ISIT'2013 1569713101

Universal Coding in Order Structures and Isotone Regression

Flemming Topsøe University of Copenhagen
Department of Mathematical Sciences
Universitetsparken 5,
2100 Copenhagen, Denmark
Email: topsoe@math.ku.dk

***Abstract*—The two problems indicated in the title are studied and a connection between them established.**

***Index Terms*—Universal coding, isotone regression, pool structure, PAV-algorithm.**

I. Introduction

Let $\Omega = (\Omega, \leq)$ be a finite partially ordered set.

The associated universal coding problem is to determine the universal code for the model $\mathcal{P}_a$ of antitone probability distributions over Ω, i.e. for the model of all distributions P such that $a \leq b \Rightarrow P(a) \geq P(b)$ for $a, b \in \Omega$. Universality is understood in the usual minimax-redundancy sense.

For isotone regression, extra structure is needed, viz. a strictly positive weight function w on Ω and a real-valued function y_0 on Ω, referred to as a *valuation* and thought of as *prior data*. The problem of *isotone regression*[1], is to determine the isotone valuation y^* which is closest to y_0 in L^2-norm.

The problems have unique solutions. However, solutions in closed form only exist in special cases. The problems were first solved for a linear order. Ryabko, [1], considered universal coding and Ayer, Brunk, Ewing, Reid and Silverman, [2], isotone regression. For this case, the universal coding problem is of an exceptional simple type, and Ryabko could devise a closed formula for its solution. For the isotone regression problem the situation is different and an algorithmic solution, the PAV-*algorithm* (*pool adjacent violators*) was developed by Ayer et al.

The study of the two types of problems has progressed quite differently. Apparently, there is to date only one contribution to the further theoretical development of the universal coding problem in order structures, viz. the authors recent joint work with Petersen, [3]. For the other and, admittedly, more important problem, there is a rich statistical literature, with several variants of the PAV-algorithm and numerous widespread applications – though to date not within information theory. The development may be traced from Pardalos and Xue, [4]. Newer results are discussed e.g. in de Leeuw et al, [5].

The study [3] deals with a general co-tree Ω. The algorithm developed had a certain similarity with the PAV-algorithm and this led an anonymous referee to suggest to look further into this. Though quite different in nature, a direct connection between the two problems does indeed exist, cf. Theorem 4.1, our main result. It concerns the case of co-trees. The proof relies on identification results from sections II and III.

Isotone regression problems are more basic than those of universal coding in order structures as treated here. To justify this point of view, note that the universal coding problem completely changes character if you consider trees in place of co-trees. Indeed, even for quite simple trees, there may not exist an algorithm which calculates the solution in finitely many steps. This impossibility may be realized by applying some Galois theory[2]. For isotone regression, no difficulties are involved when you change the direction of the order and go from co-trees to trees. Indeed, a change of sign in the prior data will transform the one situation to the other. These remarks indicate that the connection found between universal coding and isotone regression is special with little room for further expansion[3].

In Section V the PAV-algorithm is presented. Taken together with Section II, this may be read as an introduction to a fascinating area of computational statistics.

As pointed out by Csiszár, cf. [6], least squares optimality criteria and minimum divergence criteria may both be viewed as emerging from a common axiomatic setting. Thus (isotone) least squares regression and universal coding have something in common. This point of view may also be arrived at from game theoretical considerations as those of the present contribution. But this does not point to coincidence of the two approaches. Theorem 4.1 is, therefore, a singular somewhat surprising instance where quite different optimality criteria, suitably applied, lead to the same result.

II. Basics of isotone regression

In (most of) this section, $(\Omega, \leq, w, y_0)$ is an arbitrary finite, partially ordered set provided with a strictly positive weight function, w, and a valuation[4] y_0, referred to as the *prior*. Having a graphical representation of Ω in mind (via the *Hasse diagram*), we refer to the elements of Ω as *nodes*.

Let W be the measure with point masses given by w. Conditional averages of the prior over non-empty subsets A

[1] in the literature mainly called *isotonic regression*

[2] unpublished joint work with Peter Harremoës.

[3] One possibility though, as also briefly pointed to in [3], is to go beyond Shannon theory and work with Bregman divergencies instead of classical Kullback-Leibler divergence.

[4] A *valuation* (on Ω) is here understood to be a real-valued function on Ω.

of Ω play a special role and we introduce the notation

$$\overline{A} = \frac{1}{W(A)} \sum_{a \in A} w(a) y_0(a) \,.$$

A valuation y is *isotone*, respectively *antitone* if the implication $a \leq b \Rightarrow y(a) \leq y(b)$, respectively $a \leq b \Rightarrow y(a) \geq y(b)$ holds for all $a, b \in \Omega$. The class of isotone, respectively antitone valuations is denoted $\mathcal{I}$, respectively $\mathcal{A}$.

The *isotone regression* of y_0 is the, necessarily unique, isotone valuation y^* given by $y^* = \operatorname{Arg\,min}_{y \in \mathcal{I}} \|y - y_0\|^2$ where $\|y - y_0\|^2 = \sum_{a \in \Omega} w(a) |y(a) - y_0(a)|^2$. We write $\mathrm{Reg}_{\min} = \|y^* - y_0\|^2$ for the minimum value.

Concepts introduced are connected with a two-person zero-sum game of *updating* which has has isotone valuations (x's) as strategies for Player I, the minimizer, and arbitrary valuations (y's) as strategies for Player II, the maximizer. As objective function we take $\mathrm{U}_{|y_0}$ defined by

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

interpreted as *updating gain* for Player II when replacing the prior y_0 with the *posterior* y, assuming that Player I has chosen the strategy x. The minimax value of the game (Player I's value) is $\mathrm{Reg}_{\min}$ and the unique optimal strategy for Player-I is the isotone regression y^*.

A *lower set*, respectively an *upper set*, is a set $L \subseteq \Omega$, respectively $U \subseteq \Omega$, such that the implication $a \leq b, b \in L \Rightarrow a \in L$, respectively $a \leq b, a \in U \Rightarrow b \in U$ holds for all $a, b \in \Omega$. The paving[5] of lower sets is denoted $\mathcal{L}$, the paving of upper sets $\mathcal{U}$. A *level set* is an intersection of an upper and a lower set. A non-empty level set A is uniquely determined by the sets $A_{\max}$ and $A_{\min}$ of maximal, respectively minimal nodes in A and consists of all $b \in \Omega$ for which $a \leq b \leq c$ for some $a \in A_{\min}$ and $c \in A_{\max}$. By $\mathcal{C}$ we denote the paving of connected, non-empty level sets.

Let x be a valuation on Ω and denote by $\mathcal{C}_x$ the decomposition of Ω into maximal, connected sets of x-constancy, and denote by α_x the function on $\mathcal{C}_x$ which provides the appropriate values of x. Thus, for $A \in \mathcal{C}_x$, $x(a) = \alpha_x(A)$ for every $a \in A$. Clearly, x is uniquely characterized by the *representation* $(\mathcal{C}_x, \alpha_x)$. The sets in $\mathcal{C}_x$ are the *components* of x. If x is either isotone or antitone, $\mathcal{C}_x \subseteq \mathcal{C}$.

Theorem 2.1: [Identification of y^*]. Let y be a valuation on Ω. Then $y = y^*$ if and only if

- $\mathcal{C}_y$ is partially ordered in the natural order;
- α_y is strictly increasing on $\mathcal{C}_y$;
- for $A \in \mathcal{C}_y$ and for every pair (L, U) of a lower set which intersects A and an upper set which intersects A, the following double-inequality holds:

$$\overline{A \cap U} \leq \overline{A} \leq \overline{A \cap L} \,. \tag{1}$$

Regarding the first condition, $A \leq B$ (with $A \in \mathcal{C}_y$, $B \in \mathcal{C}_y$) means that $a \leq b$ for some $a \in A$, $b \in B$. The condition is that, unless $A = B$, we cannot have $a_1 \leq b_1$ and also $a_2 \geq b_2$ for nodes $a_1, a_2 \in A$, $b_1, b_2 \in B$. The last condition may be split in two:

$$\alpha_y(A) = \overline{A} \text{ for } A \in \mathcal{C}_y \,; \tag{2}$$

$$\alpha_y(A) \leq \overline{A \cap L} \text{ for } A \in \mathcal{C}_y \,, L \in \mathcal{L} \text{ with } A \cap L \neq \emptyset \,. \tag{3}$$

Proof: Assume first that $y = y^*$. Clearly then, the two first conditions hold. To verify the last, we verify (2) and (3). Consider intersecting sets $A \in \mathcal{C}_y$ and $L \in \mathcal{L}$. Put

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

Then, since y_β is isotone for $\beta \leq \alpha_y(A)$ sufficiently close to $\alpha_y(A)$, $\frac{\partial}{\partial \beta} \|y_0 - y_\beta\|^2 \leq 0$ when evaluated at $\beta = \alpha_y(A)$. Writing this out, it follows that $\alpha_y(A) \leq \overline{A \cap L}$. If $A \cap L = A$, y_β is also isotone for $\beta \geq \alpha_y(A)$ sufficiently close to $\alpha_y(A)$ and then also the reverse inequality is obtained. In this way (2) as well as (3), hence also (1), are verified.

Then assume that y satisfies the three conditions stated. Clearly, y is isotone. We shall show that y is an optimal strategy for Player I as well as for Player II in the updating game introduced in Section I. By standard game-theoretical considerations, this will imply that $y = y^*$. What we have to prove is, that for every $x \in \mathcal{I}$, $\mathrm{U}_{|y_0}(x, y) \geq \|y - y_0\|^2$. So assume that $x \in \mathcal{I}$ and rewrite the term $\|x - y_0\|^2$ in $\mathrm{U}_{|y_0}(x, y)$ as $\|x - y\|^2 + \|y - y_0\|^2 + 2\langle x - y, y - y_0 \rangle$ with $\langle \cdot, \cdot \rangle$ for standard inner product. We have to show that $\langle x - y, y - y_0 \rangle$ is non-negative. Writing it as a sum over the components of y, actually *all* summands are non-negative. To show this, let $A \in \mathcal{C}_y$ and use (2) to write the contribution from A as

$$\sum_{a \in A} w(a) \Big(x(a) - \overline{A} \Big) \Big(\overline{A} - y_0(a) \Big) \,,$$

an expression which can be written in the form

$$\sum_{a \in A} \delta(a) x(a) \,, \tag{4}$$

where, for $a \in A$, we have put $\delta(a) = w(a) \Big(\overline{A} - y_0(a) \Big)$. By (2) and (3),

$$\sum_{a \in A} \delta(a) = 0 \,, \tag{5}$$

$$\sum_{a \in A \cap L} \delta(a) \leq 0 \text{ for every } L \in \mathcal{L} \,. \tag{6}$$

Let $\alpha_0 < \alpha_1 < \cdots < \alpha_k$ denote the values assumed by x on A. Since x is isotone, each set $L_i = \{x < \alpha_i\}$ is a lower set. With $1_{\cdot}$ as notation for indicator functions, we have

$$x \cdot 1_A = \alpha_k \cdot 1_A - \sum_{i=1}^{k} (\alpha_i - \alpha_{i-1}) \cdot 1_{A \cap L_i} \,.$$

Inserting this for x in the expression (4), it follows readily from (5) and (6), that the expression is non-negative. Putting things together, $y = y^*$ follows. ∎

An example, adapted from Lee [7], is indicated in Fig.1. There, $\Omega = \{1, 2, 3, 4\}^2$ in the pointwise ordering. The prior y_0 is displayed and so are the four components of y^*. The corresponding four values of y^* can be obtained from (2). That

[5] A *paving* (in Ω) is a non-empty class of subsets of Ω, referred to as *stones*.

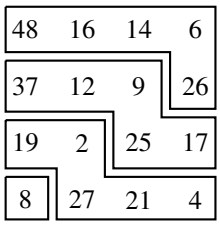

Figure 1. Product order with prior and components of the regression.

y^* is indeed the isotone regression of y_0 is easily checked by Theorem 2.1.

We note that though you may have a candidate for y^*, it can be a complex task to check this via Theorem 2.1 since there may be even exponentially many lower sets, measured relative to the size of Ω. If Ω has some special structure this may change. Thus, if Ω is a co-tree, checking of (3) is only necessary for lower sets which are left sections and hence checking in such cases is a task of linear complexity.

Finally, we comment on special facts and notation which apply to a co-tree Ω. Then, sets of the form $A_{\max}$ with $A \in \mathcal{C}$, contain only one node, the *top-node* of A, denoted τ_A. Certain issues are most conveniently discussed for antitone, rather than isotone valuations. The cone $\mathcal{A}_+$ of non-negative antitone valuations is a *simplicial cone* with the indicator functions of left sections as *basis*. More precisely, every $y \in \mathcal{A}_+$ can be written in one and only one way as a linear combination $y = \sum_{a\in\Omega} \alpha(a) 1_{a^\downarrow}$ with all coefficients $\alpha(a)$ non-negative – and any such combination defines a valuation in $\mathcal{A}^+$. By $\sigma(y)$, the *spectrum* of $y \in \mathcal{A}_+$, we understand the set of $a \in \Omega$ for which $\alpha(a)$ is positive.

The spectrum can be characterized in terms of the components of y. We note that $\mathcal{C}_y$ is a co-tree itself in the natural order. If $A \in \mathcal{C}_y$ is non-maximal, A^+ denotes the immediate successor of A. It is the unique component which contains the node τ_A^+.

Lemma 2.1: Let Ω be a co-tree and $y \in \mathcal{A}_+$. Then $\sigma(y) = \{\tau_A | A \in \mathcal{C}_y\}$.

Proof: The essential observation is, that with coefficients $(\alpha(A))_{A\in\mathcal{C}_y}$ given by

$$\alpha(A) = \begin{cases} y(A) \text{ if } A \text{ is maximal} \\ y(A) - y(A^+) \text{ otherwise}, \end{cases}$$

$y = \sum_{A\in\mathcal{C}_y} \alpha(A) 1_{\tau_A^\downarrow}$. Details are left to the reader. ∎

III. Basics of universal coding in co-trees

We repeat some introductory material from [3].

As indicated in the introduction, it is hopeless to look for efficient algorithms for general partially ordered sets or even for trees. The proper setting appears to be that of co-trees. Therefore, in this section, $\Omega = (\Omega, \leq)$ is a finite co-tree Thus, there is only one maximal node, the *top node* and every other node has a unique successor[6]. The successor of $a \in \Omega$ is denoted a^+. For $a \in \Omega$, $a^\downarrow = \{b | b \leq a\}$ is the *left section* of a. In general, $|\cdots|$ denotes "number of nodes in $\cdots$". We write $N(a)$ in place of $|a^\downarrow|$ and put $N'(a) = N(a) \ln N(a)$. By a^- we denote the set of *immediate predecessors* of a. Thus $N(a) = 1$ and $a^- = \emptyset$ if a is a minimal node of Ω.

The uniform distribution over a left-section $a^\downarrow$ is denoted U_a. As is easily seen, $\mathcal{P}_a$ is a simplex with the U_a's as extremal elements. Let $P \in \mathcal{P}_a$ and determine weights $(s_a)_{a\in\Omega}$ such that $P = \sum_{a\in\Omega} s_a U_a$. By the *spectrum* $\sigma(P)$ of P we understand the set of $a \in \Omega$ with $s_a > 0$.

A *code* is identified with a *code length function*, i.e. a valuation κ such that *Kraft's equality*, $\sum_{a\in\Omega} \exp\big(-\kappa(a)\big) = 1$ holds. The set of codes is denoted $\mathcal{K}$ and the subset of isotone codes, $\mathcal{K}_i$. The equations $P(a) = \exp\big(-\kappa(a)\big)$ and $\kappa(a) = \ln \frac{1}{P(a)}$ establish a well known and useful one-to-one correspondence between distributions and codes. For instance, we may talk about the distribution *associated* with a code. Under this correspondence, $\mathcal{P}_a$ and $\mathcal{K}_i$ are associated with each other. For $\kappa \in \mathcal{K}_i$, the *spectrum* of κ, $\sigma(\kappa)$, is defined as the spectrum of the associated distribution.

Average code length is denoted $\langle \kappa, P \rangle$, *Shannon entropy* by $\mathrm{H}(P)$ and *Kullback-Leibler divergence* by $\mathrm{D}(P\|Q)$. For $P \in \mathcal{P}$ and $\kappa \in \mathcal{K}$, *redundancy* $\mathrm{D}(P\|\kappa)$ is defined as $\mathrm{D}(P\|\kappa) = \mathrm{D}(P\|Q)$ with Q the distribution associated with κ. For $P \in \Omega$ and $\kappa \in \mathcal{K}$, the *linking identity* $\langle \kappa, P \rangle = \mathrm{H}(P) + \mathrm{D}(P\|\kappa)$ holds. The *fundamental inequality of information theory* is the observation that $\mathrm{D}(P\|Q) \geq 0$ with equality only for $P = Q$. Expressed in terms of redundancy, $\mathrm{D}(P\|\kappa) \geq 0$ with equality only when P and κ are associated with each other.

The *redundancy* of $\kappa \in \mathcal{K}$ with respect to $\mathcal{P}_a$ is $\mathrm{R}(\kappa) = \sup_{P\in\mathcal{P}_a} \mathrm{D}(P\|\kappa)$ and *minimax redundancy* is the quantity $\mathrm{R}_{\min} = \min_{\kappa\in\mathcal{K}} \mathrm{R}(\kappa)$. The minimum is achieved by a unique code, κ^*, the *universal code* (for $\mathcal{P}_a$). The distribution in $\mathcal{P}_a$ associated with κ^* we denote P^*. It is the *universal predictor*.

In analogy with Theorem 2.1, the following result holds, a special case of the well-known *Kuhn-Tucker theorem*:

Theorem 3.1: [Identification of κ^*]. Let $\kappa \in \mathcal{K}$. Then $\kappa = \kappa^*$ if and only if κ is isotone and, for some constant R,

$$\mathrm{D}(U_a\|\kappa) = R \text{ for } a \in \sigma(\kappa), \tag{7}$$

$$\mathrm{D}(U_a\|\kappa) \leq R \text{ for all } a \in \Omega. \tag{8}$$

When the conditions hold, $\mathrm{R}_{\min} = R$.

To us sufficiency is essential. Though well known, a quick proof is included: By convexity of $\mathrm{D}(\cdot\|\kappa)$[7] and by (8), $\mathrm{R}(\kappa) \leq R$. Now, for any $\rho \in \mathcal{K}$, we find, with $P = \sum_{a\in\Omega} s_a U_a$ the distribution associated with κ, that

$$\begin{aligned} \mathrm{R}(\rho) &= \sum_{a\in\sigma(P)} s_a R(\rho) \geq \sum_{a\in\sigma(P)} s_a \,\mathrm{D}(U_a\|\rho) \\ &= \sum_{a\in\sigma(P)} s_a \,\mathrm{D}(U_a\|\kappa) + \mathrm{D}(P\|\rho) = R + \mathrm{D}(P\|\rho) \end{aligned}$$

[6] We could allow several maximal nodes corresponding to a "co-forest", but as isotone regression problems as well as universal coding problems for co-forests are easily reduced to the corresponding problems for co-trees, we stick to the formally simpler setting.

[7] apply the linking identity and the fundamental inequality, cf. [8]

and it follows that $\mathrm{R}(\rho) \geq R$ with strict inequality if $\rho \neq \kappa$.

For the further development we note that there exists a special code which satisfies the conditions of Theorem 3.1, except that it need not be isotone. By definition, $\kappa_0 \in \mathcal{K}$ is a *Sylvester element* (here code)[8] for our universal coding problem if, for some constant R_0, the *Sylvester constant*, $\mathrm{D}(U_a\|\kappa_0) = R_0$ for all $a \in \Omega$.

Lemma 3.1: A unique Sylvester code κ_0 exists, given by

$$\kappa_0(a) = N'(a) - \sum_{b\in a^-} N'(b) + R_0\,, \tag{9}$$

valid for all $a \in \Omega$. R_0, the Sylvester constant, is

$$R_0 = \ln \sum_{a\in\Omega} \frac{\prod_{b\in a^-} N(b)^{N(b)}}{N(a)^{N(a)}}\,. \tag{10}$$

Proof: Clearly, κ_0 given by (9) and (10) defines a code. By the general identity

$$\sum_{b\leq a} \Big(f(b) - \sum_{c\in b^-} f(c)\Big) = f(a) \tag{11}$$

applied with $f(a) = N'(a)$, one finds that κ_0 is a Sylvester code with Sylvester constant R_0. To see that it is the only such code, assume that κ_1 is a Sylvester code with Sylvester constant R_1 and consider the set Ω_1 of $a \in \Omega$ for which $\kappa_1(a)$ equals the expression in (9) with R_0 replaced by R_1. Clearly, Ω_1 contains the minimal nodes of Ω. Appealing again to the identity (11), one verifies that in fact $\Omega_1 = \Omega$. Then $R_1 = R_0$ as well as $\kappa_1 = \kappa_0$ follow. ∎

IV. Universal coding from isotone regression

Theorem 4.1: Let Ω be a finite co-tree provided with uniform weights $w \equiv 1$. Consider the prior given by

$$y_0(a) = N'(a) - \sum_{b\in a^-} N'(b) \tag{12}$$

for $a \in \Omega$ and let y^* be the associated isotone regression. Then the universal code κ^* is obtained from y^* by normalization, i.e. $\kappa^*(a) = y^*(a) + \ln Z$ for $a \in \Omega$ where $Z = \sum_{a\in\Omega} \exp\big(y^*(a)\big)$.

The spectrum $\sigma(\kappa^*)$ coincides with the set of top-nodes of the components of y^*. It contains the top-node of Ω as well as every minimal node of Ω.

In other words, *the universal code is obtained by normalization of the isotone regression of the Sylvester code*.

Proof: We write $\mathcal{C}^*$ in place of $\mathcal{C}_{y^*}$ and, for $a \in \Omega$, τ_a denotes the top-node of that component in $\mathcal{C}^*$ which contains a.

Let $R = \ln Z$ and let κ be given by $\kappa(a) = y^*(a) + R$ for $a \in \Omega$. To show that $\kappa = \kappa^*$, we appeal to Theorem 3.1. Isotonicity of κ is automatic and it remains to verify (7) and (8) with $R = \ln Z$. First observe that by Lemma 2.1, $\sigma(\kappa) = \{\tau_A | A \in \mathcal{C}^*\}$. Then note that for $a \in \Omega$,

$$\mathrm{D}(U_a\|\kappa) = -\ln N(a) + R + \frac{1}{N(a)} \sum_{b\leq a} y^*(b)\,. \tag{13}$$

Regarding the sum in (13), for $A \in \mathcal{C}^*$,

$$\sum_{a\in A} y^*(a) = \sum_{a\in A} y_0(a) = N'(\tau_A) - \sum_{b\in A^-} N'(b)\,. \tag{14}$$

This follows by Theorem 2.1, cf. (2), by inserting the formula (12) for $y_0(a)$ and by exploiting the structure of A.

To verify (7), assume that $a \in \sigma(\kappa)$ or, equivalently, that $a = \tau_a$. By splitting the sum $\sum_{b\leq a} y^*(b)$ as a sum over $A \in \mathcal{C}^*$ with $A \subseteq a^\downarrow$, we find by (14) that

$$\sum_{b\leq a} y^*(b) = \sum_{A\in\mathcal{C}^*, A\subseteq a^\downarrow} N'(\tau_A) - \sum_{A\in\mathcal{C}^*, A\subseteq a^\downarrow} \sum_{b\in A^-} N'(b)\,.$$

The b's that appear in the double sum run over all top-nodes of $A \in \mathcal{C}^*$ with $A \subseteq a^\downarrow$, except for the node a. Therefore, we find that $\sum_{b\leq a} y^*(b) = N'(a)$. Inserting into (13), it follows that $\mathrm{D}(U_a\|\kappa) = R$.

Then assume that $a \notin \sigma(\kappa)$, i.e. that $a < \tau_a$. Let $A_0 \in \mathcal{C}^*$ be that component which contains a. When evaluating the sum $\sum_{b\leq a} y^*(b)$, we first consider the contribution from A_0, exploiting the inequality (3) with $A = A_0$ and $L = a^\downarrow$:

$$\sum_{b\in A_0\cap a^\downarrow} y^*(b) = |A_0 \cap a^\downarrow|\,\overline{A_0} \leq \sum_{b\in A_0\cap a^\downarrow} y_0(b)\,. \tag{15}$$

The contributions to $\sum_{b\leq a} y^*(b)$ from components $A \in \mathcal{C}^*$ with $A \subseteq a^\downarrow$ are treated as in the first part of the proof. Combining with (15) and collecting facts, the desired inequality $\mathrm{D}(U_a\|\kappa) \leq R$ follows.

Regarding the last statement of the theorem, we have already seen that $\sigma(\kappa^*) = \{\tau_A | A \in \mathcal{C}^*\}$. Clearly, $\sigma(\kappa^*)$ contains the top-node of Ω. Let a be a minimal node of Ω and A the component which contains a. Then, by (3), $\overline{A} \leq \overline{\{a\}} = 0$, hence $A = \{a\}$ as y_0 is positive on non-minimal nodes (indeed, $y_0(b) \geq N'(b) - \sum_{c\in b^-} N(c) \ln N(b) = \ln N(b)$ for all $b \in \Omega$). It follows that $a = \tau_a \in \sigma(\kappa^*)$. ∎

V. Algorithmic aspects

In this section, the problem of actual calculation of universal codes or, more generally according to Theorem 4.1, of isotone regressions is addressed.

By a *pool structure* in the finite set Ω, we understand a paving π in Ω such that (i) $\emptyset \notin \pi$, (ii) every singleton is a stone, (iii) stones are non-overlapping and (iv) every non-trivial stone can be split in two. Here, singletons are *trivial stones* and other stones are *non-trivial*. Condition (iii) is the requirement $A \in \pi, B \in \pi, A\cap B \neq \emptyset \Rightarrow A \subseteq B \vee B \subseteq A$ and (iv) the requirement that given a non-trivial stone A, there exist disjoint stones B and C such that $B \cup C = A$. This splitting in two is unique. For a pool structure π, π^* denotes the paving of maximal stones, referred to as the *components* of π. The pool structure with only trivial stones is denoted $\pi_\perp$.

A sequence $\pi_0, \pi_1, \cdots \pi_\nu$ of pool structures is a *creation* of π from π_0 if, for $i = 1, \cdots, \nu$, π_i is obtained form π_{i-1} by

[8] authors terminology, motivated by the fact that a problem of Sylvester, cf. [9], appears to be the first problem pointed to in the literature which can be handled efficiently by the kind of game theoretical reasoning which is in the background for important parts of our technical approach.

pooling two components of π_{i-1} (i.e. $\pi_i = \pi_{i-1} \cup \{A \cup B\}$ for two distinct components A and B of π_{i-1}). The creation is a *complete creation* of π if it is a creation of π from $\pi_\perp$. Every pool structure has a complete creation $\pi_\perp, \pi_1, \cdots, \pi_\nu$. It need not be uniquely determined from π, but the number ν, the *pooling index* of π is. It is given by $\nu = |\pi| - |\Omega|$. Pooling is only possible of components. By this we mean that if π and $\pi \cup \{A\}$ are distinct pool structures, then A must be a union of two distinct components of π. Thus pooling is not possible if a pool structure has only one component. Such structures are said to be *closed*[9]. The opposite process of pooling is *splitting*, replacing a non-trivial component by its two parts. This process decreases $|\pi|$ by 1 and increases $|\pi^*|$ by 1. Thus, after successive splittings you arrive at the trivial pool structure and hence you can conclude that:

Lemma 5.1: For a pool structure π, $|\pi| + |\pi^*| = 2|\Omega|$.

From now on assume that $\Omega = (\Omega, \leq, w, y_0)$ is a co-tree provided with the usual extra structure. A pool structure π in Ω is a $\mathcal{C}$- pool structure, and we write $\pi \in \Pi_\mathcal{C}$, if $\pi \subseteq \mathcal{C}$. Assume that this is the case and consider some $A \in \pi^*$. Denote by $\mathcal{L}^-(A)$ the class of *lower adjacent components* of A, components with a top node in a^- for some $a \in A_{\min}$.

The PAV-*algorithm* by which y^* can be constructed, consists of successive calls of the PAV-*sub routine* which, in turn, consists of successive calls of the PAV-*subsub routine*.

The input to the PAV-subsub routine is (π, A) with $\pi \in \Pi_\mathcal{C}$ and $A \in \pi^*$. The output, $[\pi, A] \in \Pi_\mathcal{C}$, is constructed from the sets in $\mathcal{L}^-(A)$. If $\mathcal{L}^-(A) = \emptyset$, or if $\overline{B} < \overline{A}$ for every $B \in \mathcal{L}^-(A)$, $[\pi, A] = \pi$ whereas, if not, we choose $B \in \mathcal{L}^-(A)$ (e.g. using a predetermined numbering of the nodes of Ω) with $\overline{B}$ maximal and put $[\pi, A] = \pi \cup \{A \cup B\}$. If A is understood, we may write π' for $[\pi, A]$ and if π is understood, we may write A' for A when $\pi' = \pi$ and A' for the set B above otherwise.

The PAV-sub routine has as input (π, a) with $\pi \in \Pi_\mathcal{C}$ and $\{a\} \in \pi^*$. The output is a creation of a pool structure $[\pi, a] \in \Pi_\mathcal{C}$ obtained through successive calls of the PAV-subsub routine, first with $(\pi, \{a\})$ as input, then with $(\pi', \{a\}')$ as input etc. until no new pool structure is created.

The PAV-algorithm consists of successive calls of the PAV-subroutine, first with inputs $(\pi_\perp, a)$ for all minimal nodes of Ω, then for the resulting pool structures and the nodes further up in the co-tree until finally you call the PAV-sub routine with the top node as the input-node.

Theorem 5.1: The components of the final pool structure created through the PAV-algorithm are the sought components of the isotone regression.

Space does not allow inclusion of a proof. It can be carried out by induction, at each step ensuring that a relevant regression for part of the co-tree already covered is constructed.

Let π be the final pool structure constructed by the algorithm and α the *(time) complexity* measured as the number of calls of the subsub routine. In case you only have to make one comparison when the subsub routine is called, e.g. if Ω is a linear order or, more generally, if Ω has *uniform branching* (cf. [3]) and if the prior is symmetric in a natural sense, the algorithm is of linear complexity as follows from Lemma 5.1.

By a combinatorial argument, $\alpha = \sum_{A \in \pi} |A^-| \leq \binom{n}{2}$ with $n = |\Omega|$ hence, if no modifications are made to the algorithm, it is at most of quadratic complexity in the size of the problem, independently of the nature of the prior.

Memory management can be carried out efficiently by noting that at each step in the creation of pool structures, you only need to retain the maximal stones.

[9] closed pool structures are in an obvious way of relevance for coding, e.g. for Shannon-Fano coding or Huffman coding.

VI. Conclusions

The key insight gained is that universal coding for co-trees can be carried out via the solution of a problem of isotone regression. Also, it has to be realized that much clarity results from focusing on isotone regression as the primary problem. Thus information theorists should perhaps consider to add this area of statistics to their toolbox. Further, we note that the role of game theoretical considerations has once more proved its value. Lastly, it seems that it is worth stressing the role of pool structures as introduced here.

As to limits of the research, it seems, as also briefly indicated in Section I, that we have gone as far as technically possible regarding the universal coding problem, whereas results for more general order structures than trees and co-trees are possible regarding the problem of isotone regression. In this connection, it is unclear to the author if an incremental construction via pool structures can be much extended – e.g. for the example in Fig.1 it is not clear if this is feasible (and this is not clarified in the technical paper [7]).

Acknowledgments

This goes to Henrik Densing Petersen, my co-author of [3], to Peter Harremoës (technical point re proof of Theorem 2.1), to Inge Li Gørtz and Philip Bille (general discussion of algorithms) and, foremost, to an anonymous referee of [3] who spotted a possible connection between the two areas of research addressed here.

References

[1] B. Y. Ryabko, "Encoding of a source with unknown but ordered probabilities," *Problems of Information Transmission*, vol. 15, pp. 71–77, 1979, russian original in Probl. Peredachi Inf., 1978.

[2] W. R. M. Ayer H.D. Brunk, G.M. Ewing and E. Silverman, "An empirical distribution function for sampling with incomplete information," *Ann. Math. Statist.*, vol. 26, pp. 641–647, 1955.

[3] H. D. Petersen and F. Topsøe, "Computation of Universal Objects for Distributions over Co-Trees," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7021–7035, Dec. 2012.

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

[5] P. J. de Leeuw, K. Hornik, "Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods," *J. Stat. Software*, vol. 32, no. 5, pp. 1–24, 2009.

[6] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Stat.*, vol. 19, no. 4, pp. 2032–2066, Dec. 1991.

[7] C.-I. C. Lee, "The Min-Max Algorithm and Isotonic Regression," *Ann. Statist.*, vol. 11, no. 2, pp. 467–477, 1983.

[8] F. Topsøe, "Game theoretical optimization inspired by information theory," *J. Glob. Optim.*, vol. 43, pp. 553–564, 2009.

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