}, y) + \sum_x \alpha_x \, \mathrm{D}(x, \overline{x})$$

with $\overline{x} = b(\alpha) = \sum_x \alpha_x x$. This is the compensation identity (authors terminology). It also holds for density matrices and is then called Donalds identity. It appears in some proofs, in particular of the result we now turn to.

a game between Nature and Observer

For a preparation $X_0 \subseteq X$, $\gamma(X_0)$ is the two-person zero-sum game with Φ as objective function, X_0 and Y as strategy sets for Nature (maximizer) and Observer (minimizer).

The value for Nature is the maximum entropy value :

$$\sup_{x \in X_0} \inf_{y \in Y} \Phi(x, y) = \sup_{x \in X_0} \mathrm{H}(x) = \mathrm{H}_{\max}(X_0) = \mathrm{H}_{\max} \ .$$

The value for Observer, is the minimum risk value :

$$\inf_{y \in Y} \sup_{x \in X_0} \Phi(x, y) = \inf_{y \in Y} \mathrm{R}(y) = \mathrm{R}_{\min}(X_0) = \mathrm{R}_{\min} \ .$$

$\gamma(X_0)$ is in equilibrium if $\mathrm{H}_{\max} = \mathrm{R}_{\min} < \infty$.
Optimal strategies: For Nature: $x \in X_0$ with $\mathrm{H}(x) = \mathrm{H}_{\max}$; for Observer: $y \in Y$ with $\mathrm{R}(y) = \mathrm{R}_{\min}(X_0)$.
And $x \in X$ is an $\mathrm{H}_{\max}$-attractor if $x_n \twoheadrightarrow x$ for every sequence (x_n) of elements in X_0 with $\mathrm{H}(x_n) \to \mathrm{H}_{\max}$.

an existence theorem

Theorem (MaxEnt) If X_0 is convex and $\mathrm{H}_{\max}(X_0) < \infty$, $\gamma(X_0)$ is in equilibrium and Observer has a unique optimal strategy y^*. An $\mathrm{H}_{\max}$-attractor x^* exists and $y^* = \widehat{x^*}$. All $\mathrm{H}_{\max}$-attractors are equivalent. For $x \in X_0$ and for $y \in Y$,

$\mathrm{H}(x) + \mathrm{D}(x, y^*) \leq \mathrm{H}_{\max}(X_0)$ (Pythagorean inequality)

$\mathrm{R}_{\min}(X_0) + \mathrm{D}(x^*, y) \leq \mathrm{R}(y)$ (dual Pythagorean inequality).

Corollary (updating) Assume $X = Y, \forall x : \hat{x} = x$. Let X_0 be convex closed, let $y_0 \in X$ be a prior with $\mathrm{D}(x, y_0) < \infty \forall x \in X$. Then the I-projection of y_0 on X_0, $x^* = \mathrm{argmin}_{x \in X_0} \mathrm{D}(x, y_0)$, exists and is characterized by:
$\forall x \in X_0 : \mathrm{D}(x, y_0) \geq \mathrm{D}(x, x^*) + \mathrm{D}(x^*, y_0)$ (P.I., classical form).

Proof: With $\Phi_{|y_0} = -$updating gain, i.e. $\Phi_{|y_0}(x, y) = \mathrm{D}(x, y) - \mathrm{D}(x, y_0)$, note that $\Big(\Phi_{|y_0}(x, y), -\mathrm{D}(x, y_0), \mathrm{D}(x, y)\Big)$ defines an information triple. Apply theorem above! $\square$

a geometric formulation

Using special sets: divergence balls, sets of the form $\{D^y \le r\}$, and half-spaces, sets of the form $\{\Phi^y \le h\}$, the game-theoretical notions can be expressed in "geometrical" terms. As an example, even without convexity, the condition

$$x \in X_0 \subseteq \{\Phi^{\hat{x}} \le H(x)\}$$

expresses equilibrium of $\gamma(X_0)$ in conjunction with optimality of the strategies x and $\hat{x}$.

Very illuminating is the consideration of the triple in Hilbert space with description of the form $\|x - y_0\|^2 - \|x - y_0\|^2$ for some prior y_0 and X_0 some subset of a hyperspace. This will illustrate what can happen if X_0 in previous theorem is not convex (best done on the blackboard! – I mean whiteboard).

methods for the generation of information triples

We sketch some possibilities and avoid technicalities (e.g. regarding possible indefinite values of functions). Further, we mainly have axioms 1 and 2 in mind with $X = Y$ and with response equal to the identity map.
Possibilities:

- constructions from either Φ, H or D
- restriction
- expansion
- atomic triples
- construction by integration
- equivalence
- relativization
- randomization

Prospect: use possibilities (especially atomic triples, integration and restriction) to develop representation theorem!

atomic triples: triples over real intervals

Take $X = Y = I = [0, \infty[$, say. An atomic triple over I is a triple (ϕ, h, d) such that, in order to satisfy Axiom 1:
$\phi(s,t) = h(s) + d(s,t)$; $d(s,t) \geq 0$; $d(s,t) = 0 \Leftrightarrow t = s$.

Axium 2 requires ϕ of the form $\phi(s,t) = x\kappa(t) + \xi(t)$.

Example ($I = \mathbb{R}$): With prior t_0, take
$\phi(s,t) = (s-t)^2 - (s-t_0)^2$, $h(s) = -(s-t_0)^2$,
$d(s,t) = (s-t)^2$. $\square$

Related to construction of Bregman: Let h, the Bregman generator be strictly concave and define triple (ϕ, h, d) by:
$\phi(s,t) = h(t) + (s-t)h'(t)$,
$d(s,t) = h(t) - h(s) + (s-t)h'(t)$.

Examples: 1) $h(s) = x \ln \frac{1}{x}$ on $[0, \infty[$, Shannon generator.
2) $h(s) = -(s-t_0)^2$ on $\mathbb{R}$ generates previous example. (tbc)

more examples, power generators

For $q > 0$ take $h_q(s) = \frac{1}{1-q}\Big(s^q - s\Big)$, the power generators .
Note that they are normalized , i.e. $h_q(1) = 0$, $h_q'(1) = -1$.
With $h_1(s) = s \ln \frac{1}{s}$, $\lim_{q\to 1} h_q = h_1$. We find

$$\phi_q(s,t) = \Big(\tfrac{q}{1-q} s t^{q-1} + t^q - \tfrac{1}{1-q} s\Big),$$

$$d_q(s,t) = \Big(\tfrac{q}{1-q} s t^{q-1} + t^q - \tfrac{1}{1-q} s^q\Big).$$

By integration and restriction to pairs (P, Q) of probabilities and imcomplete probabilities ($\sum_i q_i \leq 1$, q_i's ≥ 0) we find

$$\Phi_q(P,Q) = \sum_i \Big(\tfrac{q}{1-q} p_i q_i^{q-1} + q_i^q - \tfrac{1}{1-q} p_i\Big),$$

$$\mathrm{H}_q = \tfrac{1}{1-q} \sum_i (p_i^q - p_i),$$

$$\mathrm{D}_q(P,Q) = \sum_i \Big(\tfrac{q}{1-q} p_i q_i^{q-1} + q_i^q - \tfrac{1}{1-q} p_i^q\Big).$$

H_q suggested by Havrda and Charvát , rediscovered by others, including Lindhard and Nielsen, as well as Tsallis for statistical physics, where they generated immense interest.

equivalence, relativization, triples via divergence

If $(\Phi, \mathrm{H}, \mathrm{D})$ satisfies the axioms, so does $(\Phi + f, \mathrm{H} + f, \mathrm{D})$ when f is affine on X. We talk about equivalent triples. As f we can use all y-sections $\Phi^y : x \mapsto \Phi(x, y)$ and linear combinations of these. Related to previous discussion consider relativization with y_0 as prior arises with $f = -\Phi^{y_0}$ (assumed finite). Quantities in the new relativized triple only depend on the original divergence D. Therefore, we ask: What is needed for some function D to generate information triples as above?

Theorem Given $\mathrm{D} : X \times Y \to [0, \infty[$ with $\mathrm{D}(x, y) = 0 \Leftrightarrow y = \hat{x}$. Then, a necessary and sufficient condition that the triples $\left(\Phi_{|y_0}, -\mathrm{D}^{y_0}, \mathrm{D}\right)$ with $\Phi_{|y_0}(x, y) = \mathrm{D}(x, y) - \mathrm{D}(x, y_0)$ define valid information triples for every choice of prior $y_0 \in Y$ is that D satisfies the compensation identity – or, equivalently, that $\sum \alpha_x \mathrm{D}(x, y) - \mathrm{D}(\sum \alpha_x x, y)$ is independent of y for $\alpha \in \mathrm{MOL}(X)$.

a problem of Sylvester treated by randomization

Sylvester: "To determine the point in the plane with the smallest maximal distance to a given set of points".

Let $\Delta \subseteq \mathbb{R}^n$ be finite, take $X = \mathrm{MOL}(\Delta)$, $Y = co(\Delta)$ and "barycenter of" as response. Define an information triple by:

$\Phi(\alpha, y) = \sum_{x\in\Delta} \alpha_x \|x - y\|^2$
$\mathrm{H}(\alpha) = \sum_{x\in\Delta} \alpha_x \|x - b(\alpha)\|^2$
$\mathrm{D}(\alpha, y) = \|b(\alpha) - y\|^2$.

You can then derive a "Kuhn-Tucker type theorem":

Theorem If $\alpha^* \in \mathrm{MOL}(\Delta)$ and $y^* = b(\alpha^*)$ are such that, for some constant r, $\|x - y^*\| \leq r$ for all $x \in \Delta$ with equality for x's with $\alpha^*_x > 0$, then y^* is the solution to Sylvester's problem and r is the minimax distance sought.

The proof is modelled after a proof of the Gallager-Ryabko theorem (or the capacity-redundancy theorem) and can be formulated in a more abstract version containing both results.

$\boxed{\text{thank you, this is all – or is it?}}$