On the generation of measures of entropy, divergence and complexity

[Question:

Complexity := Description cost ?]

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences

Presentation at the Entropy workshop in Lausanne,
September 8-9, 2008

Classical Information Theory: Complexity, entropy and divergence: **either**

$$\Phi(x,y) = \sum x_i \, \textit{ln} \, \frac{1}{y_i}\,,$$
$$\mathsf{H}(x) = \sum x_i \, \textit{ln} \, \frac{1}{x_i}\,,$$
$$\mathsf{D}(x,y) = \sum x_i \, \textit{ln} \, \frac{x_i}{y_i}\,.$$

over $X = Y = M_+^1(\mathbb{A})$ **or (often better!)**

$$\Phi(x,y) = \sum x_i y_i \,,$$
$$\mathsf{H}(x) = \Phi(x,\widehat{x})\,,$$
$$\mathsf{D}(x,y) = \Phi(x,y) - \mathsf{H}(x)\,.$$

over $X = M_+^1(\mathbb{A})$, $Y = K(\mathbb{A})$ and with response $x \curvearrowright \widehat{x} = y$ defined by $y_i = \textit{ln}\,\frac{1}{x_i}$ where $K(\mathbb{A})$ is the set of code length functions over $\mathbb{A}$, functions y satisfying Kraft's inequality $\sum e^{-y_i} \leq 1$.

x's: "truth" ; y's: Belief, expectation, descriptor...

Axioms for Complexity, entropy, divergence.

Strategy sets are X, Y, a map $x \curvearrowright \widehat{x}$ of X into Y gives the response.
$\mathsf{MOL}(X) = \{$molecular measures$\}$
$= \{\alpha \in M_+^1(X) | \,\mathsf{supp}(\alpha)$ finite $\}$.

Axiom 1 Linking: $\Phi(x,y) = \mathsf{H}(x) + \mathsf{D}(x,y)$ with $\mathsf{D} \geq 0$ and $\mathsf{D}(x,y) = 0 \Leftrightarrow y = \widehat{x}$.
Axiom 2 Affinity: X is convex and Φ affine in its first variable: For $y \in Y$, $\alpha \in \mathsf{MOL}(X)$,

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

First consequences: Introduce barycentre $b(\alpha) = \sum_{x \in X} \alpha_x x$, and associated information rate

$$\mathrm{I}(\alpha) = \sum_{x \in X} \alpha_x \,\mathsf{D}(x, \widehat{b(\alpha)})\,.$$

Concavity and convexity properties:

Let $\alpha \in \mathsf{MOL}(X)$. Then

$$\mathsf{H}\Big(\sum_{x\in X}\alpha_x x\Big) = \sum_{x\in X}\alpha_x\,\mathsf{H}(x) + \mathrm{I}(\alpha)$$

and, if $\mathsf{H}(b(\alpha)) < \infty$, then, for every $y \in Y$,

$$\sum_{x\in X}\alpha_x\,\mathsf{D}(x,y) = \mathsf{D}\Big(\sum_{x\in X}\alpha_x x, y\Big) + \mathrm{I}(\alpha) \quad (*)$$

(*) is the compensation identity. Only depends on D!

Special case of information rate gives Jensen-Shannon divergence:

$\mathsf{JSD}(x_1,x_2) = \frac{1}{2}\mathsf{D}(x_1,\hat{x}) + \frac{1}{2}\mathsf{D}(x_2,\hat{x})$ with $x = \frac{1}{2}x_1 + \frac{1}{2}x_2$. Often defines the square of a metric!

Problems/ opportunities

1. good examples (+proofs!) and counterexamples
2. isometrically embeddable in Hilbert space?
3. new non-standard entropy inequalities!
4. quantum case?

Proposition JSD is the square of a metric if and only if, for every x_1, x_2, x_3

$$\sum_{k=1}^{3} \Big([ij]^2 - 2[ik][jk] + 2[ij][k] - [i][j] \Big) \leq 0$$

where

$$[ij] = \mathsf{H}(\frac{1}{2}x_i + \frac{1}{2}x_j) \text{ and } [i] = [ii] = \mathsf{H}(x_i)\,.$$

Models and exponential families For $X_0 \subseteq X$, $\gamma_\Phi(X_0)$ denotes two-person zero-sum game over $X_0 \times Y$ with Φ as objective function, Player I as maximizer and Pl. II as minimizer. Write $\gamma_\Phi(X_0) \in \mathsf{GTE}(x, y)$ if γ_Φ is in equilibrium with (x, y) as optimal strategies.
From Nash's saddle-value theorem:

> **Theorem** A given pair (x_0, y_0) is an optimal pair for a subgame in equilibrium *iff* $\Phi(x_0, y_0) \in \mathbb{R}$ and $y_0 = \hat{x}$. If so, the possible models are all X_0 with $\{x_0\} \subseteq X_0 \subseteq \{\Phi^{y_0} \leq h\}$ with $h = \Phi(x_0, y_0)$.

Natural models (genus-1 case): are the non-empty level-sets: $\mathsf{L}^f(h) = \{\Phi^f = h\} = \{x | \Phi(x, f) = h\}$
Let $\mathcal{L}^f =$ class of non-empty models of the form $L^f(h)$.
The associated exponential family is the family
$\mathcal{E}(f) = \{y | \forall L \in \mathcal{L}^f \exists c \in \mathbb{R} : L \subseteq \mathsf{L}^y(h)\}$.

> $y \in \mathcal{E}(f)\,,\ y = \hat{x} \Rightarrow L^f(\Phi^y(x)) \in \mathsf{GTE}(x, y)$

Problems: Generalized notions needed, relation to standard theory, to weaker notions of equilibrium etc.

Reminder: Games, some general considerations

$\Phi: X \times Y \to \overline{\mathbb{R}}$ defines a two-person zero-sum game, γ_Φ. It has Φ as objective function (complexity!).
Player I, a maximizer, chooses $x \in X$,
Player II, a minimizer, chooses $y \in Y$.
Specific and global values:

$$\mathsf{val}_{\mathrm{I}}(x) = \inf_{y \in Y} \Phi(x,y) = \inf \Phi_x \Big(\text{entropy! } \mathsf{H}(x) \Big)$$

$$\mathsf{val}_{\mathrm{II}}(y) = \sup_{x \in X} \Phi(x,y) = \sup \Phi^y \Big(\text{risk! } \mathsf{R}(y) \Big)$$

$$\mathsf{val}_{\mathrm{I}} = \sup_{x \in X} \mathsf{val}_{\mathrm{I}}(x)\,, \quad \mathsf{val}_{\mathrm{II}} = \inf_{y \in Y} \mathsf{val}_{\mathrm{II}}(y)\,.$$

y is an optimal response to x (or x matches y) if $y \in \widehat{x} = \mathrm{argmin}\Phi_x$.

Redundancy: Compare the potentially possible with the actually achieved to obtain Player-I redundancy and Player-II redundancy:

$$\delta_{\mathrm{I}}(x,y) = \mathsf{val}_{\mathrm{II}}(y) - \Phi(x,y)\,,$$
$$\delta_{\mathrm{II}}(x,y) = \Phi(x,y) - \mathsf{val}_{\mathrm{I}}(x) \Big(\text{divergence! } \mathsf{D}(x,y)\Big).$$

With span$(x, y) = \text{val}_{\text{II}}(y) - \text{val}_{\text{I}}(x)$,
span$(x, y) = \delta_{\text{I}}(x, y) + \delta_{\text{II}}(x, y)$, hence:
$\text{val}_{\text{I}} \leq \text{val}_{\text{II}}$ (minimax inequality).

Game Theoretical Equilibrium: if $\text{val}_{\text{I}} = \text{val}_{\text{II}} \in \mathbb{R}$.
Ideally: GTE applies and optimal strategies exist, say (x_0, y_0). Notation: $\gamma_\Phi \in \text{GTE}(x_0, y_0)$.

Saddle-value theorem (Nash): Assume that $\Phi(x_0, y_0) \in \mathbb{R}$. Then $\gamma_\Phi \in \text{GTE}(x_0, y_0)$ *iff*
$\forall (x, y): \Phi(x, y_0) \leq \Phi(x_0, y_0) \leq \Phi(x_0, y)$.
(FT): If so, abstract pythagorean inequalities hold:
$\forall x: \text{val}_{\text{I}}(x) + \delta(x, y_0) \leq \text{val}(\gamma_\Phi)$ (forward ineq.),
$\forall y: \text{val}(\gamma_\Phi) + \delta(x_0, y) \leq \text{val}_{\text{II}}(y)$ (backward ineq.).
Here, $\delta = \delta_{\text{I}}$, δ_{II} or even $\delta_{\text{I}} + \delta_{\text{II}}$. [symmetry!]

Proof: With $\delta = \delta_{\text{I}} + \delta_{\text{II}}$, the inequalities become identities! $\square$

Corollary: Assume that $\Phi(x_0, y_0) \in \mathbb{R}$. Then, if y_0 is an optimal response to x_0 and if $\Phi(x, y_0)$ is independent of $x \in X$, $\gamma_\Phi \in \text{GTE}(x_0, y_0)$. [asymmetry!]

Creation of Information Triples

Atomic Triples, Integration

$(\phi, \mathsf{h}, \mathsf{d})$ with $X = Y$ =real interval, and response the identity leads to atomic information triples.

Example 1 y_0 a prior,

$$\phi(x,y) = (x-y)^2 - (x-y_0)^2\,,$$
$$\mathsf{h}(x) = -(x-y_0)^2\,,$$
$$\mathsf{d}(x,y) = (x-y)^2\,.$$

Example 2

$$\phi(x,y) = x\;\mathit{ln}\,\frac{1}{y}\,,$$
$$\mathsf{h}(x) = x\;\mathit{ln}\,\frac{1}{x}\,,$$
$$\mathsf{d}(x,y) = x\;\mathit{ln}\,\frac{x}{y}\,.$$

Examples are of Bregman type: for "smooth" strictly concave h, $(\phi, \mathsf{h}, \mathsf{d})$ with ϕ and d defined by

$$\phi(x,y) = \mathsf{h}(y) + (x-y)\,\mathsf{h}'(y)\,,$$
$$\mathsf{d}(x,y) = \mathsf{h}(y) - \mathsf{h}(x) + (x-y)\,\mathsf{h}'(y)\,,$$

is an atomic information triple.

A natural process of integration leads to more general triples. Given measure μ on set T and then some function space $X \subseteq I^T$, take identity as response and define $(\Phi, \mathsf{H}, \mathsf{D})$ by integration, i.e.

$$\Phi(x,y) = \int_T \phi\Big(x(t), y(t)\Big) d\mu(t)$$

and similarly for H and D. ...

By integration, Example 1 extends to a triple over Hilbert space:

$$\Phi(x,y) = \|x-y\|^2 - \|x-y_0\|^2\,,$$
$$\mathsf{H}(x) = -\|x-y_0\|^2\,,$$
$$\mathsf{D}(x,y) = \|x-y\|^2\,.$$

And similarly, Example 2 leads to standard discrete information theory by integration w.r.t. counting measure over an "alphabet".

Equivalence, Relativization

Equivalence results from adding to both Φ and to H an affine function defined on X

If $(\Phi, \mathsf{H}, \mathsf{D})$ is given and you add $x \curvearrowright -\Phi(x, y_0)$, you obtain the relativized triple with y_0 as prior :

$$\begin{aligned} \tilde{\Phi}(x,y) &= \mathsf{D}(x,y) - \mathsf{D}(x,y_0) \\ \tilde{\mathsf{H}}(x) &= -\,\mathsf{D}(x,y_0) \\ \tilde{\mathsf{D}}(x,y) &= D(x,y)\,. \end{aligned}$$

(for this, it suffices that D satisfies the compensation identity). Leads to Kullback's minimum information discrimination principle , related to the problem of proper updating .

Randomization

Start with $(\Phi, \mathsf{H}, \mathsf{D})$. Allow randomized strategies $\alpha \in MOL(X)$ for Player I. Put $b(\alpha) = \sum_{x \in X} \alpha_x x$. Randomization then gives:

$$\begin{aligned}
\widehat{\alpha} &= \widehat{b(\alpha)}\,, \\
\tilde{\Phi}(\alpha, y) &= \sum_{x \in X} \alpha_x \Phi(x, y)\,, \\
\tilde{\mathsf{H}}(\alpha) &= \sum_{x \in X} \alpha_x \Phi\Big(x, \widehat{b(\alpha)}\Big)\,, \\
\tilde{\mathsf{D}}(\alpha, y) &= \mathsf{D}(b(\alpha), y)\,.
\end{aligned}$$

By equivalence you obtain:

$$\begin{aligned}
\tilde{\Phi}_0(\alpha, y) &= \sum_{x \in X} \alpha_x\, \mathsf{D}(x, y)\,, \\
\tilde{\mathsf{H}}_0(\alpha) &= \sum_{x \in X} \alpha_x\, \mathsf{D}(x, \widehat{b(\alpha)})\,, \\
\tilde{\mathsf{D}}_0(\alpha, y) &= \mathsf{D}(b(\alpha), y)\,.
\end{aligned}$$

Singling out special entropy functions

Put yourself in the shoes of the physicist who is planning observations and see if you can accept the considerations below.

1 Events have three kinds of assignments, related to, respectively, truth, belief and experience. Truth- and belief assignments are numbers in $[0, 1]$.

2 A characteristic feature of my world is that there is an interaction between truth and belief expressed by a function π on $[0, 1] \times [0, 1]$. The idea is (see table!) that $\pi_i = \pi(x_i, y_i)$.

$\mathbb{A}$	Truth	Belief	Experience
.	.	.	.
.	.	.	.
.	.	.	.
i	x_i	y_i	π_i
.	.	.	.
.	.	.	.
.	.	.	.

Example A: The classical world is a world of "no interaction", hence $\pi(x, y) = x$.

Example B: The black hole is a world of "no information", hence $\pi(x, y) = y$.

3 I believe that my world is consistent in the sense that $\sum_{i\in\mathbb{A}} \pi_i = 1$ whenever $(x_i)_{i\in\mathbb{A}}$ and $(y_i)_{i\in\mathbb{A}}$ are probability assignments and $\pi_i = \pi(x_i, y_i)$.

Note: Then interaction must be sound, i.e. a perfect match gives no change: For all $x \in [0, 1]$, $\pi(x, x) = x$.

4 Any event I may observe entails a certain effort on my part. The effort must only depend on my belief, y, and is denoted by $\kappa(y)$. The function κ, is the coder (or descriptor). Of course: $\kappa(1) = 0$.

5 Separability applies: My total effort related to observations from a particular situation is the sum of individual contributions. Weights must be assigned to each contribution according to the weight with which I will experience the various events. The total effort is the complexity (or description cost), Φ. Thus:

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

with $x = (x_i)_{i \in \mathbb{A}}$ the truth- and $y = (y_i)_{i \in \mathbb{A}}$ the belief-assignments associated with the events.

6 I will attempt to minimize complexity and shall appeal to the principle that complexity is the smallest when belief matches truth, ($(y_i)_{i\in\mathbb{A}} = (x_i)_{i\in\mathbb{A}}$). As

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

represents my frustration, the principle says that frustration is the least, in fact disappears, when $(y_i)_{i\in\mathbb{A}} = (x_i)_{i\in\mathbb{A}}$.

Note: Given $x = (x_i)_{i\in\mathbb{A}}$, minimal complexity is what I am aiming at. It is an important quantity. I will call it entropy:

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

Frustration too looks important. Perhaps I better call it divergence:

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

Can you accept all this? If so, you can conclude:

Theorem: Modulo regularity conditions and a condition of normalization, $q = \pi(1, 0)$ must be nonnegative and π and κ uniquely determined from q by:

$$\pi(x, y) = qx + (1 - q)y\,,$$
$$\kappa(y) = \mathit{ln}_q \frac{1}{y}\,,$$

where the q-logarithm is given by

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

Hence entropy is given by

$$\mathrm{H}(x) = \sum_{i \in \mathbb{A}} x_i \, \mathit{ln}_q \frac{1}{x_i}\,.$$

Challenges:

- explain interaction on physical grounds,
- suggest possibilities for an accompanying process of real coding,
- illuminate the good sense (if any :-)) of the views put forward in well studied concrete cases (possibly distinguishing between the cases $0 < q < 1$, $1 < q \leq 2$ and $q > 2$).

Let us look into the following:

- proof of theorem
- connection with Bregman generation
- relaxing the condition of consistency.

Indication of proof of main result

Functions π and κ are assumed continuous on their domains and continuously differentiable and finite valued on the interiors of their domains. Normalization of κ means that $\kappa(1) = 0$ and that $\kappa'(1) = -1$.

You can exploit the consistency condition to show that, for all $(x, y) \in [0, 1]^2$,

$$\pi(x, y) = qx + (1 - q)y$$

with $q = \pi(1, 0)$.

Consider a fixed finite probability vector $(x_i)_{i \in \mathbb{A}}$ with all x_i positive. Varying $(y_i)_{i \in \mathbb{A}}$ we find, via the introduction of a Lagrange multiplier, that f given by

$$f(x) = \frac{\partial \pi}{\partial y}(x, x)\kappa(x) + \pi(x, x)\kappa'(x)$$

is constant on $\{x_i | i \in \mathbb{A}\}$. Exploiting this for three-element alphabets $\mathbb{A}$ shows that $f \equiv -1$. Then the formula for κ is readily derived.

Bregman generation: Look at concave generator h_q and associated "Bregman quantities":

$$\begin{cases} h_q(x) = x \, \mathit{ln}_q \frac{1}{x} \, , \\ \phi_q(x,y) = h_q(y) + (x-y)h'_q(y) \, , \\ d_q(x,y) = h_q(y) - h_q(x) + (x-y)h'_q(y) \, , \\ \Phi_q(P,Q) = \sum_{a\in\mathbb{A}} \phi_q(p_i,q_i) \, , \\ \mathsf{H}_q(P) = \sum_{a\in\mathbb{A}} h_q(p_i) \, , \\ \mathsf{D}_q(P,Q) = \sum_{a\in\mathbb{A}} d_q(p_i,q_i) \, . \end{cases}$$

-compare with "interaction quantities":

$$\begin{cases} \pi_q(x,y) = qx + (1-q)y \text{ (interaction)} \, , \\ \kappa_q(x) = \mathit{ln}_q \frac{1}{x} \text{ (coder)} \, , \\ \xi(x,y) = y - x \, , \text{ (corrector)} \, , \\ \Phi_q(P,Q) = \sum_{a\in\mathbb{A}} \pi_q(p_i,q_i)\kappa_q(q_i) \\ \quad = \sum_{a\in\mathbb{A}} \Big(\pi_q(p_i,q_i)\kappa_q(q_i) + \xi(p_i,q_i) \Big) \, , \\ \mathsf{H}_q(P) = \sum_{a\in\mathbb{A}} p_i\kappa_q(p_i) \, , \\ \mathsf{D}_q(P,Q) = \sum_{a\in\mathbb{A}} \Big(\pi_q(p_i,q_i)\kappa_q(q_i) - p_i\kappa_q(p_i) \Big) \\ \quad = \sum_{a\in\mathbb{A}} \Big(\pi_q(p_i,q_i)\kappa_q(q_i) - p_i\kappa_q(p_i) + \xi(p_i,q_i) \Big) \, . \end{cases}$$

Here, ξ is the corrector introduced so that the Bregman- and interaction- quantities are synchronized. Indeed, then the individual quantities coincide, in particular,

$$\pi_q(p_i, q_i)\kappa_q(q_i) + \xi(p_i, q_i) = \phi_q(p_i, q_i)\,.$$

Note that the corrector is independent of q. When seeking further physically founded explanations for the whole set-up it may well be important to take the corrector into account.

Quantities written out:

$$\Phi(P,Q) = \frac{1}{1-q}\Big(-1 + \sum_{i\in\mathbb{A}}\big(qp_iq_i^{q-1} + (1-q)q_i^q\big)\Big)\,,$$

$$\mathsf{H}(P) = \frac{1}{1-q}\Big(-1 + \sum_{i\in\mathbb{A}} p_i^q\Big)\,,$$

$$\mathsf{D}(P,Q) = \frac{1}{1-q}\sum_{i\in\mathbb{A}}\Big(qp_iq_i^{q-1} - p_i^q + (1-q)q_i^q\Big)\,.$$

Relaxing the condition of consistence: If we only assume that π is sound, i.e. that $\pi(x,x) = x$ for $0 \leq x \leq 1$, then other forms of interaction may leed to Tsallis-entropy as well. This happens with

$$\pi(x,y) = x^q y^{1-q}\,.$$

Thus, many quite different forms of interaction may give the same entropy function. But of course, the complexity- and divergence-functions will be different.

References in brief:

- Havrda and Charvát (1967): first appearence in the mathematical literature
- Lindhard and Nielsen (1971) and Lindhard (1974): first appearence in the physical literature
- Tsallis (1988): well known (:-)) take-off point which triggered much research and debate.

As recent contributions relevant for the present research, I mention Naudts (2008) and my own contribution from (2007).