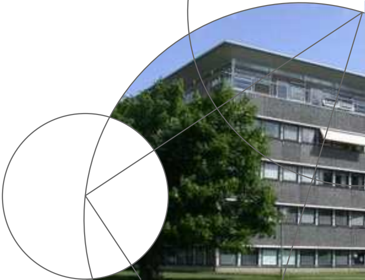
Beyond Shannon with examples from Geometry, Statistics and Information Theory

Flemming Topsøe, topsoe@math.ku.dk
Department of Mathematical Sciences, University of Copenhagen



NOTE: Essentially same slides used for internal talk:

Isotone Regression and Universal Coding

(September 3rd, University of Copenhagen).

Outline

Goal: To develop abstract theory inspired by information theoretical thinking, aiming at fascillitating inference in situations involving elements of cognition.

Approach: Via two-person zero-sum games with **N**ature, holder of "truth" , and **O**bserver, steared by belief, as players.

What do players fight about: Either effort $\Phi = \Phi(x,y)$ with **N** as maximizer and **O** as minimizer, or gain (same as utility) $\mathrm{U} = \mathrm{U}(x,y)$ with **N** as minimizer and **O** as maximizer.

Inference: Will be based on theorems of (Nash-) equilibrium. Results only briefly indicated on next slide.

Focus: This will be on three problems which serve as examples and indicate the range of possible applications.

Indication of basic elements of abstract theory

Effort based information triple $(\Phi, \mathrm{H}, \mathrm{D})$ should satisfy $\Phi(x,y) = \mathrm{H}(x) + \mathrm{D}(x,y)$: : Effort=entropy+divergence.

Information triple based on gain $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ should satisfy $\mathrm{U}(x,y) = \mathrm{M}(x) - \mathrm{D}(x,y)$: Gain = maxgain - divergence.

In both cases, divergence should satisfy the fundamental inequality: $\mathrm{D}(x,y) \geq 0$ with $\mathrm{D}(x,y) = 0 \Leftrightarrow y = x$.

Adding assumptions (like affinity of $\Phi(\cdot, y)$ or $\mathrm{U}(\cdot, y)$) one finds that *"TYPICALLY"*:

- If **N**- and **O**-strategies are of the same kind, there exists a bi-optimal strategy $x^* = y^*$, i.e. one which is optimal for both players.
- If **N**- and **O**-strategies are of different kinds, unique optimal strategies, x^* and y^*, still exist and in this case y^* is derived directly from x^* in a natural way (e.g. y^* = barycentre of x^*).

1.st problem: From Geometric Location Theory

Sylvester (1857):"It is required to find the least circle which shall contain a given system ($\mathcal{P}$) of points in the plane".

So, task is to minimize $\max_{x \in \mathcal{P}} \|x - y\|$. Can we take $\Phi(x, y) = \|x - y\|$? No!

1.st modification: May as well replace $\| \cdot \|$ with any increasing function f of $\| \cdot \|$ (say with $f(0) = 0$). Still no chance of equilibrium as maximin-value is 0.

2.nd modification: Randomize! Combining with 1., you should look at $\Phi(\alpha, y) = \sum_{x \in \mathcal{P}} \alpha_x \|x - y\|^2$ with α's certain weights.

This works!

2.nd probl.: From universal prediction and coding

Let $\Omega = (\Omega, \leq)$ be a finite partially ordered set and consider the model $\mathcal{A}$ of all antitone probability distributions over Ω ($a \leq b \Rightarrow x(a) \geq x(b)$). Requested is the distribution $y = y^*$, the universal predictor, with

$$\sup_{x \in \mathcal{A}} \mathrm{D}(x, y) = \sup_{x \in \mathcal{A}} \sum_{a \in \Omega} x(a) \ln \frac{x(a)}{y(a)}$$

minimal. As you see, here D is Kullback-Leibler divergence .

Similar remarks as those to Sylvester's problem apply.

The equivalent problem of universal coding is to find the code length function $\kappa = \kappa^*$ associated with y^*, i.e. $\kappa^*(a) = \ln \frac{1}{y^*(a)}$ for $a \in \Omega$.

Note: κ's are isotone ($a \leq b \Rightarrow \kappa(a) \leq \kappa(b)$).

3.rd problem: From statistics, isotone regression

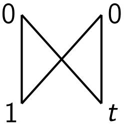
Again, Ω is a finite partially ordered set. Given is a weight function W (e.g. the uniform distribution) and a prior, y_0, just any function on Ω, referred to as a valuation. Sought is $y = y^*$, the posterior, required to be the isotone valuation closest to y_0 in squared norm.
Thus we should minimize

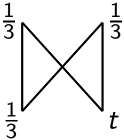
$$\|y_0 - y\|^2 = \sum_{a \in \Omega} W(a)|y_0(a) - y(a)|^2 \qquad (1)$$

subject to a requirement on y of isotonicity.

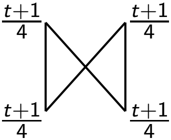
Existence and uniqueness of y^* is pretty evident. It is the isotone regression of y_0.
This problem we shall study in more detail. Let us start with a simple example:

(Ω, y_0): Butterfly set with prior y_0 depending on parameter t; listing also its isotone regression y^*

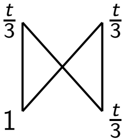




Case 1: $t < \frac{1}{3}$



Case 2: $\frac{1}{3} \leq t \leq 3$



Case 3: $t > 3$

NOTE: Every valuation y is determined by a decomposition $\mathcal{S}_y$ of Ω in maximal connected sets of y-constancy and the associated values $\alpha_y = \big(\alpha_y(S)\big)_{S \in \mathcal{S}_y}$. So, problem is to determine $\mathcal{S}_{y^*}$ and α_{y^*}. In fact, only $\mathcal{S}_{y^*}$ is needed:

Below, a lower set L is a union of left sections (write $L \in \mathcal{L}$). For $A \subseteq \Omega$, conditional y_0-average over A is denoted $\overline{A}$, i.e. $\overline{A} = \sum_{a \in A} W(a|A)y_0(a) = \frac{1}{W(A)} \sum_{a \in A} W(a)y_o(a)$.

Theorem [Identification] With y_0 as prior, assume that a valuation y is given in terms of $\mathcal{S}_y$ and α_y. Then, necessary and sufficient for y to be the isotone regression of y_0 ($y = y^*$) is that the following holds:

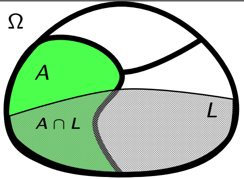
(i) [ordering] $\mathcal{S}_y$ is partially ordered in the natural ordering;

(ii) [monotonicity] $A < B \Rightarrow \alpha(A) < \alpha(B)$ for $A, B \in \mathcal{S}_y$;

(iii) [proper values] for $A \in \mathcal{S}_y$, $\alpha(A) = \overline{A}$;

(iv) [boundedness] for $A \in \mathcal{S}_y$ and $L \in \mathcal{L}$, $\alpha(A) \leq \overline{A \cap L}$.

Proof: Necessity (No game needed for this part!):
(i) and (ii): Obvious! (iii) and (iv): Consider $A \in \mathcal{S}_y$ and a lower set L which intersects A. Define a distortion y_β of y by



$$y_\beta(a) = \begin{cases} \beta \text{ if } a \in A \cap L \\ y(a) \text{ otherwise.} \end{cases}$$

Observe that for $\beta < \alpha_y(A)$ sufficiently close to $\alpha_y(A)$, isotonicity is preserved. Thus

$$\frac{d}{d\beta}\|y_\beta - y_0\|^2 \text{ evaluated at } \beta = \alpha_y(A) \text{ is } \leq 0\,, \text{ hence}$$

$$\sum_{a \in A \cap L} 2W(a)\big(\beta - y_0(a)\big) \leq 0 \text{ at } \beta = \alpha_y(A) \text{ and}$$

$$W(A \cap L)\alpha_y(A) - \sum_{a \in A \cap L} W(a) y_0(a) \leq 0 \text{ follows,}$$

hence $\alpha_y(A) \leq \overline{A \cap L}$. This proves (iv).

In case $A \cap L = A$, e.g. if $L = \Omega$, also a slight *increase* of the value of y on A preserves isotonicity. Hence, the derivative from before must vanish. Thus $\alpha_y(A) = \overline{A}$, i.e. (iii) holds.

Sufficiency: For this part we appeal to a game. Let **N**=NATURE choose an isotone valuation x (write $x \in \mathcal{I}$) and **O**=OBSERVER choose any valuation y. Consider the game of updating with updating gain $\mathrm{U} = \mathrm{U}_{|y_0}$ given by

$$\mathrm{U}_{|y_0}(x, y) = \|x - y_0\|^2 - \|x - y\|^2 .$$

Such games are in (Nash-) equilibrium and both players have unique optimal strategies which coincide. Since, for every $x \in \mathcal{I}$, $\max_y \mathrm{U}_{|y_0}(x, y) = \|x - y_0\|^2$, the optimal strategy for **N** is the sought isotone regression.

Now assume that our y satisfies (i)–(iv). Put, in anticipation, $x^* = y^* = y$. We have to verify optimality of these strategies. For this we verify Nash's inequalities, which express that **O**'s value under the strategy y^* is equal to **N**'s value under the strategy x^*.

As one inequality is obvious (by the general minimax ineq.), we need only check the other direction, i.e. that

$$\inf_{x\in\mathcal{I}} \mathrm{U}_{|y_0}(x,y^*) \geq \|x^*-y_0\|^2$$

or, for every $x\in\mathcal{I}$, that

$$\|x-y_0\|^2 - \|x-y^*\|^2 \geq \|x^*-y_0\|^2\,,$$

and, as $x^*=y^*$, this inequality is equivalent to

$$\langle x-y^*, y^*-y_0\rangle \geq 0\,. \qquad (2)$$

This has a clear geometric interpretation (...). We shall analyze it analytically and, for a given $x\in\mathcal{I}$, rewrite (2) as

$$\begin{aligned}
&\sum_{A\in\mathcal{S}_y}\sum_{a\in A} W(a)\big(x(a)-\overline{A}\big)\big(\overline{A}-y_0(a)\big)\\
&=\sum_{A\in\mathcal{S}_y}\sum_{a\in A} W(a)x(a)\big(\overline{A}-y_0(a)\big)\\
&=\sum_{A\in\mathcal{S}_y}\sum_{a\in A} x(a)\delta_A(a)\ \text{(by definition of } \delta_A)\,.
\end{aligned}$$

Fix $A \in \mathcal{S}_y$. We shall prove that $\sum_{a \in A} x(a)\delta_A(a) \geq 0$. As

$$\delta_A(a) = W(a)\big(\overline{A} - y_0(a)\big)\ ; \ \text{for } a \in A$$

and by (iii) and (iv), we have $\sum_{a \in A} \delta_A(a) = 0$ as well as $\sum_{a \in A \cap L} \delta_A(a) \leq 0$ for every lower set L.

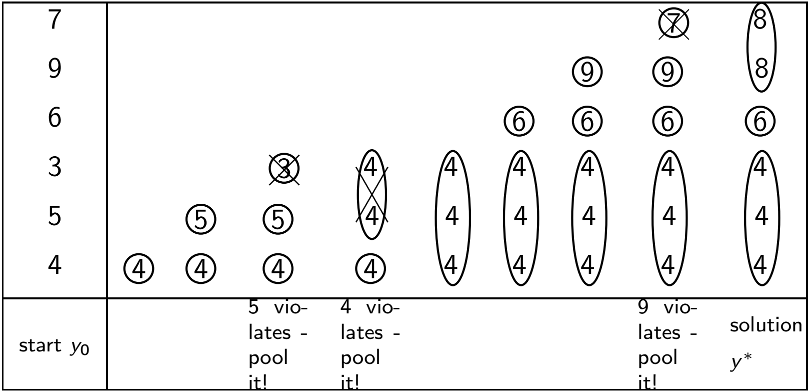
Now, let $\alpha_0 < \alpha_1 < \cdots < \alpha_n$ be the values assumed by x and write x in the form

$$x = \alpha_n - \sum_{i=1}^{n} (\alpha_i - \alpha_{i-1}) 1_{L_i} \tag{3}$$

with L_i the lower set $L_i = \{x < \alpha_i\}$. From this representation and above observations, the desired result $\sum_{a \in A} x(a)\delta_A(a) \geq 0$ follows. Putting things together, sufficiency is proved. □

This is not the end! More results are needed, especially we want to know how to *construct* the sought regression.

A toy example: linear order with equal weights



7								~~7~~	8
9							9	9	8
6						6	6	6	6
3			~~3~~	~~4~~	4	4	4	4	4
5		5	5	~~4~~	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4
start y_0			5 violates - pool it!	4 violates - pool it!				9 violates - pool it!	solution y^*

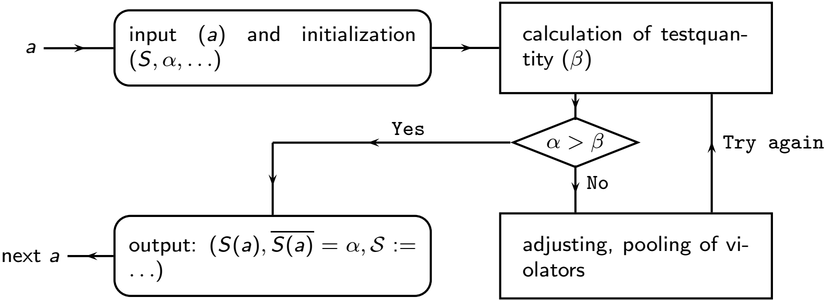
An algorithm for a general co-tree

For an arbitrary co-tree Ω, there is an efficient algorithm, a pool adjacent violaters algorithm, which rather precisely follows the ideas of the toy example.

We construct, incrementally, a decomposition $\mathcal{S}$ of Ω. At each step, $\hat{a}$ denotes the set in the relevant part of $\mathcal{S}$ which contains a. We start from the bottom, the minimal elements, and continue up the co-tree. At each stage the basic subroutine (see next slide) is called with focus on a node a for which all objects related to nodes further down the co-tree have already been calculated. Aim is to:

- calculate $S(a) = \hat{a}$ as if the co-tree was $a^{\downarrow} = \{b|b \leq a\}$
- calculate $\alpha(a) = \overline{S(a)}$
- dismiss all previous $S(b)$'s with $b < a$ which are "swallowed up" by the new set $S(a)$.

So then, here is the basic subroutine:



a
input (a) and initialization ($S, \alpha, \ldots$)
calculation of testquantity (β)
$\alpha > \beta$
Yes
No
Try again
adjusting, pooling of violators
output: $(S(a), \overline{S(a)} = \alpha, \mathcal{S} := \ldots)$
next a

- at time of input: $S(b)$'s and $\alpha(a) = \overline{S(b)}$'s with $b < a$ known; the $S(b)$'s constitute a decomposition of $a^{\downarrow} \setminus \{a\}$;
- initialization: $S := \{a\}, \alpha := y_0(a)$, $\mathcal{D} := \emptyset$ (placeholder for sets to be dismissed), $\partial := a^{-}$, set of nearest neighb. $b < a$;
- testquantity: $\beta := \max_{b \in \partial} \overline{S(b)}$
- adjusting, pooling: $\partial_0 = \{b \in \partial | \overline{S(b)} = \beta\}$, $S := S \cup \bigcup_{b \in \partial_0} S(b)$; $\alpha := \overline{S}$; $\partial := (\partial \setminus \partial_0) \cup \bigcup_{b \in \partial_0} \partial(S(b))$; $\mathcal{D} := \{S(b) | b \in \partial_0\}$
- output: $S(a) = S$, $\alpha(a) = \alpha$, $\mathcal{S} := \big(\mathcal{S} \setminus \mathcal{D}\big) \cup \{S\}$; $\partial(S(a)) = \partial$.

regarding the proof

The proof that the algorithm works, i.e. that the monotonicity (obvious) and the boundedness (more difficult) properties hold for the final decomposition $\mathcal{S}$, depends on some properties of conditional expectations and on co-trees. It is not shown here.

We end by showing a surprising connection between the problems of universal coding and isotone regression.

From isotone regression to universal coding

Theorem. With $N(a)$ for the number of $b \in \Omega$ with $b \leq a$, take as prior the valuation y_0 given by

$$y_0(a) = N(a) \ln N(a) - \sum_{b \in a^-} N(b) \ln N(b)$$

and denote by y^* the isotone regression of y_0, providing Ω with the uniform distribution. Then the universal code κ^* is:

$$\kappa^*(a) = y^*(a) + \ln Z \text{ where } Z = \sum_{a \in \Omega} \exp(-y^*(a)) .$$

Presently, the proof is via an algorithm developed for the calculation of κ^* (joint work with Henrik Densing Petersen). The key is that the model $\mathcal{A}$ is a simplex and that, for the distribution P associated with y_0 ($P(a)$ =constant$\cdot \exp(-y_0(a))$), $\mathrm{D}(Q\|P)$ is constant for every extremal element in $\mathcal{A}$.

Now, to conclusions:

Can we conclude that: yes, here is a sound theory, worth while developing further which has a proven potential of incorporating basic classical optimization problems of Shannon Theory as well as going beyond that?

I hope so! Relations to other theories of learning, of complexity, of MDL, ... (with several developers present here) should be investigated, further consolidation carried out, e.g. thorough exposition of the basic theory (on its way), axiomatization (initiated), new developments (e.g. re algorithms), more applications ...

Acknowledgments: To Boris Ryabko (discussions re universal prediction, especially in order structures), Henrik Densing Petersen (co-author re algorithm for co-trees), NN (referee of the paper with HDP who pointed to the possible connection to isotone regression), Peter Harremoës (step in proof of identification theorem) and, then to WITMSE-organizers for providing a platform of exchange which, now for the third time, has allowed me to present ideas to a group of specialists.