BEYOND SHANNON – EXAMPLES FROM GEOMETRY, INFORMATION THEORY AND STATISTICS

Flemming Topsøe

University of Copenhagen
Department of Mathematical Sciences
Universitetsparken 5, dk-2100 Copenhagen, Denmark, topsoe@math.ku.dk

ABSTRACT

In previous contributions to WITMSE, [1] and [2], an abstract theory of cognition, inspired by information theory but going beyond classical Shannon theory in certain respects was outlined. See also [3]. Here, we continue the work by presenting three concrete problems: Sylvester's problem from geometric location theory, a problem of universal coding from information theory and the problem of isotone regression from statistics. At first, we focus on non-technical, philosophically oriented considerations. A more complete analysis of isotone regression follows and finally we point out a surprising connection between this problem and the one from universal coding.

1. THREE PROBLEMS

First geometry: In 1857 Sylvester wrote "It is required to find the least circle which shall contain a given system of points in the plane." In fact, this is the full text of [4]! Thus, if X denotes the set of points in the plane, $\|\cdot - \cdot\|$ Euclidean distance and $\mathcal{P} \subseteq X$ a given system – here assumed finite – of points in X, we seek a point $y = y^*$ in X which minimizes the quantity

$$\max_{x\in\mathcal{P}} \|x - y\|. \tag{1}$$

For the two remaining problems, $\Omega = (\Omega, \leq)$ denotes a finite partially ordered set provided with a *weight function* W. Little is lost if you take W to be the uniform distribution (and this will be assumed if no special mention of W is made). A real-valued function f on Ω is *isotone* if, for $a, b \in \Omega$, the implication $a \leq b \Rightarrow f(a) \leq f(b)$ holds. And f is *antitone* if $-f$ is isotone.

The problem from information theory which we shall deal with concerns the *model* $\mathcal{A}$ of all antitone probability distributions over Ω. Requested is the distribution $y = y^*$ which best represents $\mathcal{A}$ in the sense that

$$\sup_{x\in\mathcal{A}} \mathrm{D}(x\|y) \tag{2}$$

is minimized. Here D stands for *Kullback-Leibler divergence*, i.e. $\mathrm{D}(x\|y) = \sum_{a\in\Omega} x(a) \ln \frac{x(a)}{y(a)}$. This is a problem of *universal prediction*.

The corresponding problem of *universal coding* is to find a suitable *code length function* (in the sequal simply a *code*), κ^*, which can be taken as the base for actual coding of observations from a source emitting independent outputs from Ω, generated by a distribution known only to lie in $\mathcal{A}$. Appealing to standard information theoretical insight, the sought *universal code* is κ^* given from y^* by $\kappa^*(a) = \ln \frac{1}{y^*(a)}$ for $a \in \Omega$ (the good sense of this also involves an idealization and a replacement of logarithms to the base 2 with natural logarithms). Our codes satisfy *Kraft's equality*: $\sum_{a\in\Omega} \exp\big(-\kappa(a)\big) = 1$.

As our final problem we take *isotone least squares regression* (below just *isotone regression*), an important problem from statistics. Given is a real-valued function y_0 on Ω, referred to as a *valuation*. Sought is the isotone valuation $y = y^*$ which is closest in mean-squared norm to the given valuation y_0. Thus, we should minimize

$$\|y_0 - y\|^2 = \sum_{a\in\Omega} W(a)|y_0(a) - y(a)|^2 \tag{3}$$

subject to a requirement on y of isotonicity. Just as with the two previous problems, existence and uniqueness of the sought object is pretty evident. We refer to it as the *isotone regression of* y_0 (or just the *isotone regression*).

2. A COMMON FRAMEWORK

There exists a common framework which allows an efficient treatment of problems as those presented and of many others – e.g. from information theory, one could point to problems of maximum entropy determination, information projections and capacity determination. The reader is referred to [1] and [2] (or to a more comprehensive study, not yet in final form). Rather than spending time here on technicalities, we shall emphasize some features of the underlying theory as seen in the light of the three problems above.

The problems presented are all *optimization problems*. The first two are quite similar, technically. Euclidean distance stands out for the first, Kullback-Leibler divergence for the second. One should, however, note that optimization as in (1) and (2), does not uniquely tell us which are the basic quantities as any strictly increasing function of the appearing quantities could also be used. As we

shall argue below – and not all that surprising – squared Euclidean distance is adequate for the first problem and Kullback-Leibler divergence itself for the second.

A guiding principle for the choice of appropriate basic quantities is that – as recognized since long in optimization theory and convex analysis – one benefits from treating along with a given problem, also a *dual* problem. For this to work out conveniently, one needs certain strict relationships to hold which essentially involve conditions of linearity or affinity. Theoretically, introductory considerations can be carried out without imposing such strict conditions, cf. [1] and [2]. However, when it comes to actually treating concrete problems of interest, you need to be more specific.

In order to motivate necessary restrictions for a successful model building, we claim that the "two-ness" of duality considerations is best expressed by choosing a game-theoretical setting involving certain asymmetric *two-person zero-sum games*. For these games, the players have quite different roles. The first player, considered female, is conceived as "*Nature*". Nature chooses a strategy which reflects "*truth*", whereas the second player is a much more easily understood being, "*you*" or "*Observer*" – a mere mortal person, male we reckon, seeking the truth but restricted to "*belief*". Analyzing these thoughts, you find that though tempting to imagine Nature as a rational being reflecting "*absolute truth*", really, this is naive and what is involved is more sensibly thought of as another side of yourself. The "*zero-sumness*" of the games you are led to consider express an insight consistent with ideas of Jaynes from the mid-fifties, cf. [5], viz. that acting in a way which would contradict the zero-sum character would reflect that "you have known something more" and, therefore, your model building would be incomplete and should be adjusted.

An essential restriction in our model building then is that the games considered should, typically, be in *equilibrium*, i.e. the *minimax* and *maximin* values should coincide. In many cases this is not so at first sight. E.g., for the two first problems, where a minimax-value is sought, we find that the corresponding maximin-value is uninformative, indeed it vanishes identically. This may be remedied if suitable extensions of the allowed strategise for Nature can be deviced. For the two problems pointed to, this can be achieved by allowing *randomized strategies* for Nature (and, regarding (1), replacing norm by squared norm). In this way a common game theoretical base for the treatment of these problems can be found. This also applies to the third problem, though it is of a different type. There it pays to consider the given valuation y_0 as a parameter, cf. Section 3.

One has to be realistic as to what can be expectecd of a common theoretical base. In fact, though problems we are able to deal with typically have unique solutions, e.g. none of the three concrete problems considered allow solutions in closed form. One has to be satisfied with numerical algorithms or turn to special cases where solutions can be written down in closed form or, more realistically, where finite state algorithms of low complexity leeds to the solution. Such algorithms are special. Often Galois theory shows that even rather "small" problems have solutions which cannot be expressed quantitatively using the basic algebraic operations applied to the natural quantitative specifications of the problems.

Thus, an appeal to game theory does not in itself lead to solutions of the problems at hand. But it does help to characterize what is required of a solution. Such results of *identification* are often derived from an application of the *saddle-value inequalities* now associated with Nash's name. An example of this follows in the next section.

The overall theme of our investigations, that of establishing a useful theoretical base going "beyond Shannon", has been pursued by several authors in one way or another and appears right now to be gaining momentum, cf. also [6]. Shannon himself was aware of the need to broaden the theory he had initiated, e.g., in 1953 he writes "It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field", cf. [7].

3. ISOTONE REGRESSION

Let us leave the airy considerations of the foregoing section and turn to a closer study of isotone regression. The key to a game-theoretical formulation is the binary function $\mathrm{U}_{|y_0} = \mathrm{U}_{|y_0}(x, y)$ given by

$$\mathrm{U}_{|y_0}(x, y) = \|x - y_0\|^2 - \|x - y\|^2 . \qquad (4)$$

This is interpreted as the *updating gain*, when the *prior* y_0 is updated by Observers choice of the *posterior* y, assuming that the strategy chosen by Nature is x. In (4), x runs over the set X of all isotone valuations. These are the strategies of Nature. The strategies of Observer may be taken to be the set of all valuations, but it may also be restricted to X.

If Nature chooses x, the best response by Observer is also to choose x. The resulting value of $\mathrm{U}_{|y_0}$ will then be $\|x - y_0\|^2$ and it follows that the *optimal strategy* for Nature is to choose the sought isotone regression.

Comparing with Section 3 of [2], you realize that all conditions stated there are fulfilled. In particular, the squared norm satisfies the *compensation identity* (13) of [2]. From Theorems 2 and 3 of [2], it follows that Nature and Observer both have unique optimal strategies x^* and y^* and that these strategies coincide: $x^* = y^*$. A key problem is, therefore, to determine this common *bi-optimal strategy*. A suitable result of identification for this problem will now be derived.

Let $x^* = y^*$ be a given isotone valuation, from the outset not known to be the sought bi-optimal strategy. Then, by the general theory, this *is* the sought strategy if and only if the non-trivial part of Nash's inequalities holds:

$$\mathrm{U}_{|y_0}(\xi, y^*) \geq \|x^* - y_0\|^2 \text{ for every } \xi \in X . \qquad (5)$$

Expressing squared norm via the associated inner product defined by $\langle f, g\rangle = \sum_{a\in\Omega} W(a)f(a)g(a)$, and recalling that $y^* = x^*$, we transform the requirement to the

condition

$$\langle \xi - x^*, x^* - y_0 \rangle \geq 0 \text{ for every } \xi \in X\,. \qquad (6)$$

For the further analysis, we note that any valuation f induces a special decomposition of Ω, denoted $\mathcal{S}_f$. The sets in $\mathcal{S}_f$ are the *maximal connected sets of f-constancy*, i.e. the connected subsets of Ω on which f assumes the same value and which are maximal with respect to these properties. Further, we note that in case f is isotone, the sets in $\mathcal{S}_f$ are partially ordered in a natural way, viz. by defining $A < B$ to mean that, firstly, $A \neq B$ and, secondly, that $a < b$ for some (a, b) with $a \in A$ and $b \in B$.

Any valuation f is specified by the decomposition $\mathcal{S}_f$ and the associated function values. For the isotone regression only the decomposition $\mathcal{S}^* = \mathcal{S}_{x^*}$ needs to be specified as the function values can then be identified as *conditional averages*. Indeed, denoting by $\overline{A}_{|y_0}$ (or simply $\overline{A}$) the conditional average of the prior y_0 over A, i.e.

$$\overline{A} = \sum_{a \in A} W(a|A) y_0(a) = \frac{1}{W(A)} \sum_{a \in A} W(a) y_0(a)\,, \quad (7)$$

then, for the isotone regression x^*,

$$\text{for all } A \in \mathcal{S}^*,\ x^* = \overline{A} \text{ on } A\,. \qquad (8)$$

In fact, this is easy to prove by a differential argument based on the considerations of valuations obtained from x^* by varying the value on A and keeping other values fixed. The argument can be refined, yielding another central property of $\mathcal{S}^*$, *boundedness*. This is the property, that for each $A \in \mathcal{S}^*$ and each *lower set* L which intersects A – a lower set being a set such that $a < b \in L$ implies $a \in L$ – it holds that

$$\overline{A}_{|y_0} \leq \overline{A \cap L}_{|y_0}\,. \qquad (9)$$

Theorem 1 (Identification) *Let x be a valuation with associated decomposition $\mathcal{S}$ and associated function-values $\alpha(A)$; $A \in \mathcal{S}$. Then a necessary and sufficient condition that $x = x^*$, the sought isotone regression of y_0, is that the following conditions hold: (i) [ordering]: $\mathcal{S}$ is partially ordered; (ii) [monotonicity]: if $A, B \in \mathcal{S}$ and $A < B$, then $\alpha(A) < \alpha(B)$; (iii) [proper values]: $\alpha(A) = \overline{A}_{|y_0}$ for each $A \in \mathcal{S}$ and (iv) [boundedness]: for every $A \in \mathcal{S}$ and every lower set L which meets A,* (9) *holds.*

Proof A proof that the stated conditions are necessary was indicated above. In order to establish sufficiency, assume that the conditions hold. The essential point is to establish the validity of (6). An indication has to suffice: First, write the inner product in (6) as a sum and then split the sum in a sum over each of the classes in $\mathcal{S}$. For the essential argument we may assume that $\mathcal{S} = \{\Omega\}$. Consider a fixed isotone valuation ξ. Let $\alpha_0 < \cdots < \alpha_n$ be the values assumed by ξ and write ξ in the form

$$\xi = \alpha_n - \sum_{i=1}^{n} (\alpha_i - \alpha_{i-1}) 1_{\{\xi < \alpha_i\}}\,. \qquad (10)$$

Consider the valuation δ defined by

$$\delta(a) = W(a)\big(\overline{\Omega}_{|y_0} - y_0(a)\big)\,. \qquad (11)$$

Then $\sum_{a \in \Omega} \delta(a) = 0$ and $\sum_{a \in L} \delta(a) \leq 0$ for each lower set L. By (10) it follows that $\sum_{a \in \Omega} \xi(a)\delta(a) \geq 0$, which is the required result. □

A discussion is in order. The reasoning demonstrates that though Nash's inequalities in principle contain the essentials, this may be in a somewhat concealed form and require quite a bit of extra work until a transformation into a manageable form has been obtained. We may also note that though the identification result is easy to use in examples of moderate size – see, e.g. the butterfly set discussed in Figures 1 and 2 – the necessary checking of condition (iv) of Theorem 1 may be forbidding for more elaborate partially ordered sets as the number of lower sets may be of exponential size in the number of parameters necessary to specify the partial order.

Thus one should ask for further results aiming at the actual construction of the isotone regression. Often, this is not feasible but, fortunately, the problem dealt with is one for which satisfactory results exist, cf. [8] and references referred to there, especially [9].

The problem is greatly simplified if we restrict attention to tree-like structures. We shall assume from now on that Ω is a *co-tree*, i.e. right sections are well ordered (or, equivalently, the reverse partial ordering is a tree). This is a significant simplification. For one thing, lower sets can then be represented as disjoint unions of left sections, thus the checking involved in the identification theorem is feasible, as only left sections need to be checked when checking the boundedness property.

Without being very specific, the existence of an efficient algorithm for the determination of the isotone regression is indicated below. The ideas are contained in the identification theorem. As it turns out, if you focus on all properties *except* boundedness and aim at construction of the classes in $\mathcal{S}^*$ "from below", then an argument (not shown here) will reveal the fact that boundedness is verified automatically. The build-up from below exploits the idea of searching for violation of the monotonicity requirement followed by pooling of adjacent already constructed classes if a violation occurs. This idea is well known from the statistical literature on isotone regression and there referred to as *pooling of adjacent violators* (PAV). The example of a linear ordering as displayed in Figure 3 explains better than many words how the intended algorithm works. And generalizing to a arbitrary co-tree presents no further problems.

4. A SURPRISING CONNECTION

Consider again the problem of universal coding. The assumption, still in force, that Ω is a co-tree, implies that the model $\mathcal{A}$ is a simplex with the uniform distributions over left sections as extremal elements. Denote by $a^{\downarrow}$ the left section determined by a, by $N(a)$ the number of elements in $a^{\downarrow}$ and by U_a the uniform distribution over $a^{\downarrow}$. Further, let a^- be the set of immediate predecessors of a.

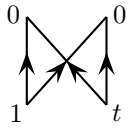
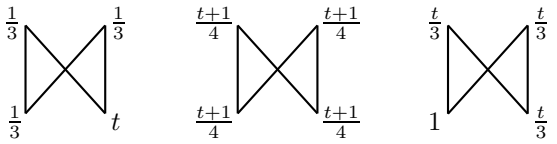


Figure 1. Butterfly with valuation depending on a parameter t.



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

Figure 2. Isotone regression for the butterfly, depending on the value of the parameter t.

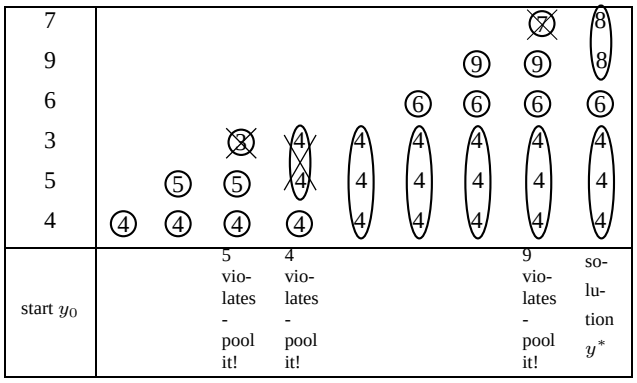


Figure 3. Algorithmic construction of the isotone regression for a 6-element linear order with valuation $y_0 = (4, 5, 3, 6, 9, 7)$.

It is easy to check that there exists a distribution Q, not necessarily isotone, such that $\mathrm{D}(U_a \| Q)$ i s independent of a. Indeed, Q is proportional to μ given by

$$\mu(a) = \frac{\prod_{b \in a^-} N(b)^{N(b)}}{N(a)^{N(a)}},\ a \in \Omega\,. \tag{12}$$

Theorem 2 *Let y_0 be the valuation given by*

$$y_0(a) = \ln \frac{1}{\mu(a)}\,;\ a \in \Omega\,, \tag{13}$$

and denote by y^* the isotone regression of y_0. Then the universal code κ^* is obtained from y^* by normalization, i.e., for a suitable constant, c, $\kappa^*(a) = y^*(a) + c$ for every $a \in \Omega$.

This follows, in a rather roundabout manner, by comparing [10] with results from isotone regression. A more direct proof may well exist.

The special distribution Q with constant divergence to a set of elements which generate the relevant model may be called a *Sylvester point*. It is easy to see that the universal predictor can be obtained as the information projection of Q on the model $\mathcal{A}$. Analogous features apply to Sylvester's problem, though the existence of a Sylvester point in that setting is only possible in very special cases, e.g. for the illuminating case of a three-element model $\mathcal{P}$.

5. ACKNOWLEDGMENTS

This goes to Henrik Densing Petersen, cf. [10], to a referee of [10] who encouraged a comparison with [8] and to Peter Harremoës for a discussion of the boundedness property of Theorem 1.

6. REFERENCES

[1] F. Topsøe, "Cognition beyond Shannon," in *Proceedings of the third Workshop on Information Theoretic Methods in Science and Engineering, Tampere*, 2010, available from http://sp.cs.tut.fi/WITMSE10/Proceedings/index.html.

[2] F. Topsøe, "Cognition and Inference in an Abstract Setting," in *Proceedings of the fourth Workshop on Information Theoretic Methods in Science and Engineering, Helsinki*. 2011, Report C-2011-45, pp. 67–70, University of Helsinki, available from http://www.helsinki.fi/witmse2011/proceedings.html.

[3] F. Topsøe, "Game Theoretical Optimization inspired by Information Theory," *J. Global Optim.*, pp. 553–564, 2009.

[4] J. J. Sylvester, "A question in the geometry of situation," *Quarterly Journal of Pure and Applied Mathematics*, vol. 1, pp. 79, 1857.

[5] E. T. Jaynes, *Probability Theory - The Logic of Science*, Cambridge University Press, Cambridge, 2003.

[6] W. Szpankowski, "Algorithms, Combinatorics, Information, and Beyond," *IEEE Information Theory Society Newsletter*, vol. 62, pp. 5–20, 2012.

[7] C. Shannon, "The lattice theory of information," *IRE professional Group on Information Theory*, vol. 1, pp. 105–107, 1953.

[8] P.M. Pardalos and G. Xue, "Algorithms for a Class of Isotonic Regression Problems," *Algorithmica*, vol. 23, no. 3, pp. 211–222, Mar. 1999.

[9] C. I. C. Lee, "The Min-Max Algorithm and Isotonic Regression," *Ann. Statist.*, vol. 11, pp. 467–477, 1983.

[10] H. D. Petersen and F. Topsøe, "Computation of universal objects for distributions over co-trees," under publication in *IEEE Trans. Inform. Theory*, 2012.