Intrinsic methods for optimization problems

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences
Presentation for ISIT2008

Thesis: *when a specific problem of optimization is "canonical", i.e. works with "just the right concepts" and reflects "just the right questions", then intrinsic tools are the way forward to insight.*

MaxEnt

Consider a preparation consisting of all probability distributions over a discrete alphabet $\mathbb{A}$ with given mean "energy":

$$X_0 = \{x | \sum_{i \in \mathbb{A}} x_i E_i = \overline{E}\}\,.$$

Problem: Maximize entropy $\mathrm{H}(x)$ over X_0.

Solution 1: Introduce Lagrange multipliers!

Solution 2: First introduce more structure (next 3-4 slides):

Define the set Y of codes (code length functions) by

$$Y = \{y | \sum_{i \in \mathbb{A}} e^{-y_i} = 1\}$$

or, equivalently, $Y = \{y | \sum_{i \in \mathbb{A}} \xi_i = 1\}$ with ξ the matching distribution, also thought of as the belief corresponding to y, i.e. $\xi_i = e^{-y_i}$ for $i \in \mathbb{A}$.

Consider the complexity function:

$$\Phi(x, y) = \sum_{i \in \mathbb{A}} x_i y_i = \sum_{i \in \mathbb{A}} x_i \ln \frac{1}{\xi_i} .$$

Then $\boxed{\Phi(x, y) = \mathsf{H}(x) + \mathsf{D}(x, y)}$ (linking identity).

Here, $\mathsf{D}(x, y) = \mathsf{D}(x \| \xi) = \sum_{i \in \mathbb{A}} x_i \ln \frac{x_i}{\xi_i}$, standard Kullback-Leibler divergence.

Recall the fundamental inequality:

$$\text{D}(x,y) \geq 0 \text{ with D} = 0 \Leftrightarrow y = \widehat{x}$$

Here, y is adapted to x , $y = \widehat{x}$, if $y_i =$ *ln* $\frac{1}{x_i}$ for $i \in \mathbb{A}$. Thus entropy is minimal complexity:

$$\text{H}(x) = \min_{y \in Y} \Phi(x,y) = \min \Phi_x \, .$$

We call y robust (member of the associated exponential family $\mathcal{E}$) if $\exists h < \infty \forall x \in X_0 : \; \Phi(x,y) = h$.

Robustness lemma: If (x^*, y^*) has $y^* = \widehat{x^*}$, $x^* \in X_0$ and y^* robust, then x^* is the MaxEnt distribution.

Proof:

1: $\mathsf{H}(x^*) = \Phi(x^*, \widehat{x^*}) = \Phi(x^*, y^*) = h$,

hence, as $x^* \in X_0$, $\mathsf{H}_{\max} \geq h$,

2: For $x \in X_0 \setminus \{x^*\}$,

$\mathsf{H}(x) < \mathsf{H}(x) + \mathsf{D}(x, y^*) = \Phi(x, y^*) = h$. **qed**

Having introduced more structure, and armed with the robustness lemma, the second solution is:

- seek codes of the form $y^* = \alpha + \beta E$,
- note that, trivially, these codes are all robust,
- apply the robustness lemma.

Axiomatize !

Consider information triples (Φ, H, D), in more detail $(X, Y, x \curvearrowright \widehat{x}, \Phi, \mathsf{H}, \mathsf{D})$, satisfying:

Axiom 1: Linking+fundamental inequality,
Axiom 2: X is convex and, for all y, Φ^y is affine (Φ^y is the marginal function $x \curvearrowright \Phi(x, y)$).

Then robustness lemma holds for *any* preparation X_0. But: most natural to consider, given y, the associated natural preparation family which consists of all preparations which are level sets of Φ^y, i.e. sets of the form $X_0 = \{\Phi^y = h\}$.

These are preparations of genus 1 as only one constraint is involved. Generalization to finite genus is straight forward.

Applications

MaxEnt: As before but more general triples than "Shannon triple" are possible using Bregman construction.

Capacity-redundancy: Consider DMC and randomize to obtain an appropriate set X of distributions over the input alphabet, take as Y output distributions, as Φ expected divergence ... (if time, see details further on).

MinDiv, updating: Consider a prior and measure performance relative to this. Leads to minimum discrimination principle and to information projections. The important process is one of relativization.

Updating in Hilbert space: Consider:

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

The natural preparation family gives families of hyperplanes and the robustness lemma gives the natural projections of the prior onto these hyperplanes.

Sylvester's problem: "*It is required to find the least circle which shall contain a given system of points in the plane*". Treated as the capacity-redundancy problem ... (if time, see next pages)

... further details on my homepage or in forthcoming publications.

common treatment of capacity-redundancy (CR) and Sylvester (S) problems

Given $(P_i)_{i \in \mathbb{A}}$. Let X be the set of distributions $\alpha = (\alpha_i)_{i \in \mathbb{A}}$.

S: The P_i are the given points (in the plane or ...). Let Y be the set of *all* points in the plane (or...). $\mathrm{D}(P, Q)$ denotes $\|P - Q\|^2$.

CR: $\mathbb{A}$ is the input alphabet of the DMC, the P_i's, distributions over an output alphabet , $\mathbb{B}$, the output distributions of the DMC. Let Y be the set of *all* distributions over $\mathbb{B}$. $\mathrm{D}(P, Q)$ denotes K-B divergence.

For both cases, $b(\alpha)$ with $\alpha \in X$, denotes the barycenter $\sum_{i \in \mathbb{A}} \alpha_i P_i$.

In both cases, the compensation identity holds:

$$\sum_{i \in \mathbb{A}} \alpha_i \, \mathsf{D}(P_i, Q) = \sum_{i \in \mathbb{A}} \alpha_i \, \mathsf{D}(P_i, b(\alpha) + \mathsf{D}(b(\alpha), Q)\,.$$

holds for any $Q \in Y$. Therefore,

$$\begin{aligned} \Phi(\alpha, Q) &= \sum_{i \in \mathbb{A}} \alpha_i \, \mathsf{D}(P_i, Q)\,, \\ \mathsf{H}(\alpha) &= \sum_{i \in \mathbb{A}} \alpha_i \, \mathsf{D}(P_i, b(\alpha))\,, \\ \mathsf{D}(\alpha, Q) &= \mathsf{D}(b(\alpha), Q) \end{aligned}$$

satisfies Axioms 1 and 2. Instead of robustness you here use a more general result, Nash's saddle-value inequalities. They leed directly to the usual Kuhn-Tucker criterion which provides the intrinsic method sought for.

Insights, view points, a question

1 intrinsic solutions to natural problems is possible, depends on adequate structure and leeds to insight

2 Game theory appears appropriate as it stresses the interplay between you and the part of "nature" you are studying

3 key results (robustness, Kuhn-Tucker) follow from general game theory, especially results due to Nash

4 Never consider entropy alone

5 always consider exponential families alongside with the related natural preparation families

6 Axiomatic approach devoid, in principle, of information theory: Is it a gift from information theory to optimization theory – or should parts of information theory be much broadened and subsumed under a more general mathematical theory ?