Applications of Game Theoretical Optimization Techniques inspired by Information Theory

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences
topsoe@math.ku.dk, www.math.ku.dk/~topsoe
Poster Presentation for ISIT2007, Nice, June 2007

Abstract as communicated to ISIT2007

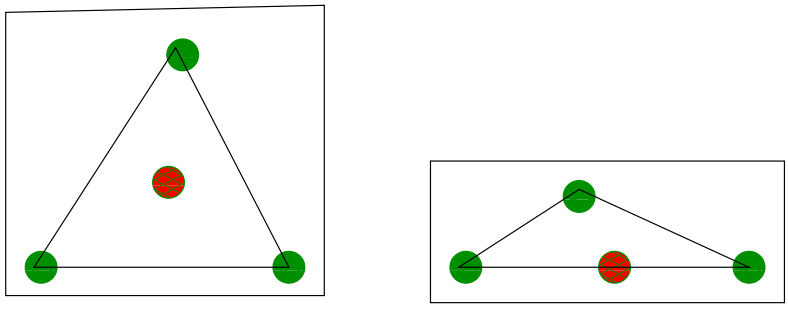
Previous joint work with Peter Harremoës points to a common basis for much of information theoretical optimization (MaxEnt, Minimum Discrimination, Capacity-redundancy as main cases). The starting point may be taken to be a kind of ”complexity function” (key example re MaxEnt is avarage code length). The results hinted at constitute important applications in their own right (publications of Csiszár, Grünwald and Dawid, the author, ...). It turns out that going outside information theory proper and applying the techniques to other kinds of ”complexity functions” interesting connections to other areas may be established. The strength of this is being explored but not yet totally clarified. There appears to be room for further investigations that may attract the interest of young researchers.

Two anniversaries: 150 years and 50 years

Sylvester 1857: start of *location theory*. He wrote:

> "*It is required to find the least circle which shall contain a given system of points in the plane*".

A man of few words. Two examples: Easy (3 *anchors*) and difficult (2 anchors and 1 *inactive* point)



Second anniversary: Jaynes 1957: MaxEnt, the *maximum entropy principle* of statistical physics (and ...):

> *Base inference in statistical physics on the principle to point to that distribution, consistent with available information, which has maximal entropy*.

What do the two problems have in common?

A notion of complexity

Let $\Phi = \Phi(P, Q)$ be a "complexity function". Assume: "diagonal condition" holds: $\Phi(P, Q) \geq \Phi(P, P)$ with strict inequality if $Q \neq P$. P and Q to vary over the *same* "basic set". But P and Q are *treated differently:*

P refers to "nature", "the system", (what you cannot know about – but want to know) and Q refers to "you" (the communicator, the statistician, the physicist, the investor, ...) and reflects what *YOU* can do about it!

Define Φ-Entropy and divergence by:

$$\mathsf{H}(P) = \min_{Q} \Phi(P, Q)\,, \mathsf{D}(P, Q) = \Phi(P, Q) - \mathsf{H}(P)$$

entropy = minimal complexity
Divergence = actual − minimal complexity

EXAMPLE (information theory): P, Q's: probability distributions over an "alphabet", here a finite set, say.

$$\Phi(P,Q) = \sum p_i \ln \frac{1}{q_i} = \langle \ln \frac{1}{Q}, P \rangle$$

(average codelength!). Then
$\mathsf{H}(P) = \sum p_i \ln \frac{1}{p_i}$ and $\mathsf{D}(P,Q) = \sum p_i \ln \frac{p_i}{q_i}$ □

EXAMPLE (geometry): P, Q's: Points in Euclidean space, say $\mathbb{R}^2$.

$$\Phi(P,Q) = \|P - Q\|^2$$

(squared Euclidean distance!). Then $\mathsf{H}(P) = 0$ (!), $\mathsf{D}(P,Q) = \Phi(P,Q)$. What is this good for? Wait and see! □

MaxEnt :
Given $\mathcal{P}$, a preparation (set of P's). MaxEnt-distribution is the $P \in \mathcal{P}$ of maximal entropy (if well defined). MaxEnt-value : $H_{\max}(\mathcal{P}) = \max_{P \in \mathcal{P}} H(P)$.

A *highly useful, trivial,* but *neglected criterion*:

If $Q \in \mathcal{P}$ is robust: $\Phi(P, Q)$ independent of $P \in \mathcal{P}$, say $\forall P \in \mathcal{P} : \Phi(P, Q) = h$, then Q is the MaxEnt-distribution and $\mathsf{H}_{\max}(\mathcal{P}) = h$.

Proof. Firstly: $\mathsf{H}(Q) = \Phi(Q, Q) = h$.
Secondly: if $P \neq Q$, $P \in \mathcal{P}$, then
$H(P) < H(P) + \mathsf{D}(P, Q) = \Phi(P, Q) = h$. □

EXAMPLE (information theory). For a function f on the alphabet and a prescribed meanvalue a, look at the preparation

$$\mathcal{P} = \{P | \langle f, P \rangle = a\}.$$

By robustness, if $Q \in \mathcal{P}$ and there are constants λ_0 and λ such that

$$\ln \frac{1}{Q} = \lambda_0 + \lambda \cdot f$$

then Q is the MaxEnt-distribution. □

The technique can be expanded and leads to (grand) canonical ensembles of statistical thermodynamics in a rather direct way and without using Lagrange multipliers. Expansion to cover MinXEnt or minimum discrimination principle of Kullback also possible (for indications/reminders, see further on).

Enter games

Game γ depending on Φ and preparation $\mathcal{P}$
Φ is the objective function
Player I chooses $P \in \mathcal{P}$ and is a "maximizer"
Player II chooses (any) Q and is a "minimizer"

Player I-value:
$\sup_{P\in\mathcal{P}} \inf_Q \Phi(P,Q) = \sup_{P\in\mathcal{P}} \mathsf{H}(P) = \mathsf{H}_{\max}(\mathcal{P})$.

Player II-value:
$\inf_Q \sup_{P\in\mathcal{P}} \Phi(P,Q) = \inf_Q \mathsf{R}(Q|\mathcal{P}) = \mathsf{R}_{\min}(\mathcal{P})$
("R" for "risk").

By minimax inequality, $\mathsf{H}_{\max} \leq \mathsf{R}_{\min}$. If equal, the game γ is in equilibrium.

Theorem Assume P's and Q's range over convex topological spaces (such as probability distributions or points in Euclidean space), that $P \curvearrowright \Phi(P, Q)$ is concave for all Q and that certain technical topological conditions are satisfied.
Then every game $\gamma = \gamma(\mathcal{P})$ with $\mathcal{P}$ convex and $\mathsf{H}(\mathcal{P}) < \infty$ is in equilibrium. Furthermore, Player II has a unique optimal strategy Q^* ($\mathsf{R}(Q^*|\mathcal{P}) = \mathsf{R}_{\min}(\mathcal{P})$), so $\sup_{P\in\mathcal{P}} \Phi(P, Q^*) = \min_Q \mathsf{R}(Q|\mathcal{P})$. Finally, $\forall(P_n) \subseteq \mathcal{P}$ with $\mathsf{H}(P_n) \to \mathsf{H}_{\max}(\mathcal{P})$: $P_n \to Q^*$. (So if e.g. $\mathcal{P}$ is compact, Q^* is the MaxEnt distribution).

In short:

"normally", if Φ is concave in the first variable, then $\mathsf{H}_{\max} = \mathsf{R}_{\min}$ for convex preparations.

Sylvesters problem, universal prediction

Take any divergence function which satisfies the compensation identity :

$$\sum \alpha_\nu \, \mathsf{D}(P_\nu, Q) = \sum \alpha_\nu \, \mathsf{D}(P_\nu, \overline{P}) + \mathsf{D}(\overline{P}, Q)$$

($\overline{P} = \sum \alpha_\nu P_\nu$, a convex mixture of the P_ν's).

e.g. D could come from a Φ which is concave in the first variable or D is "geometric divergence": $\mathsf{D}(P, Q) = \|P - Q\|^2$.

The Game based on D is *not* in equilibrium. Player I-side is trivial and uninteresting, but Player II-side is just what we are interested in (Sylvester's problem, universal coding). Thus Q^* with
$\mathsf{R}(Q^*) = \sup_{P \in \mathcal{P}} \mathsf{D}(P, Q^*)$ is what we are looking for.

Randomize "a la von Neumann": weights $\alpha = (\alpha_P)$ are considered for Player I-choices and for complexity

we take $\Phi(\alpha, Q) = \sum \alpha_P \, \mathsf{D}(P, Q)$. "Entropy" for this kind of complexity function is denoted I (information transmission rate in the information theory case):

$$\begin{aligned} \mathsf{I}(\alpha) &= \inf_Q \sum \alpha_P \, \mathsf{D}(P, Q) \\ &= \sum \alpha_P \, \mathsf{D}(P, \overline{P}) \text{ by compensation id.} \end{aligned}$$

Note: The Player II-quantities (the R's) do not change by randomization. Further, Φ satisfies conditions of the theorem. Hence Gallager-Ryabko theorem holds in abstract setting and then includes Sylvester's problem. So $\boxed{\mathsf{I}_{\max} = \mathsf{R}_{\min}}$ ("capacity"="minimal redundancy"). Also, the Kuhn-Tucker criterion applies ...

Priors (Bayesian considerations)

Given Φ and prior Q_0. Updating gain is defined as

$$\Psi_{Q_0}(P, Q) = \Phi(P, Q_0) - \Phi(P, Q) \, .$$

As $-\Psi$ behaves as a complexity function, previous theory carries over (with reversal of some signs).

Now, Player I is a minimizer, Player II as maximizer.

If Player I chooses $P \in \mathcal{P}$, he associates the value (accepting the "risk" that Player II responds optimally)

$$\sup_Q \Psi_{Q_0}(P,Q) = \sup_Q \Big(\mathsf{D}(P,Q_0) - \mathsf{D}(P,Q) \Big) = \mathsf{D}(P,Q_0)$$

with this choice and aims for a $P^* \in \mathcal{P}$ with

$$\mathsf{D}(P^*,Q_0) = \min_{P\in\mathcal{P}} \mathsf{D}(P,Q_0) = \mathsf{D}_{\min}(\mathcal{P})\,.$$

For Player II, the value associated with the choice Q as update is the "guaranteed gain"

$$\Gamma(Q) = \inf_{P\in\mathcal{P}} \Psi_{Q_0}(P,Q)$$

and the value to aim for for Player II is

$$\Gamma_{max}(\mathcal{P}) = \sup_Q \Gamma(Q)\,.$$

In information theory all this is well known and goes back to Kullback's minimum information discrimination principle, Csiszár's I-projection studies etc. In

the physical literature this is mainly referred to as the MinXEnt principle (due to Jaynes) (minimum cross entropy principle) . The novelty is the generality under which all this holds. Not that difficult, but useful. Let's see an implication for geometry and prove a classical fact by these considerations:

If $\mathcal{P} \subseteq \mathbb{R}^n$ is convex and compact, and Q_0 a point outside $\mathcal{P}$ then a hyperplane through Q_0 separates Q_0 from $\mathcal{P}$.

Proof Consider updating as above with Q_0 as prior and based on geometric complexity. Observe that

$$\Psi_{Q_0}(P, Q) = 2\langle P - Q, Q - Q_0\rangle + \|Q_0 - Q\|^2 ,$$

hence this function is affine in the first variable. Furthermore, as is easily checked, the compensation identity hold. Thus main theorem applies and $\Gamma_{\max}(\mathcal{P}|Q_0) = \mathsf{D}_{\min}(\mathcal{P}|Q_0)$. Then argue as follows:

1. As $Q_0 \notin \mathcal{P}$ and as $\mathcal{P}$ is compact, $\mathsf{D}_{\mathsf{min}}(\mathcal{P}) > 0$.

2. Hence, by the main theorem, $\Gamma_{\mathsf{max}}(\mathcal{P}) > 0$, and we may choose Q such that $\Gamma(Q) > 0$.

3. Then $Q \neq Q_0$ and we can consider the hyperplane π through Q_0 which has the linesegment Q_0Q as normal. Consider that halfspace determined by π which does not contain Q_0. Then this halfspace (including the hyperplane π) cannot contain any point from $\mathcal{P}$ since, for any point S in the halfspace concerned, $\Psi_{Q_0}(S, Q) = \|S - Q_0\|^2 - \|S - Q\|^2 < 0$. We have thus found a separating hyperplane.