Information and Games

Flemming Topsøe (topsoe@math.ku.dk)
University of Copenhagen
Department of Mathematical Sciences
"Facets of Entropy" Copenhagen, October 24-26, 2007

based on
http://www.math.ku.dk/∼topsoe/jogo2.pdf and
http://www.math.ku.dk/∼topsoe/CTnexttopsoe.pdf
Overview: existence and identification

****

So said Kolmogorov (1983):

"*Information theory*
must
precede probability theory
and not be based on it"

Let us follow him!

Nature	Observer
holds the truth!	seeks the truth
does not have a mind!	has a mind
an absolute	YOU!
passive	inventive

In so doing, I reveal nature of entropy and divergence:

- Entropy is minimal complexity
- Divergence measures discrepancy between actual and minimal complexity

– and what then is complexity ?

- Complexity is whatever complexity should be in any concrete situation.

– hmmm. So axiomatize based on game theoretical thinking! — but why games?

- because they provide a perfect setting for the modeling of conflict situations

Is that what information is about? yes(!?)

Which tools?

- the simplest are the most important: two-person zero-sum games!

Modeling of un-symmetric conflict situations between players, one having a mind, the other not - with focus: complexity (Φ), entropy (H) and divergence (D)

or: pay-off (Ψ), maximal pay-off (Π), divergence (D)

- strategy sets X and Y (e.g. X=Y),
- connection $X \to Y$ written $x \curvearrowright \widehat{x}$ (e.g. $\widehat{x} = x$),
- complexity $\Phi : X \times Y \to]-\infty, \infty]$ (or $[0, \infty]$),
- entropy $\mathsf{H} : X \to]-\infty, \infty]$ (or $[0, \infty]$),
- divergence $\mathsf{D} : X \times Y \to [0, \infty]$
- and possibly a preparation $X_0 \subseteq X$ (or more!)

Axiom 1 (linking): For $(x, y) \in X \times Y$,
$\Phi(x, y) = \mathsf{H}(x) + \mathsf{D}(x, y)$ (linking identity) and
$\mathsf{D}(x, y) = 0 \Leftrightarrow y = \widehat{x}$

Axiom 2 (affinity): X is convex and Φ affine in first variable: $\Phi(\sum a_\nu x_\nu, y) = \sum a_\nu \Phi(x_\nu, y)$

Axiom 3 (semi-continuity): reference topology (X, τ) is Hausdorff, algebraic operations continuous and, for any (x_0, y_0), $x \curvearrowright \mathsf{D}(x, y_0)$ and $x \curvearrowright \mathsf{D}(x_0, \widehat{x})$ are τ-lower semi-continuous.

Axiom 4 (weak completeness): For a sequence (x_n) in X, put $x_{n,m} = \frac{1}{2}x_n + \frac{1}{2}x_m$ and $y_{n,m} = \widehat{x_{n,m}}$. If "Cauchy property" $\mathsf{D}(x_n, y_{n,m}) \to 0$ as $n, m \to \infty$ then $\exists x$, $(x_{n_k})_{k \geq 1}$ such that $x_{n_k} \to x$

strong convergence: $x_n \twoheadrightarrow x$ if $\mathsf{D}(x_n, \widehat{x}) \to 0$

Sometimes pay-off Ψ, maximal pay-off Π and Divergence D as before are more natural to work with.
By "duality" you pass from the one system to the other: $(\Phi, \mathrm{H}, \mathrm{D}) \leftrightarrow (\Psi, \Pi, \mathrm{D})$ with $\Phi = -\Psi$, $\mathrm{H} = -\Pi$.
We talk about information triples based on complexity and information triples based on pay-off.

Example 1(classical information theory) $\mathbb{A}$ a discrete alphabet, $X = M_+^1(\mathbb{A})$, set of probability distributions over $\mathbb{A}$, $Y = K(\mathbb{A})$, set of code length functions over $\mathbb{A}$, i.e. set of $\kappa : \mathbb{A} \to [0, \infty]$ such that Krafts equality holds: $\sum_{i \in \mathbb{A}} \exp(-\kappa_i) = 1$. Let $P \curvearrowright \hat{P}$ be the bijection $P \leftrightarrow \kappa$ with $\kappa_i = \ln \frac{1}{p_i}$, $p_i = \exp(-\kappa_i)$. With Φ as average code length and with

$$\Phi(P, \kappa) = \langle \kappa, P \rangle = \sum_{i \in \mathbb{A}} p_i \kappa_i$$

$$\mathrm{H}(P) = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}$$

$$\mathrm{D}(P, \kappa) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \text{ (here } Q \leftrightarrow \kappa) \qquad \square$$

Example (geometric version of updating)

$X = Y$, a Hilbert space, with identity as connection. Let y_0 be a point in Y, the prior. With

$$\begin{aligned} \Psi(x,y) &= \|x-y_0\|^2 - \|x-y\|^2 \\ \Pi(x) &= \|x-y_0\|^2 \\ \mathrm{D}(x,y) &= \|x-y\|^2, \end{aligned}$$

Axioms 1-4 hold. □

For standard updating you will replace $\|x-y\|^2$ by a standard divergence, say Kullback-Leibler divergence. The reason why seemingly different objects can be used for the same task lies in the common identities fulfilled by these quantities. They involve entropy, divergence and information rate $\mathrm{I}(\cdot)$ (really the same as mutual information) and hold whenever Axioms 1 and 2 hold. Let $\mathrm{MOL}(X)$ be the set of probability distributions over X with finite support. Then, for $\alpha \in \mathrm{MOL}(X)$, define

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

Theorem (identities re concavity- and convexity)

(i) Let $\overline{x} = \sum_{x\in X} \alpha_x x$ be a convex combination of elements in X corresponding to $\alpha \in \mathsf{MOL}(X)$. Then

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

(ii) With notation as in (i), assume that $\mathsf{H}(\overline{x}) < \infty$ and let $y \in Y$. Then

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

(iii) For elements $\alpha_1, \cdots, \alpha_m$ in $\mathsf{MOL}(X)$ with barycentres $\overline{x_1} \ldots, \overline{x_m}$, and for any mixture $\alpha = \sum_1^m w_k \alpha_k$ with a barycentre $\overline{x}$ of finite entropy, the following identity holds:

$$\mathrm{I}\Big(\sum_{k=1}^m w_k\alpha_k\Big) = \sum_{k=1}^m w_k \,\mathrm{I}(\alpha_k) + \sum_{k=1}^m w_k \,\mathsf{D}(\overline{x_k}, \overline{x})\,.$$

Proof: (i) is trivial by linking and (i) $\Rightarrow$ (ii) $\Rightarrow$. (iii) $\square$

Consider $(\Phi, \mathsf{H}, \mathsf{D})$ satisfying Axioms 1-4 and a preparation X_0. This defines two-person zero-sum game $\gamma(X_0)$ with Φ as objective function, Observer as minimizer and Nature as optimizer, but with restriction to strategies in X_0. Some important notions:

- $x \in X_0$: consistent strategy
- $\sup_{x\in X_0} \inf_{y\in Y} \Phi(x,y) = \sup_{x\in X_0} \mathsf{H}(x)$: maximum entropy value, $\mathsf{H}_{\max} = \mathsf{H}_{\max}(X_0)$
- $\inf_{y\in Y} \sup_{x\in X_0} \Phi(x,y) = \inf_{y\in Y} \mathsf{R}(y)$: minimal risk, $\mathsf{R}_{\min} = \mathsf{R}_{\min}(X_0)$.
- minimax inequality: $\mathsf{H}_{\max} \le \mathsf{R}_{\min}$.
- *equilibrium* means that $\mathsf{H}_{\max} = \mathsf{R}_{\min} < \infty$.
- x MaxEnt-strategy: consistent x with $\mathsf{H}(x) = \mathsf{H}_{\max}$.
- A sequence (x_n) of consistent strategies is asymptotically optimal if $\lim_{n\to\infty} \mathsf{H}(x_n) = \mathsf{H}_{\max}$.
- $x \in X$ is a $\mathsf{H}_{\max}$-attractor if $x_n \twoheadrightarrow x$ for every asymptotically optimal sequence (x_n).
- $y \in Y$ $\mathsf{R}_{\min}$-strategy if $\mathsf{R}(y) = \mathsf{R}_{\min}(X_0)$.
- (x^*, y^*) optimal pair: x^* MaxEnt-strat., y^* $\mathsf{R}_{\min}$-strat.

Main Theorem X_0 convex, $\mathsf{H}_{\mathsf{max}}(X_0) < \infty$. Then:

- Observer has unique optimal strategy y^*
- an $\mathsf{H}_{\mathsf{max}}$-attractor x^* exists and $y^* = \widehat{x^*}$.
- $\mathsf{H}_{\mathsf{max}}$-attractors are equivalent, hence unique if connection is injective
- the game is in equilibrium: $\mathsf{H}_{\mathsf{max}}(X_0) = \mathsf{R}_{\mathsf{min}}(X_0)$
- for $x \in X_0$, $y \in Y$ strong inequalities hold:

$$\mathsf{H}(x) + \mathsf{D}(x, y^*) \leq \mathsf{H}_{\mathsf{max}}(X_0)$$
$$\mathsf{R}_{\mathsf{min}}(X_0) + \mathsf{D}(x^*, y) \leq \mathsf{R}(y)\,.$$

Application to updating: Consider $(\Phi, \mathsf{H}, \mathsf{D})$ satisfying Axioms 1-4 with $x \curvearrowright \hat{x}$ injective and, to simplify, Φ finite. Let $X_0 \subseteq X$ be convex and let $y_0 \in Y$ be a "prior". Define:

$$\Psi(x, y) = \mathsf{D}(x, y_0) - \mathsf{D}(x, y)$$
$$\Pi(x) = \mathsf{D}(x, y_0)\,.$$

Consider corresponding game. For $y \in Y$, let

$$\Gamma(y) = \inf_{x \in X_0} \Psi(x, y)$$

guaranteed updating gain associated with y and let

$$\Gamma_{\max} = \sup_{y \in Y} \Gamma(y)\,,\ \mathrm{D}_{\min} = \inf_{x \in X_0} \mathrm{D}(x, y_0)$$

A strategy $x^* \in X$ is the generalized I-projection of y_0 on X_0 if $x_n \twoheadrightarrow x^*$ for every sequence (x_n) in X_0 which is asymptotically optimal in the sense that $\lim_{n\to\infty} \mathrm{D}(x, y_0) = \mathrm{D}_{\min}$.

Theorem The game is in equilibrium: $\Gamma_{\max} = \mathrm{D}_{\min}$. There is a unique generalized I-projection x^* of y_0 on X_0 and $y^* = \widehat{x^*}$ is the unique optimal strategy for Observer: $\Gamma(y^*) = \Gamma_{\max}$. Furthermore, for $(x, y) \in X_0 \times Y$,

$$\mathrm{D}(x, y_0) \geq \mathrm{D}_{\min} + \mathrm{D}(x, y^*)\,,$$
$$\Gamma(y) + \mathrm{D}(x^*, y) \leq \Gamma_{\max}\,.$$

This follows from applying main theorem to

$$(x, y) \curvearrowright -\Psi(x, y) = \mathrm{D}(x, y) - \mathrm{D}(x, y_0)$$
$$x \curvearrowright \mathrm{H}(x) - \Phi(x, y_0) = -\mathrm{D}(x, y_0)$$
$$(x, y) \curvearrowright \Phi(x, y) - \mathrm{H}(x) = \mathrm{D}(x, y)\,.$$

Theorem Let X_0 be a closed convex subset of the Hilbert space Y and let $y_0 \notin X_0$. Then there exists a hyperplane which separates y_0 from X_0.

Proof: Take (Ψ, Π, D) as before:

$$\begin{aligned} \Psi(x,y) &= \|x-y_0\|^2 - \|x-y\|^2 \\ \Pi(x) &= \|x-y_0\|^2 \\ D(x,y) &= \|x-y\|^2\,, \end{aligned}$$

Conclude from main Theorem that the associated game is in equilibrium. The value for Nature is $\inf_{x \in X_0} \|x-y_0\|^2$ which is positive by assumption. Therefore, the value of the game for Observer must also be positive, i.e.

$$\sup_{y \in Y} \inf_{x \in X_0} \left(\|x-y_0\|^2 - \|x-y\|^2 \right) > 0\,.$$

We conclude that there exists $y \in Y$ such that

$$\|x-y_0\| > \|x-y\| \text{ for all } x \in X_0\,.$$

This shows that the hyperplane of all x with the same distance to y as to y_0 separates y_0 and X_0. □

Change focus from existence to identification. Key to this: Nash equilibrium! Requires pair $(x^*, y^*) \in (X_0, Y)$ such that saddle value inequalities $\Phi(x, y^*) \leq \Phi(x^*, y^*) \leq \Phi(x^*, y)$ hold for all $(x, y) \in X_0 \times Y$. If also $\Phi(x^*, y^*) < \infty$, equilibrium follows with (x^*, y^*) a (MaxEnt,R_{min})-pair. With our special assumptions we find:

Theorem Let $x^* \in X_0$, put $y^* = \widehat{x^*}$. Then the game is in equilibrium with (x^*, y^*) as (MaxEnt,R_{min})-pair *iff* $H(x^*) < \infty$ and Nash's inequality
$\Phi(x, y^*) \leq \Phi(x^*, y^*)$ holds for all $x \in X_0$.

Nash inequality gives strong inequalities (Pythagorean ineq.).

Useful corollaries: Kuhn-Tucker type theorems (not discussed here) and robustness lemma below. Given preparation X_0, $y \in Y$ is robust if $\exists c < \infty \forall x \in X_0$: $\Phi(x, y) = c$. Define exponential family:
$\mathcal{E} = \mathcal{E}(X_0) = \{y | y \text{ robust }\}$.

Lemma If x^* is consistent and $y^* = \widehat{x^*}$ robust then (x^*, y^*) is a (MaxEnt,R_{min})-pair.

Only Axiom 1 required for this! Proof is easy.

Return to standard probabilistic setting:

Discrete alphabet $\mathbb{A}$, both strategy sets $= M_+^1(\mathbb{A})$, connection=identity. Now use notation: P, Q rather than x, y and $\mathcal{P}$ for typical preparation. We aim at discussing preparations given by linear constraints. Given set $f = (f_\nu)_{1\leq\nu\leq k}$ of real functions on $\mathbb{A}$. The natural preparations are

$$\mathcal{P}_a = \{P | \langle f_\nu, P\rangle = a_\nu \text{ for } 1 \leq \nu \leq k\}$$

with $a = (a_\nu)_{1\leq\nu\leq k} \in \mathbb{R}^k$.
Natural exponential family:

$$\mathcal{E} = \{Q |\ Q \text{ robust for all the } \mathcal{P}_a\}\,.$$

Question: which complexity measures ?
try to simplify search for distributions in $\mathcal{E}$!

key idea: Take Φ of the form

$$\Phi(P,Q) = \xi_Q\Big(\langle\overline{\kappa}(Q),P\rangle\Big) \text{ where}$$

$$\Big(\overline{\kappa}(Q)\Big)(i) = \kappa(q_i) \text{ for } i \in \mathbb{A} \text{ thus}$$

$$\langle\overline{\kappa}(Q),P\rangle = \sum_{i\in\mathbb{A}} p_i\kappa(q_i)\,.$$

with assumptions:

- The ξ_Q's increasing and concave (e.g. linear)
- the coder κ is smooth, decreasing, convex, $\kappa(1) = 0$
- Φ satisfies Axiom 1

classical: ξ_Q's the identity map, $\kappa(q) = \ln\frac{1}{q}$
Then κ^{-1} is restriction of $x \curvearrowright \exp(-x)$ to $[0,\infty]$. Entropy generated by this measure of complexity is standard BGS-entropy.

From previous page:

$$\Phi(P,Q) = \xi_Q\Big(\langle\overline{\kappa}(Q), P\rangle\Big) \text{ where}$$

$$\langle\overline{\kappa}(Q), P\rangle = \sum_{i\in\mathbb{A}} p_i \kappa(q_i)\,.$$

Trivial but key observation: any Q for which $\overline{\kappa}(Q)$ is a linear combination of the constant function 1 and the given functions $f_1, \cdots, f_k$, i.e. of the form

$$\overline{\kappa}(Q) = \lambda_0 + \lambda_1 f_1 + \cdots + \lambda_k f_k = \lambda_0 + \lambda \cdot f$$

for certain constants λ_0 and $\lambda = (\lambda_1, \cdots, \lambda_k)$, is a member of $\mathcal{E}$.

Motivated by this, fix constants $\lambda = (\lambda_1, \cdots, \lambda_k)$ and ask: $\exists \lambda_0, Q : \overline{\kappa}(Q) = \lambda_0 + \lambda \cdot f$? This amounts to $q_i = \kappa^{-1}(\lambda_0 + \lambda \cdot f(i))$. Summarizing what this leads to we find:

Theorem (MaxEnt calculus) Let $\lambda = (\lambda_1, \cdots, \lambda_k)$ be given constants. Then "normally", the equation

$$\sum_{i \in \mathbb{A}} \kappa^{-1}\Big(\lambda_0 + \lambda \cdot f(i)\Big) = 1$$

has a solution, necessarily unique, and Q given by

$$q_i = \kappa^{-1}\Big(\lambda_0 + \lambda \cdot f(i)\Big) \text{ for } i \in \mathbb{A}$$

has the stipulated form, hence belongs to the exponential family $\mathcal{E}$. This distribution is the MaxEnt-distribution for $\mathcal{P}_a$ with $a = (a_1, \cdots, a_k)$ given by

$$a_\nu = \sum_{i \in \mathbb{A}} q_i f_\nu(i) \text{ for } \nu = 1, \cdots, k$$

and, for this value of a,

$$\mathrm{H}_{\max}(\Phi, \mathcal{P}_a) = \xi_Q(\lambda_0 + \lambda \cdot a)\,.$$

Theorem replaces and expands the standard recipe for MaxEnt-calculations. Focus on λ_0 rather than on the classical partition function.

Which complexity measures?

Example The complexity measures

$$\Phi^B(P,Q) = \frac{1}{q-1} + \sum \left(q_i^q - \frac{q}{q-1} p_i q_i^{q-1} \right)$$
$$\Phi^C(P,Q) = \frac{1}{1-q} \sum p_i^q (1 - q_i^{1-q})$$
$$\Phi^R(P,Q) = \frac{1}{1-q} \left(\frac{\sum p_i^q}{\sum p_i^q q_i^{1-q}} - 1 \right).$$

all give Tsallis entropy. Only Φ^B is good! □

First restrict form of Φ: From

$$\Phi(P,Q) = \xi_Q\Big(\langle \overline{\kappa}(Q), P \rangle\Big) \text{ to}$$

$$\Phi(P,Q) = \langle \overline{\kappa}(Q), P \rangle + \overline{\xi}(Q) \text{ with}$$

$$\overline{\xi}(Q) = \sum_{i \in \mathbb{A}} \xi(q_i)\,.$$

κ is the coder, ξ the corrector.

Generation of information triples à la Bregman:

Bregman generator: a strictly concave and smooth real function h defined on $[0,1]$ with $\mathsf{h}(0)=\mathsf{h}(1)=0$ and $\mathsf{h}'(1)=-1$.

From h we generate two more functions, $\phi=\phi(p,q)$, and $\mathsf{d}=\mathsf{d}(p,q)$:

$$\phi(p,q)=\mathsf{h}(q)+(p-q)\,\mathsf{h}'(q)\,,$$
$$\mathsf{d}(p,q)=\mathsf{h}(q)-\mathsf{h}(p)+(p-q)\,\mathsf{h}'(q)\,.$$

Consider the internal functions, $\Phi=\Phi_{\mathsf{h}}$, $\mathsf{H}=\mathsf{H}_{\mathsf{h}}$ and $\mathsf{D}=\mathsf{D}_{\mathsf{h}}$ generated by ϕ, h and d:

$$\Phi(P,Q)=\sum_{i\in\mathbb{A}}\phi(p_i,q_i)\,,$$
$$\mathsf{H}(P)=\sum_{i\in\mathbb{A}}\mathsf{h}(p_i)\,,$$
$$\mathsf{D}(P,Q)=\sum_{i\in\mathbb{A}}\mathsf{d}(p_i,q_i)\,.$$

[the slides ended here – orally I ended by stressing the importance of not considering entropy alone, and the essentials of using the method of generation à la Bregman]