Modeling information – abstract and probabilistic; the truth-belief-data interaction

[Excuse clumsy title!]

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences

Presentation at the “beyond Shannon”workshop in
Venice, December 29-30, 2008

Part I: Information triples, non-probabilistic modeling

Kolmogorov ($\approx$ 1970): *information theory must precede probability theory and not be based on it.*
Claim: For standard tasks, mainly related to optimization problems, this may be achieved in a theory operating with description cost $\Phi = \Phi(x,y)$, entropy $\mathsf{H} = \mathsf{H}(x)$ and divergence $\mathsf{D} = \mathsf{D}(x,y)$.

Example (Shannon type, pointwise or local form)

$$\Phi = x \; \mathit{ln}\frac{1}{y} + y\,, \quad \mathsf{H} = x \; \mathit{ln}\frac{1}{x} + x \quad \mathsf{D} = x \; \mathit{ln}\frac{x}{y} + y - x\,.$$

Example (Shannon type, standard (?) accum. form)

$$\Phi = \sum\left(x_i \; \mathit{ln}\frac{1}{y_i} + y_i\right), \quad \mathsf{H} = \sum\left(x_i \; \mathit{ln}\frac{1}{x_i} + x_i\right),$$

$$\mathsf{D} = \sum x_i \; \mathit{ln}\frac{x_i}{y_i}\,.$$

Associated problem: MaxEnt! (model selection!)

Q “adjusted “ Φ and H? Then $\boxed{\text{adjusted entropy} \geq 1!?}$

Example (change basic model to one with prior)

$$\Phi = \Phi(x,y) - \Phi(x,y_0)\,, \quad \mathsf{H} = -\,\mathsf{D}(x,y_0)$$
$$\mathsf{D} = \mathsf{D}(x,y)\,.$$

Then $-\Phi$ is updating gain. This is of Shannon type if D= KL-divergence. OBS: Only depends on D. Associated optimization problem: MinDiv! (problem of updating).

This is of non-Shannon type if D=squared Euclidean distance. Then $\Phi = \|x-y\|^2 - \|x-y_0\|^2$, $\mathsf{H} = -\|x-y_0\|^2$ and $\mathsf{D} = \|x-y\|^2$.

In all examples strategy sets $X = Y$ are involved, the identity $x \curvearrowright \widehat{x}$ gives the response (in more general modeling, $X \neq Y$ is allowed):

Axiom 1 Linking: $\Phi(x,y) = \mathsf{H}(x) + \mathsf{D}(x,y)$ with $\mathsf{D} \geq 0$ and $\mathsf{D}(x,y) = 0 \Leftrightarrow y = \widehat{x}\,.$

This is the basic axiom. Invites for two-person zero-sum games $\gamma(X_0)$ with objective function Φ and preparation $X_0 \subseteq X$.
Philosophy, features in brief:

X_0 is the strategy set for Player I, nature ,
Y is the strategy set for Player II, you!
Player-I value $=$ MaxEnt-value , $\sup_{x \in X_0} \inf_{y \in Y} \Phi(x, y)$
$= \sup_{x \in X_0} \mathrm{H}(x) = \mathrm{H}_{\max}(X_0)$,
Player-II value $=$ MinRisc-value ,
$\mathrm{R}_{\min}(X_0) = \inf_Y \sup_{X_0}$-value,
$\mathrm{H}_{\max}(X_0)$ always $\leq \mathrm{R}_{\min}(X_0)$, equilibrium if equal,
Nash conditions (saddlevalue inequalities), a key tool.

Preparations, exponential families, $\mathcal{E}(\cdot)$ and optimal strategies (x^*, y^*).
Basic case:
Given X_0, associated exponential family defined as:
$\mathcal{E}(X_0) = \{y^* | \exists h \forall x \in X_0 : \Phi(x, y^*) = h\}$.

$\boxed{(x^*, y^*) \text{ optimal if } x^* \in X_0,\ y^* \in \mathcal{E}(X_0) \text{ and } y^* = \widehat{x^*}}$

Advanced: Connected with Q: what *can* we know? Natural preparations (genus-1 case) are the level sets, sets $\neq \emptyset$ of the form $L^\eta(h) = \{x|\Phi(x,\eta) = h\}$ for $\eta \in Y$ and h a constant. Define $\mathcal{E}(\eta) = \bigcap_h \mathcal{E}(L^\eta(h))$.

(x^*, y^*) optimal for $\gamma(L^\eta(h))$ if: $x^* \in L^\eta(h)$, $y^* = \widehat{x^*}$ and $y^* \in \mathcal{E}$. Pythagorean inequalities hold.

Illustrative "beyond Shannon" example.
$\Phi = \|x-y\|^2 - \|x-y_0\|^2$. Fix η. Then $\mathcal{E}(\eta)$ consists of hyperplanes with $y_0 - \eta$ as normal. And updating reduces to standard projection. This and Shannon examples satisfy axiom of affinity:

Axiom 2 X is convex and Φ affine in its first variable: For $y \in Y$, α molecular probability measure over X,

$$\Phi\Big(\sum_{x\in X} \alpha_x x, y\Big) = \sum_{x\in X} \alpha_x \Phi(x,y)\,.$$

Leads to important concavity- and convexity results for entropy, information transmission and divergence.

Part II: Special entropy functions

Think as a physicist, planning experiments:

1: I focus on Truth, belief and experience on the way to information .
I seek the truth, am restricted by my beliefs and will know by experience through the data how truth manifests itself to me.
I ask *why should not what I see in terms of data depend not only on truth but also on belief?* I assume $z = \Pi(x, y)$. Here, x, y and z are truth-, belief- and data instances , objects associated with any particular situation I may be interested in. Π is the global interactor. It is a characteristic of the world of which I am a part.

Examples: The classical or Shannon world is characterized by $\Pi(x, y) = x$.
A black hole is characterized by $\Pi(x, y) = y$. In such a world, I can only get out what I myself put in.

$\mathbb{A}$	Truth (x)	Belief (y)	Experience (z)
$\cdot$	$\cdot$	$\cdot$	$\cdot$
i	x_i	y_i	z_i
$\cdot$	$\cdot$	$\cdot$	$\cdot$

2: Will focus on concepts which are independent of semantic content. Therefore, I apply probabilistic reasoning across semantic differences. This will also enable quantitative reasoning. Thus, instances x, y and z in a specific situation will be probability vectors $(x_i)_{i\in\mathbb{A}}$, $(y_i)_{i\in\mathbb{A}}$ and $(z_i)_{i\in\mathbb{A}}$ over the alphabet $\mathbb{A} = \{i|\cdots\}$ with i's representing basic events. I assume that the global interactor acts locally, i.e. $\Pi(x,y) = (\pi(x_i,y_i))_{i\in\mathbb{A}}$ for some real valued function π defined on $[0,1] \times [0,1]$. This function is the local interactor or just the interactor.

In Shannon world: $\pi(x,y) = x$.
In black hole: $\pi(x,y) = y$.

3: The interactor must be sound: $\pi(x,x) = x$ for $x \in [0,1]$. I assume it is even consistent, i.e. $\sum_{i\in\mathbb{A}} z_i = 1$ with $z_i = \pi(x_i,y_i)$ for all probability vectors x and y.

4: Any event I may observe entails a certain effort $\kappa(y_i)$ which only depends on the belief-value. $\kappa : y \curvearrowright \kappa(y)$ is the descriptor. Clearly, $\kappa(1) = 0$ and as normalization condition I take $\kappa'(1) = -1$.

5: Description cost, denoted Φ, is the total effort taking into account the weights with which I will experience the various basic events:

$$\Phi(x,y) = \sum_{i\in\mathbb{A}} \pi(x_i,y_i)\kappa(y_i)\,. \qquad (1)$$

6: I will minimize description cost and appeal to the variational principle that the smallest value is obtained when there is a perfect match between truth and belief, i.e. when $y = x$. This is the perfect match principle. The quantity

$$\sum_{i \in \mathbb{A}} \pi(x_i, y_i)\kappa(y_i) - \sum_{i \in \mathbb{A}} x_i \kappa(x_i) \tag{2}$$

represents my frustration, as it compares the actual description cost with the smallest possible cost, had I only known the truth. The perfect match principle says that frustration disappears, when $y = x$.
Theoretically, if I knew $x = (x_i)_{i \in \mathbb{A}}$, minimal description cost is what I aim at, I call it entropy:

$$\mathsf{H}(x) = \inf_{y=(y_i)_{i \in \mathbb{A}}} \Phi(x,y) = \sum_{i \in \mathbb{A}} x_i \kappa(x_i)\,.^{a} \tag{3}$$

The quantity (2) I call divergence:

$$\mathsf{D}(x,y) = \Phi(x,y) - \mathsf{H}(x)\,. \tag{4}$$

[a] to allow a singular case, the infimum should be restricted to run over probability distributions y with a support which contains the support of x.

Theorem Assuming consistency and suitable regularity conditions, $q = \pi(1,0) \geq 0$. To each $q \in [0, \infty[$, there is only one interactor and one descriptor which fulfill the conditions imposed. These functions, π_q and κ_q, are determined by

$$\pi_q(x,y) = qx + (1-q)y\,, \tag{5}$$

$$\kappa_q(y) = \textit{ln}\,_q\frac{1}{y}\,, \tag{6}$$

where the q-*logarithm* is given by

$$\textit{ln}\,_q x = \begin{cases} \textit{ln}\, x \text{ if } q = 1, \\ \frac{x^{1-q}-1}{1-q} \text{ if } q \neq 1\,. \end{cases} \tag{7}$$

Outline of proof: (5) follows by consistency. Then, (6), follows from variational principle via technique with Lagrange multipliers, which leads you to the differential equation

$$(1-q)\kappa(x) + x\kappa'(x) = -1\,. \tag{8}$$

Final step: To show that with (5) and (6) the perfect match principle holds, follows from (14) below.

The accompanying quantities, description cost, entropy and divergence are denoted Φ_q, H_q and D_q, respectively. We find

$$\Phi_q = \sum_{i\in\mathbb{A}} \pi_q(x_i,y_i)\kappa_q(y_i)\,, \tag{9}$$

$$\mathsf{H}_q = \sum_{i\in\mathbb{A}} x_i\kappa_q(x_i)\,, \tag{10}$$

$$\mathsf{D}_q = \sum_{i\in\mathbb{A}} \Big(\pi_q(x_i,y_i)\kappa_q(y_i) - x_i\kappa_q(x_i)\Big)\,. \tag{11}$$

Shannon type quantities for $q=1$: Kerridge inaccuracy, Shannon entropy and Kullback-Leibler divergence . For $q \neq 1$:

$$\Phi_q = \sum_{i\in\mathbb{A}} \left(\frac{q}{1-q}x_i y_i^{q-1} + y_i^q - \frac{1}{1-q}x_i\right), \tag{12}$$

$$\mathsf{H}_q = \frac{1}{1-q}\sum_{i\in\mathbb{A}}(x_i^q - x_i) = \frac{1}{1-q}\sum_{i\in\mathbb{A}} x_i^q - 1\,, \tag{13}$$

$$\mathsf{D}_q = \sum_{i\in\mathbb{A}} \left(\frac{q}{1-q}x_i y_i^{q-1} + y_i^q - \frac{1}{1-q}x_i^q\right). \tag{14}$$

In (12) the linearity in x is evident. This is important as it leads to a relatively easy approach to key optimization problems. In (13) we recognize the family of Tsallis entropies. For $q = 0$ (black hole), $\mathrm{H}_0(x) = n - 1$ ($n =$ size of support of x).

In (14) the summands are non-negative. This can be exploited to give an easy proof of the "q-version" of the fundamental inequality of information theory: $\mathrm{D}_q(x, y \geq 0$ with equality if and only if $x = y$. This is valid for any $q > 0$. Note the pointwise version of the fundamental inequality:
$\boxed{\pi(x,y)\kappa(y) + y \geq x\kappa(x) + x.}$ For $q = 0$, one finds that $\mathrm{D}_0 \equiv 0$. (14)also points to possible extensions to continuous distributions.

The general formulas (1), (3) and (4) indicate that for the determination of the quantities involved one needs to know the interactor π as well as the descriptor κ. Two facts should be emphasized. Firstly, through the perfect match principle, the descriptor is uniquely determined from the interactor. Therefore, in principle, only the interactor needs to be known. Secondly, different interactors may well determine the same descriptor. Thus, knowing only the descriptor, you cannot determine divergence or description cost. But you *can* determine the entropy function.

Outstanding questions: Physical mechanisms behind interaction, coding interpretations of description cost.

Hints to the literature

J. Havrda and F. Charvát, 1967.
J. Lindhard and V. Nielsen, 1971.
J. Lindhard, 1974.
J. Naudts, 2008.
F. Topsøe, 2007.
C. Tsallis, 1988.