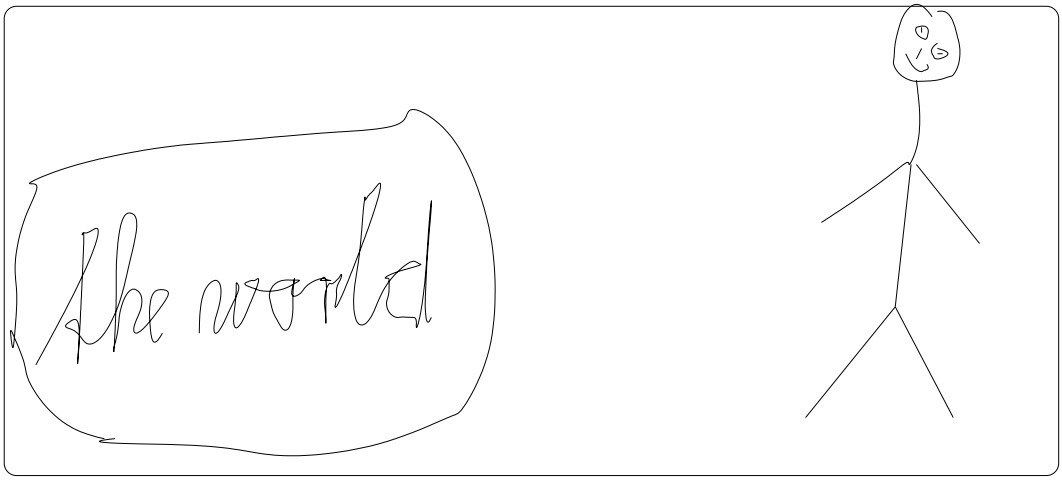
Optimization inspired by Information Theory

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences
Seminar, Statistics and Probability Theory Group,
April 9th, 2008

Inspired by previous work on information theoretical optimization, an axiomatic approach to certain special two-person zero-sum games is developed. One of the players (the statistician, physicist, investor, planner, or ...), is imagined to have a "mind", the other (the data, the physical system, the market, the costumers, ...) not. Somewhat provacatorially, it is the ambition of the speaker that after the talk, the audience will (finally!) understand what lies behind the notion of an exponential family.

A source of inspiration



Natures side
Player I

Observers side (you!)
Player II

A situation of conflict between two “persons”:
Player I who does not have a mind,
Player II – you – who does.

... and another

Motto:
When you seek an extremum, *don't* differentiate!

Example: minimum of $x^2 - 6x + 1$=?

1.st method: Differentiate
– and you are done.

2.nd method: Write expression as $(x-3)^2 - 8$
– and you are done

Best method? 2.nd! And more clear, e.g. you do indeed find a *minimum* this way.

Motto turned positive:
If your problem is *natural* ,
there is an *intrinsic method !*

Games!

(X, Y, Φ) defines a *two-person zero-sum game* with Φ as *objective function* (or *complexity*) if $\Phi : X \times Y \to \mathbb{R}$ (or $\overline{\mathbb{R}}$) and we associate notions of *optimal strategies* and *equilibrium* with this function.

Players and strategies:

Player I is a maximizer, , chooses x

Player II is a minimizer, chooses y.

Specific and *global values:*

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

$$\mathrm{val}_{\mathrm{I}} = \sup_{x \in X} \mathrm{val}_{\mathrm{I}}(x)\,,$$

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

$$\mathrm{val}_{\mathrm{II}} = \inf_{y \in Y} \mathrm{val}_{\mathrm{II}}(y)\,.$$

$y \in Y$ is an *optimal response* *to* $x \in X$ if $\Phi(x, y) = \mathrm{val}_{\mathrm{I}}(x)$. We put $\widehat{x} = \mathrm{resp}_{\mathrm{II}}(x) = \mathrm{argmin} \Phi_x$.

Terminology: y *is adapted to* x or x is a *matching strategy* to y (matches Pl.-II aim to respond optimally).

$$y \in \mathrm{resp}_{\mathrm{II}}(x) \Leftrightarrow x \in \mathrm{match}_{\mathrm{I}}(y) \quad \Leftrightarrow y \in \mathrm{argmin}\Phi_x\,,$$
$$x \in \mathrm{resp}_{\mathrm{I}}(y) \Leftrightarrow y \in \mathrm{match}_{\mathrm{II}}(x) \Leftrightarrow x \in \mathrm{argmax}\Phi^y\,.$$

Visualization: e.g. as a matrix

	y_1	y_2	y_3	y_4	y_5	$\mathrm{val}_{\mathrm{I}}(\cdot)$
x_1	0	0	5	1	3	0
x_2	3	5	1	1	1	**1**
x_3	2	0	1	5	2	0
$\mathrm{val}_{\mathrm{II}}(\cdot)$	**3**	5	5	5	**3**	

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

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

Define $\mathrm{span}(x,y) = \mathrm{val}_{\mathrm{II}}(y) - \mathrm{val}_{\mathrm{I}}(x)$, then:
$\mathrm{span}(x,y) = \delta_{\mathrm{I}}(x,y) + \delta_{\mathrm{II}}(x,y)$, hence:
minimax inequality holds: $\mathrm{val}_{\mathrm{I}} \le \mathrm{val}_{\mathrm{II}}$.

Game Theoretical Equilibrium: If $\mathrm{val_I} = \mathrm{val_{II}} \in \mathbb{R}$. Ideally: γ_Φ is in game theoretical equilibrium and has optimal strategies, say (x_0, y_0). Notation: $\gamma_\Phi \in \mathrm{GTE}(x_0, y_0)$. If (x_0, y_0) not specified, write $\gamma_\Phi \in \mathrm{GTE}^*$.

If $\Phi(x_0, y_0) \in \mathbb{R}$, then (Player-I satisfaction):

$$\Phi(x, y_0) \leq \Phi(x_0, y_0) \text{ for all } x \in X\,,$$
$$\mathrm{val_{II}}(y_0) = \Phi(x_0, y_0)\,,$$
$$\delta_{\mathrm{I}}(x_0, y_0) = 0\,,$$
$$x_0 \in \mathrm{resp_I}(y_0)\,,$$
$$y_0 \in \mathrm{match_{II}}(x_0)$$

are equivalent and so are (Player-II satisfaction):

$$\Phi(x_0, y_0) \leq \Phi(x_0, y) \text{ for all } y \in Y\,,$$
$$\mathrm{val_I}(x_0) = \Phi(x_0, y_0)\,,$$
$$\delta_{\mathrm{II}}(x_0, y_0) = 0\,,$$
$$y_0 \in \mathrm{resp_{II}}(x_0)\,,$$
$$x_0 \in \mathrm{match_I}(y_0)\,.$$

saddle-value theorem: Let $(x_0, y_0) \in X \times Y$ and assume that $\Phi(x_0, y_0) \in \mathbb{R}$. Then the following conditions are equivalent:

$$
\begin{aligned}
& \gamma_\Phi \in \mathsf{GTE}(x_0, y_0)\,, \\
\rightarrow\ & \forall (x, y):\ \Phi(x, y_0) \leq \Phi(x_0, y_0) \leq \Phi(x_0, y)\,, \\
& \mathsf{val}_{\mathrm{I}}(x_0) = \Phi(x_0, y_0) = \mathsf{val}_{\mathrm{II}}(y_0)\,, \\
& \delta_{\mathrm{I}}(x_0, y_0) = \delta_{\mathrm{II}}(x_0, y_0) = 0\,, \\
& x_0 \in \mathsf{resp}_{\mathrm{I}}(y_0) \text{ and } y_0 \in \mathsf{resp}_{\mathrm{II}}(x_0)\,, \\
& y_0 \in \mathsf{match}_{\mathrm{II}}(x_0) \text{ and } x_0 \in \mathsf{match}_{\mathrm{I}}(y_0)\,.
\end{aligned}
$$

If so, $\mathsf{val}(\gamma_\Phi) = \Phi(x_0, y_0)$.

In this case, we talk about: *Nash equilibrium, Nash equilibrium pair, saddle-value inequalities, saddle point*.

Thus $\gamma_\Phi \in \mathsf{GTE}^* \Leftrightarrow \exists$ saddle point.

...

(figures very helpful - plan to have some on the blackboard)

Subgames, preparations

Introduce *subgames* by restricting strategy set for Player I: $\gamma_\Phi(X_0)$ corresponding to a *preparation*, X_0. Let Γ be class of all subgames. Expand notation: $\mathsf{val}_{\mathrm{I}}(X_0)$, $\mathsf{val}_{\mathrm{II}}(y|X_0)$ and $\mathsf{val}_{\mathrm{II}}(X_0)$ if necessary.

Level- and sub-level sets become important: $\mathsf{L}^y(h) = \{\Phi^y = h\}$, $\mathsf{SL}^y(h) = \{\Phi^y \leq h\}$, in full:

$$\mathsf{SL}^y(h) = \{x \in X | \Phi^y(x) \leq h\} = \{x \in X | \Phi(x,y) \leq h\}$$

From the saddle-value theorem:

(x_0, y_0) are optimal strategies for a subgame in equilibrium *iff* $\Phi(x_0, y_0) \in \mathbb{R}$ and y_0 is adapted to x_0. If so, the possible preparations are all X_0 with

$$\{x_0\} \subseteq X_0 \subseteq \mathsf{SL}^{y_0}(h) \text{ with } h = \Phi(x_0, y_0)\,.$$

The sets $(\mathsf{SL}^y(h))_{y,h}$ with the level h some finite value of Φ^y for an argument which matches y are thus the maximal preparations.

Typically the maximal preparations are not "practically feasible" – but level sets are. Given a preparation X_0, define the exponential family (Player-II domain), by

$$\mathcal{E}_{\mathrm{II}}(X_0) = \{y \in Y | \exists h \in \mathbb{R} : \; X_0 \subseteq \mathsf{L}^y(h)\} \, .$$

An $y \in \mathcal{E}_{\mathrm{II}}(X_0)$ is a robust Pl.-II strategy and h is the level of robustness.

> If y_0 is robust and adapted to $x_0 \in X_0$, then $\gamma_\Phi(X_0) \in \mathsf{GTE}(x_0, y_0)$ and $\mathsf{val}(\gamma_\Phi(X_0)) = \Phi(x_0, y_0)$.

Exponential family for a *preparation family* $\mathcal{X}$:

$$\mathcal{E}_{\mathrm{II}}(\mathcal{X}) = \bigcap_{X_0 \in \mathcal{X}} \mathcal{E}_{\mathrm{II}}(X_0) \, .$$

If $\mathcal{E}_{\mathrm{II}}$ is an exponential family (Pl.-II domain), the corresponding exponential family (Pl.-I domain) is

$$\mathcal{E}_{\mathrm{I}} = \{x | \, \mathsf{resp}_{\mathrm{II}}(x) \cap \mathcal{E}_{\mathrm{II}} \neq \emptyset\} = \mathsf{match}_{\mathrm{I}}(\mathcal{E}_{\mathrm{II}}) \, .$$

From $\mathcal{E}_{\mathrm{II}}$ define family $\mathcal{X}$ of associated preparations:

$$\mathcal{E}_{\mathrm{II}}^{\perp} = \bigcup_{y \in \mathcal{E}_{\mathrm{II}}} \{\mathsf{L}^y(h) | h \in \mathbb{R}\} \, .$$

More another time, to be worked out (geometry etc.)

Axioms for Complexity, entropy, divergence.

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

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

Axiom 2 X is convex and Φ affine in first variable: For $y \in Y$, $\alpha \in \mathrm{MOL}(X)$,

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

Axiom 3 X is topological, algebraic operations continuous and, for each $(x_0, y_0) \in X \times Y$, $x \curvearrowright \mathsf{D}(x, y_0)$ and $x \curvearrowright \mathsf{D}(x_0, \widehat{x})$ are lower semi-continuous.

Axiom 4 Every D-Cauchy sequence has a convergent subsequence.

$(x_n)_{n \geq 1}$ D-*Cauchy* means:
$\lim_{n,m \to \infty} \mathsf{D}\left(x_n, \widehat{(\frac{1}{2}x_n + \frac{1}{2}x_m)}\right) = 0.$

First consequences

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

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

Concavity and convexity properties:

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

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

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

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

Last identity is the compensation identity.

Main theorem

Assume Axioms 1-4 are satisfied and let X_0 be a convex preparation.

A sequence $(x_n) \subseteq X_0$ is *asymptotically optimal* if $\lim_{n\to\infty} \mathsf{H}(x_n) = \mathsf{H}_{\max}$. A strategy $x \in X$ (not necessarily in X_0) is an $\mathsf{H}_{\max}$*-attractor* if $\mathsf{D}(x_n, \widehat{x}) \to 0$ for every asymptotically optimal sequence (x_n).

If X_0 is convex and $\mathsf{H}_{\max}(X_0) < \infty$, then Pl. II has a unique optimal strategy y^*, and a $\mathsf{H}_{\max}$-attractor x^* exists and $y^* = \widehat{x^*}$. The game is in equilibrium and for each $x \in X_0$ and each $y \in Y$:

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

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,

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

Example 2

$$\begin{aligned}\phi(x,y) &= x\,\mathit{ln}\,\frac{1}{y}\,,\\ \mathsf{h}(x) &= x\,\mathit{ln}\,\frac{1}{x}\,,\\ \mathsf{d}(x,y) &= x\,\mathit{ln}\,\frac{x}{y}\,.\end{aligned}$$

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* .

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:

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

By equivalence you obtain:

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

Classical Information Theory

Let $\mathbb{A}$, the *alphabet*, be discrete, put $X = M_+^1(\mathbb{A})$, and $Y = K(\mathbb{A})$, the set of *code length functions* over $\mathbb{A}$, i.e. the set of $\kappa : \mathbb{A} \to [0, \infty]$ such that *Krafts equality*

$$\sum_{i \in \mathbb{A}} e^{-\kappa_i} = 1$$

holds. The response $P \curvearrowright \kappa$ is defined by $\kappa_i = \mathit{ln}\, \frac{1}{p_i}$; $i \in \mathbb{A}$ and for Φ we take *average code length*, i.e.

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

Then

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

$$\mathsf{D}(P, Q) = \sum_{i=1}^{n} p_i \, \mathit{ln}\, \frac{p_i}{q_i} \,.$$

Consider preparation of the form

$$\mathcal{P} = \{P | \langle f_k, P \rangle = a_k \text{ for } k = 1, \cdots, m\}$$

and seek MaxEnt-distribution. Lagrange multipliers one possibility. Better: Seek $\kappa \in \mathcal{E}_{\mathrm{II}}$. Clearly, κ of the form

$$\kappa = \alpha + \beta_1 f_1 + \cdots + \beta_m f_m$$

works. By Krafts equality this requires that

$$\alpha = \mathit{ln}\, Z(\beta_1, \cdots, \beta_m)$$

with the partition function given by

$$Z(\beta_1, \cdots, \beta_m) = \sum_{i \in \mathbb{A}} \exp\left(-\beta_1 f_1(i) - \cdots - \beta_m f_m(i)\right).$$

Adjust coefficients so that the matching distribution is consistent ($\in \mathcal{P}$), and you are done.

.....

Further examples include separation, location theory, universal coding, determination of capacity, duality and more ... Next time!