Entropy splitting for antiblocking pairs and perfect graphs

by

I. Csiszár[1], J. Körner[2], L. Lovász[3] K. Marton[1] and G. Simonyi[2]

[1] Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary H-1053

[2] Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary H-1053 and and Départment Informatique, ENST, Paris

[3] Department of Computer Science, Eötvös University, Budapest, Hungary H-1088 and Department of Computer Science, Princeton University, Princeton, N.J. 08544

Research was partially sponsored by the Hungarian National Foundation, Scientific Research Grants No 1806 and 1812.

Abstract

We characterize pairs of convex sets A, B in the k-dimensional space with the property that every probability distribution $(p_1, \ldots, p_k)$ has a representation

$$p_i = a_i \cdot b_i, \quad a \in A,\ b \in B.$$

Minimal pairs with this property are antiblocking pairs of convex corners. This result is closely related to a new entropy concept. The main application is an information theoretic characterization of perfect graphs.

1. Introduction

The concept of antiblocking pairs of polyhedra was introduced by Fulkerson [4]; it can be extended to non-polyhedral convex sets in a straightforward way (see [6]). Let $\mathbb{R}_+^k$ denote the non-negative orthant of the k-dimensional Euclidean space. If $a = (a_1, \ldots, a_k)^T,\ b = (b_1, \ldots, b_k)^T \in \mathbb{R}_+^k$, then $b^T \cdot a$ denotes their inner produced, and we write $a \leq b$ if $a_i \leq b_i$ for all i.

Definition. A set $A \subseteq \mathbb{R}_+^k$ is called a *convex corner* if it is compact, convex, has non-empty interior, and for every $a \in A, a' \in \mathbb{R}_+^k$ with $a' \leq a$ we have $a' \in A$. The *antiblocker* of the convex corner A is the convex corner

$$A^* = \{b \in \mathbb{R}_+^k :\ b^T \cdot a \leq 1\ \forall a \in A\}.$$

If $B = A^*$ then (A, B) is called an *antiblocking* pair. It is well knwon that $(A^*)^* = A$ and hence if (A, B) is an antiblocking pair then so is (B, A).

If $A, B \subseteq \mathbb{R}_+^k$ are convex corners and $A \subseteq B$ then $B^* \subseteq A^*$ and $(A^*)^* = A$.

A vector $p \in \mathbb{R}_+^k$ is called a *probability distribution* if its coordinates add up to 1. We are interested in pairs of sets in $\mathbb{R}_+^k$ that generate all probability distributions in the following sense.

Definition. For $a, b \in \mathbb{R}_+^k$, let $a \circ b$ denote the vector $(a_i \cdot b_i : i = 1, \ldots, k)$. For two sets $A, B \subseteq \mathbb{R}^k$, we put $A \circ B = \{a \circ b : \ a \in A, \ b \in B\}$. (If A and B are convex corners then $A \circ B$ is not necessarily a convex corner.) A pair of sets $A, B \subseteq \mathbb{R}_+^k$ is called a *generating pair* if every probability distribution $p \in \mathbb{R}_+^k$ can be represented as

$$p = a \circ b, \quad a \in A, \ b \in B. \tag{1}$$

For convex corners, this is equivalent to saying that $S \subseteq A \circ B$. We shall prove that a pair of convex corners $A, B \subseteq \mathbb{R}_+^k$ is a generating pair iff $A^* \subseteq B$ (which is equivalent to $B^* \subseteq A$). Also, if (A, B) is an antiblocking pair then the representation (1) is essentially unique.

These results are closely related to a new entropy concept.

Definition. Let $A \subseteq \mathbb{R}_+^k$ be a convex corner, and $p \in \mathbb{R}_+^k$ a probability distribution. The *entropy* of p with respect to A is

$$H_A(p) = \min_{a \in A} -\sum_{i=1}^{k} p_i \log a_i$$

(the log's are taken to the base 2).

Observe that the function to minimize is convex, tends to ∞ at the boundary of the non-negative orthant but it tends monotone to $-\infty$ along rays from the origin. Hence the minimum is always achieved and finite, and is assumed at the boundary of A but in the interior of the non-negative orthant. It also follows easily that each coordinate a_i of the minimizing vector a is uniquely determined provided $p_i > 0$.

To justify the name "entropy" for this quantity, let us remark that the entropy $H_S(p)$ of a probability distribution p with respect to the *unit corner* $S = \{x \geq 0, \sum_i x_i \leq 1\}$ is just the Shannon entropy $H(p) = -\sum_i p_i \log p_i$.

There is another way to obtain this value. Consider the mapping $\Lambda : \ \mathrm{int}\, \mathbb{R}_+^n \to \mathbb{R}^n$ defined by

$$\Lambda(x) = (-\log x_1, \ldots, -\log x_n).$$

It is easy to see using the concavity of the log function that if A is a convex corner then $\Lambda(A)$ is a closed, convex, full-dimensional set, which is *up-monotone*, i.e., $a \in \Lambda(A)$, $a' \geq a$ imply $a' \in \Lambda(A)$. Now $H_A(p)$ is the minimum of the linear objective function $\sum_i p_i x_i$ over $\Lambda(A)$.

Our main result is the following.

1.1 Theorem. *For convex corners $A, B \subseteq \mathbb{R}_+^k$ the following three conditions are equivalent:*

(i) $A^* \subseteq B$;

(ii) (A, B) *is a generating pair;*

(iii) $H(p) \geq H_A(p) + H_B(p)$ *for every probability distribution $p \in \mathbb{R}_+^k$.*

This and related results will be proved in Section 2. As a main application, in Section 3 we shall prove the following characterization of perfect graphs: *a graph is perfect iff it "splits graph entropy"*. Graph entropy, introduced by Körner [7], is an information theoretic functional on a graph with a probability distribution given on its vertex set; it may also be considered as a probabilistic refinement of the notion of chromatic number.

Definition. Let $G = (V, \mathcal{E})$ be a graph with vertex set V and edge set $\mathcal{E}$, and let p be a probability distribution on V. Let $G^{(n)} = (V^n, \mathcal{E}^{(n)})$ denote the n-th conormal power of G, i.e., V^n is the set of sequences of length n from V, and

$$\mathcal{E}^{(n)} = \{(x, y) \in V^n \times V^n : \ \exists i : \ (x_i, y_i) \in \mathcal{E}\}.$$

Define the probability distribution p^n on V^n by $p^n(x) = \prod_{i=1}^n p(x_i)$. For $U \subseteq V^n$, let $G^{(n)}(U)$ denote the subgraph induced by U in $G^{(n)}$, and let $\chi(G^{(n)}(U))$ denote its chromatic number. Then for every $0 < \varepsilon < 1$, the limit

$$H(G, p) = \lim_{n\to\infty} \frac{1}{n} \min_{p^n(U)\geq 1-\varepsilon} \log \chi(G^{(n)}(U))$$

exists and is independent of ε. $H(G, p)$ is called the *graph entropy* of the graph G with respect to the probability distribution p.

We may view the elements of V as an alphabet, two letters being connected by an edge iff they are "distinguishable". The elements of V^n are words of length k. Two such words are connected in the n-th conormal power iff they are distinguishable (i.e., they are distinguishable in at least one position). We want to encode these words by 0-1 words of length as smal;l as possible, so that distinguishable words get different codes (two words that are indistinguishable anyway may get the same code). Such an encoding corresponds to a coloring of the n-th conormal power. However, we only want to encode the "majority" of words, i.e., a fraction of ε is allowed not to get codes. Then for large n, the optimum encoding uses words with length about $H(G, p) \cdot n$.

[7] contains a non-asymptotic formula for $H(G, p)$, from which we shall derive (see Lemma 3.1 below) that $H(G, p)$ is the entropy of p with respect to the so-called vertex packing polytope of G (which is a convex corner).

Graph entropy can be used to obtain lower bounds for the minimum number of graphs of a given type needed to cover the edge set of a fixed graph (cf. [9], [11]). This is based on the following sub-additivity property of graph entropy. Let $F = (V, \mathcal{E}_1), G = (V, \mathcal{E}_2)$ be graphs on the same vertex set V; their union is the graph

$$F \cup G = (V, \mathcal{E}_1 \cup \mathcal{E}_2).$$

In [9] it is proved that

$$H(F \cup G, p) \leq H(F, p) + H(G, p)$$

for all probability distributions p on V. In particular, for a graph G

$$H(p) = H(G \cup \overline{G}, p) \leq H(G, p) + H(\overline{G}, p). \tag{2}$$

(The fact that the entropy of the complete graph with respect to the probability distribution p is $H(p)$, follows from the formula for $H(G,p)$ given in [7] or from Lemma 3.1 below.)

Körner and Longo [10] introduced the following notion.

Definition. A graph $G = (V, \mathcal{E})$ is *strongly splitting* if for every probability distribution p on V, (2) holds with equality, i.e.,

$$H(p) = H(G,p) + H(\overline{G},p) \tag{3}$$

for every probability distribution p on V. A graph is called *weakly splitting* if (3) holds for at least one probability distribution $p > 0$.

The results [8] and [10] show that every perfect graph is weakly splitting, but there are weakly splitting graphs that are not perfect. A graph-theoretic characterization of weakly splitting graphs is contained in [10] and [11]: a graph is weakly splitting if and only if it is *normal*, i.e., it contains a a family $\mathcal{A}$ of independent sets and a family $\mathcal{B}$ of cliques, both covering all points, such that every $A \in \mathcal{A}$ intersects every $B \in \mathcal{B}$.

Körner and Marton [11] showed that bipartite graphs are strongly splitting while odd cycles are not. They conjectured the following characterization, to be proved in section 3:

1.2 Theorem. *A graph is strongly splitting iff it is perfect.*

We use Theorems 1.1 and 1.2 to derive, for strongly splitting (i.e., perfect) graphs, a stronger version of normality.

In Section 4 we generalize this characterization to some families of subsets of a given set. We give an information theoretic interpretation of the entropy of a probability distribution with respect to the convex corner spanned by the indicator vectors of a family of subsets of a given set. Another example of entropy with respect to a convex corner is a probabilistic version of the functional $\theta(G)$ introduced by Lovász [14] to bound graph capacity from above. This example will be studied in detail in [16].

In Section 5 we prove additivity and subadditivity properties of entropy with respect to a convex corner. For graph entropy these are known results, but for the functional of [16] the additivity is quite surprising.

2. Generating pairs of convex corners

We start with a simple lemma about entropy of convex corners.

2.1 Lemma. *For two convex corners $A, C \subseteq \mathbb{R}_+^k$, we have $H_A(p) \geq H_C(p)$ for all p if and only if $A \subseteq C$.*

Proof. The "if" part is obvious. Assume that $H_C(p) \leq H_A(p)$ for all p. As remarked above, we have $H_A(p) = \min\{p^T x : x \in \Lambda(A)\}$, and hence it follows that we must have

$\Lambda(A) \subseteq \Lambda(C)$. This clearly implies $A \subseteq C$. ∎

In particular, it follows from this lemma that a convex corner A is completely determined if we know $H_A(p)$ for all p. Note that $H_A(p)$ may be negative or larger than $H(p)$. However, Lemma 2.1 has the following

2.2 Corollary. *We have $0 \le H_A(p) \le H(p)$ for every probability distribution p iff A contains the unit corner and is contained in the unit cube.*

Our next lemma relates entropy to antiblocking pairs.

2.3 Lemma. *Let $A, B \subseteq \mathbb{R}^k_+$ be convex corners and $p \in \mathbb{R}^n_k$, a probability distribution. Then*

(a) *If $p = a \circ b$ for some $a \in A$ and $b \in B$ then*

$$H(p) \ge H_A(p) + H_B(p),$$

with equality if and only if a and b achieve $H_A(p)$ and $H_B(p)$.

(b) *If $A^* \supseteq B$ then*

$$H(p) \le H_A(p) + H_B(p).$$

with equality iff $p = a \circ b$ for some $a \in A$, $b \in B$.

Proof. (a) We have

$$\begin{aligned} H(p) &= -\sum_i p_i \log a_i b_i = -\sum_i p_i \log a_i - \sum_i p_i \log b_i \\ &\ge H_A(p) + H_B(p). \end{aligned}$$

We have equality here if and only if a and b achieve $H_A(p)$ and $H_B(p)$.

(b) Let $a \in A$ and $b \in B$ achieve $H_A(p)$ and $H_B(p)$, respectively. Then the strict concavity of the log function and the relation $b^T a \le 1$ imply

$$H_A(p) + H_B(p) - H(p) = -\sum_i p_i \log \frac{a_i b_i}{p_i} \ge -\log \sum_i a_i b_i \ge 0.$$

Equality holds if and only if $a_i b_i = p_i$ whenever $p_i > 0$. But then by $1 \ge \sum_i a_i b_i \ge \sum_i p_i = 1$, equality also holds for those indices with $p_i = 0$. ∎

Proof of Theorem 1.1.

(i)⇒(ii): We have to show that if $A^* \subseteq B$ then every probability distribution $p \in \mathbb{R}^k_+$ has a representation (1). Let $a \in A$ minimize $f(x) = -\sum_i p_i \log x_i$ over A (i.e., achieve $H_A(p)$). If $p_i > 0$ then obviously $a_i > 0$, so the vector $b = (b_i)$,

$$b_i = \begin{cases} p_i/a_i, & \text{if } p_i > 0 \\ 0, & \text{otherwise.} \end{cases}$$

is well defined, and all we have to show is that $b \in B$.

Observe that the convex sets A and $\{x \in \mathbb{R}_+^k : f(x) < f(a)\}$ are disjoint and so they can be separated by a hyperplane. But the two sets touch at the point a and the second one is smooth there, so this separating hyperplane must be its tangent there. Now the gradient of $-f$ at a is $(p_1/a_1, \ldots, p_k/a_k) = b$ and so this separating hyperplane is $b^T x = 1$. But this means that $b^T x \leq 1$ for every $x \in A$, i.e., $b \in A^* \subseteq B$.

(ii)⇒(iii) follows from Lemma 2.3(a).

(iii)⇒(i): Notice first that, by the (already established) implication (i)⇒(iii) and by Lemma 2.3(b), we have

$$H(p) = H_A(p) + H_{A^*}(p)$$

for every probability distribution $p \in \mathbb{R}_+^k$, and hence $H_{A^*}(p) \geq H_B(p)$ for every p. By Lemma 2.1, this implies that $A^* \subseteq B$. ∎

Notice that we do not know how to decide for an arbitrary pair of convex corners $A, B \subseteq \mathbb{R}_+^k$, and a given probability distribution p whether p has a representation (1). Lemma 2.3 answers this question if $A^* \supseteq B$.

Theorem 1.1 and Lemma 2.3 also imply the following characterization of antiblocking pairs:

2.4 Corollary. *Let $A, B \subseteq \mathbb{R}_+^k$ be convex corners. (A, B) is an antiblocking pair iff*

$$H(p) = H_A(p) + H_B(p)$$

for every probability distribution $p \in \mathbb{R}_+^k$. ∎

The next assertion describes, for an antiblocking pair (A, B), the pairs (a, b) needed in representations of probability distributions. For a convex corner A, let A' denote the closure of that part of the boundary of A that is not contained in any of the coordinate hyperplanes $x_i = 0$.

2.5 Corollary. *Let A be a convex corner in $\mathbb{R}_+^k$ and $a \in A'$. Let $b \geq 0$ be the normal vector to a supporting hyperplane to A through a, normalized by $b^T a = 1$. Then $b \in A^*$, and $a \circ b$ is a probability distribution. Every probability distribution $p \in \mathbb{R}_+^k$ has a representation by pairs (a, b) obtained this way, and if $p_i > 0$ for all i then this representation is unique.* ∎

Let us remark that this proposition motivates an alternative proof of the generating property which is topological and essentially different from the one given above. Here is a sketch. Let A be a convex corner. Assume that A' is smooth. Then through any point $a \in A'$ there is a unique tangent hyperplane, and, consequently, a unique normal vector $b = b(a) \in A^*$ satisfying $b^T a = 1$. The function $\varphi(a) = a \circ b(a)$ is then a continuous mapping from A' into the simplex of probability distributions in $\mathbb{R}_+^k$. Using Brouwer's Fixed Point Theorem, one can establish that φ is onto. The case when A' is not smooth follows by a compactness argument.

3. Perfect graphs and entropy splitting

Let us first recall the definition of perfect graphs and of certain polytopes associated with graphs. From now on, we assume $V = \{1, \dots, k\}$.

Definition. A graph G is *perfect* if for every induced subgraph G' of G, the chromatic number of G' equals the maximum size of a clique in G'.

Perfect graphs have been introduced by Berge; cf. Berge [1] and Lovász [15]. We need a pair of other important notions from graph theory ([5, 14]; see also [6]):

Definition. The *vertex packing polytope* $VP(G)$ of the graph G is the convex hull of the indicator vectors of the independent sets of G. The *fractional vertex packing polytope* of G is defined as

$$FVP(G) = \{b \in \mathbb{R}_+^k : \sum_{i \in K} b_i \leq 1 \text{ for all cliques } K \text{ of } G\}.$$

It is easy to see that $VP(G)$ and $FVP(G)$ are convex corners. Moreover, $FVP(G) = [VP(\overline{G})]^*$, and $VP(G) \subseteq FVP(G)$ for every graph G. Equality holds here if and only if the graph is perfect (Fulkerson [5], Chvátal [2]). We can express the graph entropy $H(G, p)$ as the entropy of p with respect to $VP(G)$:

3.1 Lemma. *For every graph $G = (V, \mathcal{E})$ and every probability distribution p on V,*

$$H(G, p) = H_{VP(G)}(p).$$

Proof. We have to use some elementary concepts from information theory. The interested reader may consult [3] or [17]. If X is a random variable with values in the set $V = \{1, 2, \dots, k\}$ and distributed according to the probability distribution p then the *Shannon entropy* of X is $H(X) = H(p) = -\sum_{i \in V} p_i \log p_i$. If (X, Y) is a pair of random variables having finite range then the *mutual information* of X and Y is

$$I(X \wedge Y) = H(X) + H(Y) - H(X, Y).$$

Here we consider (X, Y) as one random variable, and $H(X, Y)$ stands for the entropy of this random variable. For graph entropy the following formula was proved in [7]:

$$H(G, p) = \min\{I(X \wedge Y) : \ \mathrm{dist}(X) = p, \ X \in Y \in \mathcal{F}(G)\}. \tag{4}$$

Here dist(X) denotes the distribution of X, $\mathcal{F}(G)$ is the family of independent sets of G, and "$X \in Y \in \mathcal{F}(G)$" means that (X, Y) is a random pair, Y takes values in $\mathcal{F}(G)$, and the random vertex X is bound to belong to the random set Y. First we prove

$$H(G,p) \geq \min_{a \in VP(G)} -\sum_{i=1}^{k} p_i \log a_i.$$

Let the min in (4) be achieved by a random pair (X, Y), $\mathrm{dist}(X) = p$, $X \in Y \in \mathcal{F}(G)$. Let q denote the conditional distribution of Y given X, and let r be the distribution of Y. By the definition of mutual information and some trivial identities,

$$H(G,p) = I(X \wedge Y) = -\sum_i p_i \sum_{i \in F \in \mathcal{F}(G)} q(F \mid i) \log \frac{r(F)}{q(F \mid i)}.$$

By the concavity of the log function, the inner sum is at most $\log \sum_{i \in F \in \mathcal{F}(G)} r(F)$. Define the vector a by $a_i = \sum_{i \in F \in \mathcal{F}(G)} r(F)$; then $a \in VP(G)$, and $H(G,p) \geq -\sum_i p_i \log a_i$.

To prove the reverse inequality, fix a point $a \in VP(G)$, say, $a_i = \sum_{i \in F \in \mathcal{F}(G)} s(F)$, where s is a probability distribution on $\mathcal{F}(G)$, and define the transition probabilities

$$q(F \mid i) = \begin{cases} s(F)/a_i, & \text{if } i \in F \\ 0, & \text{if } i \notin F \end{cases}$$

($i \in V, F \in \mathcal{F}(G)$). We have

$$H(G,p) \leq \sum_{i,F} p_i q(F \mid i) \log \frac{q(F \mid i)}{r(F)}, \tag{5}$$

where $r(F) = \sum_i p_i q(F \mid i)$. By the concavity of the log function,

$$-\sum_F r(F) \log r(F) \leq -\sum_F r(F) \log s(F),$$

and hence

$$-\sum_{i,F} p_i q(F \mid i) \log r(F) \leq -\sum_{i,F} p_i q(F \mid i) \log s(F).$$

Thus (5) can be continued:

$$H(G,p) \leq \sum_{i,F} p_i q(F \mid i) \log \frac{q(F \mid i)}{s(F)} = -\sum_i p_i \log a_i.$$

∎

Now we can characterize not only the strongly splitting graphs, but also those for which (3) holds for a given p:

3.2 Lemma. *For a probability distribution p on V, we have $H(p) = H(G,p)+H(\overline{G},p)$ iff $H_{VP(G)}(p) = H_{FVP(G)}(p)$.*

Proof. We have $[VP(\overline{G})]^* = FVP(G)$. Thus Lemma 3.1 and Corollary 2.4 imply

$$\begin{aligned} H(G,p) + H(\overline{G},p) - H(p) =& H_{VP(G)}(p) + H_{VP(\overline{G})}(p) - H(p) \\ =& H_{VP(G)}(p) - H_{FVP(G)}(p). \end{aligned}$$

∎

Proof of Theorem 1.2. By Lemmas 3.2 and 2.1, G is strongly splitting iff $VP(G) = FVP(G)$. This is equivalent to the perfectness of G. ∎

Let $\mathcal{F}(\mathcal{G})$ and $\mathcal{K}(\mathcal{G})$ denote the families of the independent sets and cliques, respectively, of the graph G. By definition, a vector $a \in \mathbb{R}^k_+$ belongs to $VP(G)$ iff there exists a probability distribution q on $\mathcal{F}(G)$ such that the coordinates of a can be written as

$$a_i = \sum_{i \in F \in \mathcal{F}(G)} q(F).$$

Thus, by Lemma 2.2, Theorem 1.2 can be stated in the following equivalent, and perhaps more transparent form (c.f. [11], [12]):

3.3 Corollary. *The graph $G = (V, \mathcal{E})$ is perfect iff for every probability distribution p on V there exist probability distributions q on $\mathcal{F}(G)$ and r on $\mathcal{K}(G)$ such that for all $i \in V$,*

$$p_i = \sum_{i \in F \in \mathcal{F}(G)} q(F) \sum_{i \in K \in \mathcal{K}(G)} r(K).$$

By Corollary 2.5, q and r are concentrated on the maximal independent sets and maximal cliques of G, respectively, whenever $p_i > 0$ for all i. In contrast to the uniqueness of the representation (1), q and r are not uniquely determined.

Another way to put this result is the following. [4] It follows from Theorem 1.1 that for each graph G,

$$S = VP(G) \circ FVP(\overline{G}) = FVP(G) \circ VP(\overline{G})$$

(where S is the unit corner). Hence

$$VP(G) \circ VP(\overline{G}) \subseteq S \subseteq FVP(G) \circ FVP(\overline{G}).$$

[4] We are grateful to the referee of our paper for this remark.

Now corollary 3.3 asserts that G is perfect if and only if $VP(G) \circ VP(\overline{G}) = S$. This may be contrasted with a result of Fulkerson [4] that can be phrased as follows: G is perfect if and only if $FVP(G) \circ FVP(\overline{G}) \subseteq S$, i.e., that $l \cdot w \leq 1$ for every $l \in FVP(G)$ and $w \in FVP(\overline{G})$ (this inequality is sometimes called the *length-width inequality* or *max-max inequality*. In view of the inequality above, this is equivalent to saying that $FVP(G) \circ FVP(\overline{G}) = J$.

By studying the structure of the representation in Corollary 3.4, we can derive the following strengthening of the normality of perfect graphs.

3.4 Theorem. *Let G be a perfect graph. Then G contains a family $\mathcal{A}$ of independent sets and a family $\mathcal{B}$ of cliques with the following properties:*
(a) $|\mathcal{A}| + |\mathcal{B}| = k + 1$*;*
(b) *the sets in $\mathcal{A}$ ($\mathcal{B}$) cover all points;*
(c) *the incidence vectors of sets in $\mathcal{A}$ ($\mathcal{B}$) are linearly independent;*
(d) *every $A \in \mathcal{A}$ intersects every $B \in \mathcal{B}$.*

Proof. For every probability distribution $p > 0$, we have a family $\mathcal{A}$ of independent sets and a family $\mathcal{B}$ of cliques, and non-negative reals λ_A ($A \in \mathcal{A}$) and μ_B ($B \in \mathcal{B}$) such that $\sum_A \lambda_A = 1$, $\sum_B \mu_B = 1$, and for each $i \in V$,

$$\sum_{i \in A} \sum_{i \in B} \lambda_A \mu_B = p_i. \tag{6}$$

We may assume here that $\lambda_A, \mu_B > 0$ and that the incidence vectors $a_1, \ldots, a_s$ of the members of $\mathcal{A}$ as well as the incidence vectors $b_1, \ldots, b_t$ of the members of $\mathcal{B}$ are affinely independent. Adding up (6) for each i, we get that

$$\begin{aligned} 1 = \sum_i p_i &= \sum_{A \in \mathcal{A}} \sum_{B \in \mathcal{B}} \lambda_A \mu_B |A \cap B| \\ &\leq \sum_{A \in \mathcal{A}} \sum_{B \in \mathcal{B}} \lambda_A \mu_B = \Big(\sum_{A \in \mathcal{A}} \lambda_A \Big) \cdot \Big(\sum_{B \in \mathcal{B}} \mu_B \Big) \\ &= 1. \end{aligned}$$

Hence we see that we must have $|A \cap B| = 1$ for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$, i.e., (d) holds. Since $p > 0$, (b) is obvious. (c) follows by observing that every a_i satisfies $a_i^T b_1 = 1$ and hence the affine independence of the a_i implies their linear independence: any linear dependence relation $\sum_i \alpha_i a_i = 0$ would imply $\sum_i \alpha_i = \sum_i \alpha_i a_i^T b_1 = 0$, i.e., it would be an affine dependence. This proves (c). It is an easy linear algebra that (c) and (d) imply "one half" of (a): for each b_j, (d) provides a linear equation $b_J^T x = 1$ satisfied by the a_i, and since these relations are linearly independent by (c), the affine hull of the a_i has dimension at most $k - |\mathcal{B}|$. Since they are affine independent, the number of the a_i is at most $k - |\mathcal{B}| + 1$. This proves that $|\mathcal{A}| + |\mathcal{B}| \leq k + 1$.

To show that equality can be achieved here, we have to use that every p has such a representation. Note that there are only finitely many possible pairs $\mathcal{A}, \mathcal{B}$, and each fixed pair provides a representation of the form (6) for a closed set of probability distributions p.

Hence there must be a pair $\mathcal{A}, \mathcal{B}$ that provides representation for a $(k-1)$-dimensional set of probability distributions. But (6) can be viewed as a polynomial mapping of the direct product of a simplex with $|\mathcal{A}|$ vertices and a simplex with $|\mathcal{B}|$ vertices into the simplex of probability distributions. Since such a mapping does not increase dimension, this implies that

$$(|\mathcal{A}| - 1) + (|\mathcal{B}| - 1) \geq k - 1,$$

which proves (a). ∎

Unfortunately, this theorem does not characterize perfect graphs: for example, the 9-cycle has the property stated in the theorem. In fact, let $V = \{1, \ldots, 9\}$ be the vertex set of the 9-cycle, and consider the cliques

$$\{1,2\},\ \{2,3\},\ \{4,5\},\ \{5,6\},\ \{7,8\},\ \{8,9\}$$

and the independent sets

$$\{2,5,8\},\ \{1,3,5,8\},\ \{2,4,6,8\},\ \{2,5,7,9\}.$$

To conclude this section, we consider briefly a notion analogous to graph entropy, but defined using the normal, rather than the conormal, powers of a graph. Theorem 2 will imply that, for perfect graphs, the value of the entropy is independent of the graph multiplication involved.

Definition. The n-th *normal power* $G^{[n]} = (V^n, \mathcal{E}^{[n]})$ of the graph $G = (V, \mathcal{E})$ is defined by

$$\mathcal{E}^{[n]} = \{(x,y) \in V^n \times V^n :\ x \neq y,\ \forall i :\ (x_i, y_i) \in \mathcal{E} \text{ or } x_i = y_i\}.$$

Note that the normal and conormal powers are related by complementation: $\overline{G^{[n]}} = \overline{G}^{(n)}$.

Definition [10]. The *π-entropy* of the graph $G = (V, \mathcal{E})$ with respect to the probability distribution p on V is defined as

$$H_\pi(G,p) = \lim_{\delta \to 0} \lim_{n \to \infty} \min_{\substack{U \subseteq V^n \\ p^n(U) \geq 1-\delta}} \frac{1}{n} \log \chi(G^{[n]}(U)).$$

It was noted in [10] that $H(p) \leq H_\pi(G,p) + H(\overline{G}, p)$. Moreover, evidently, $H_\pi(G,p) \leq H(G,p)$. Thus Theorem 1.2 implies the following.

3.5 Corollary. *If G is perfect then*

$$H_\pi(G,p) = H(G,p) \tag{7}$$

for every probability distribution p on V. ∎

In [10] the problem of characterization of the graphs satisfying (7) was raised. Though this Corollary gives some information, we still do not know whether there exist non-perfect graphs with the property (7). Note that no non-asymptotic formula is known for $H_\pi(G,p)$ in general; indeed, such a formula would imply a formula for graph capacity (c.f. [16]). In [16] a lower bound is given for $H_\pi(G,p)$.

4. Families of subsets of a given set

Here we discuss the limits of validity of Corollary 3.3 if $\mathcal{F}(G)$ and $\mathcal{K}(G)$ are replaced by arbitrary families of subsets of the set $V = \{1, 2, \ldots, k\}$. We only allow families the union of which covers the whole set V.

Definition. The families $\mathcal{F}$ and $\mathcal{K}$ of subsets of the set V are said to form a *generating pair* if for every probability distribution p on V there exist probability distributions q on $\mathcal{F}$ and r on $\mathcal{K}$ such that

$$p_i = \sum_{i \in F \in \mathcal{F}} q(F) \sum_{i \in K \in \mathcal{K}} r(K), \tag{8}$$

for every $i \in V$.

First note that only the maximal elements of $\mathcal{F}$ and $\mathcal{K}$ occur in the representation (8), and so we may assume that $\mathcal{F}$ and $\mathcal{K}$ are hereditary families, i.e., together with any member they contain all subsets of it. (We might as well assume that they are clutters, i.e., they do not contain any comparable pairs of sets; but this will be more convenient.)

Denote by $C(\mathcal{F})$ the convex hull of the indicator vectors of the members of $\mathcal{F}$; this is a convex corner. Our definition says that $\mathcal{F}$ and $\mathcal{K}$ form a generating pair if $(C(\mathcal{F}), C(\mathcal{K}))$ is a generating pair in $\mathbb{R}^k_+$. Thus Theorem 2.1 implies the following:

4.1 Corollary. *$(\mathcal{F}, \mathcal{K})$ is generating if and only if $C(\mathcal{F})$ and $C(\mathcal{K})$ contain each other's antiblockers.*

Unfortunately, this characterization of generating pairs of families of sets is not easy to use. In the case when we have that $|F \cap K| \leq 1$ for each $K \in \mathcal{K}$ and $F \in \mathcal{F}$, we can give the following more definite description. (Note, however, that since the class of perfect graphs is not well-characterized in the sense of complexity theory, even in this special case the answer is not complete.)

4.2 Theorem. *For each pair of hereditary set-systems $\mathcal{F}$ and $\mathcal{K}$ on the same set V, the following conditions are equivalent:*

(i) *$|F \cap K| \leq 1$ for all $F \in \mathcal{F}$ and $K \in \mathcal{K}$, and $(\mathcal{F}, \mathcal{K})$ is a generating pair;*

(ii) *$(C(\mathcal{F}), C(\mathcal{K}))$ is an antiblocking pair;*

(iii) *$H_{C(\mathcal{F})}(p) + H_{C(\mathcal{K})}(p) = H(p)$ for all p on V;*

(iv) *there exists a perfect graph $G = (V, \mathcal{E})$ such that $\mathcal{F}$ and $\mathcal{K}$ are exactly the independent sets and the cliques of G, respectively.*

Proof (i)$\Leftrightarrow$(ii): By Corollary 4.1, $(\mathcal{F}, \mathcal{K})$ is generating iff $C(\mathcal{F})^* \subseteq C(\mathcal{K})$. On the other hand, $|F \cap K| \leq 1$ for every $F \in \mathcal{F}$ and $K \in \mathcal{K}$ means in polyhedral terms that $u^T v \leq 1$ holds for every vertex u of $C(\mathcal{F})$ and v of $C(\mathcal{K})$. This is clearly equivalent to the

same relation holding when u and v are arbitrary points in $C(\mathcal{F})$ and $C(\mathcal{K})$, respectively. This is equivalent to saying that $C(\mathcal{F})^* \supseteq C(\mathcal{K})$.

(ii)$\Leftrightarrow$(iii) by Corollary 2.4.

(i)$\Rightarrow$(iv): Define the graph $G = (V, \mathcal{E})$, connecting two vertices iff they are contained in a common $K \in \mathcal{K}$.Every $K \in \mathcal{K}$ becomes then a clique, whereas, by (i), every $F \in \mathcal{F}$ becomes an independent set. Hence

$$C(\mathcal{F}) \subseteq VP(G) \text{ and } C(\mathcal{K}) \subseteq VP(\overline{G}).$$

By (iii) and Lemma 3.1, we have for any probability distribution p on V

$$H(p) = H_{C(\mathcal{F})}(p) + H_{C(\mathcal{K})}(p) \geq H(G, p) + H(\overline{G}, p) \geq H(p),$$

i.e., G is strongly splitting, and so perfect. Moreover, by Lemma 2.2, $C(\mathcal{F}) = VP(G)$, i.e., the maximal independent sets of G coincide with the maximal sets in $\mathcal{F}$, and similarly for $\mathcal{K}$.

(iv)$\Rightarrow$(i) by Corollary 3.3. ∎

Finally, we mention an information theoretic interpretation of $H_{C(\mathcal{F})}(p)$, where $\mathcal{F}$ is a family of subsets of the set V (cf. [3], Chapter 2, §2). Let U be a finite set, and $d : V \times U \to \{0, 1\}$ a function called *distortion function*. We assume that, for every $i \in V$, there exists a $j \in U$ with $d(i, j) = 0$. Let us consider a discrete memoryless source emitting symbols from V according to a probability distribution p. A number R is called an *achievable rate* (at distortion level 0) if for every $\varepsilon > 0$ and sufficiently large n, there exists a function $f : V^n \to U^n$ (called *coding function*) such that the size of the range of f is at most 2^{nR}, and for every $x \in V^n$,

$$\Pr\{d(x, f(x)) > 0\} < \varepsilon.$$

Here

$$d(x, y) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i)$$

for $x = (x_1, \ldots, x_n) \in V^n$, $y = (y_1, \ldots, y_n) \in U^n$.

In data compression, one is interested in the infimum of the achievable rates. Let $R_d(p)$ denote this infimum. (In information theory, $R_d(p)$ is called the *value of the rate distortion function* of the source with probability distribution p, at distortion level 0.)

Now define the family $\mathcal{F} = \mathcal{F}_d$ as follows. Write

$$F_u = \{v \in V : \ d(v, u) = 0\}, \quad (u \in U)$$

and

$$\mathcal{F} = \mathcal{F}_d = \{X : \ X \subseteq F_u, \ u \in U\}.$$

Then we have

Lemma 4.3. $R_d(p) = H_{C(\mathcal{F}_d)}(p)$.

Proof. This is a generalization of Lemma 3.1. To prove it, we need the following generalization of (4):

$$R_d(p) = \min\{I(X \wedge Y) : \ \mathrm{dist}(X) = p, \ X \in Y \in \mathcal{F}_d\}. \tag{9}$$

(9) is an equivalent form of the following well known formula (c.f. [3 Chapter 2, §2]):

$$\begin{aligned} R_d(p) = \min\{&I(X \wedge Y) : \\ &\mathrm{dist}(X) = p, \ Y \text{ takes values in } U, \ E\{d(X,Y)\} = 0\}. \end{aligned}$$

(Here E denotes mathematical expectation.) Lemma 4.3 follows from (9) in exactly the same way as Lemma 3.1 from (4). ∎

Notice that $H_{C(\mathcal{F})}(p)$ equals the entropy of the "probabilistic hyperclub $(V, \mathcal{F}, p)$" defined in [11] as the right-hand-side of formula (9).

In the light of this interpretation of $H_{C(\mathcal{F}_d)}(p)$, Theorem 6 can be considered as a characterization of those pairs $(d, \overline{d})$ of $(0,1)$-valued distortion functions for which, given any source distribution p on V, there exists an essentially error-free loss-less two-step encoding of the corresponding discrete memoryless source in the sense of [10] and [13].

5. Additivity and sub-additivity

If $a \in \mathbb{R}_+^k$ and $b \in \mathbb{R}_+^l$ then their *tensorial product* $a \times b \in \mathbb{R}_+^{kl}$ is defined by

$$(a \times b)_{ij} = a_i \cdot b_j, \qquad i = 1, \dots, k, j = 1, \dots, l.$$

Note that if p and q are probability distribution then $p \times q$ is the usual product distribution. Recall that if $k = l$ then also the *dyadic product* $a \circ b \in \mathbb{R}_+^k$ is defined by

$$(a \circ b)_i = a_i \cdot b_i, \qquad i = 1, \dots, k.$$

Definition. Let $A \subseteq \mathbb{R}_+^k$ and $B \subseteq \mathbb{R}_+^l$ be convex corners. Their *tensorial product* $A \otimes B \subseteq \mathbb{R}_+^{kl}$ is the convex corner spanned by the tensorial products $a \times b$, $a \in A$, $b \in B$. The *dyadic product* $A \odot B$ of the convex corners $A, B \subseteq \mathbb{R}_+^k$ is the convex corner in that same space spanned by the vectors $a \circ b, a \in A, b \in B$. In other words, $A \odot B = \mathrm{conv}(A \circ B)$.

5.1 Theorem. (i) *Let $A \subseteq \mathbb{R}_+^k$, $B \subseteq \mathbb{R}_+^l$ be convex corners, and $p \in \mathbb{R}_+^k, q \in \mathbb{R}_+^l$ probability distributions. Then*

$$H_{A \otimes B}(p \times q) = H_A(p) + H_B(q) = H_{(A^* \otimes B^*)^*}(p \times q).$$

(ii) *Let $A, B \subseteq \mathbb{R}_+^k$ be convex corners, and $p \in \mathbb{R}_+^k$ a probability distribution. Then*

$$H_{A \odot B}(p) \leq H_A(p) + H_B(p).$$

Proof. (i) For $a \in A$, $b \in B$, we have $a \times b \in A \otimes B$, implying

$$\begin{aligned} H_{A \otimes B}(p \times q) &\leq -\sum_{i=1}^{k} \sum_{j=1}^{l} p_i q_j \log a_i b_i \\ &= -\sum_{i=1}^{k} p_i \log a_i - \sum_{j=1}^{l} q_j \log b_j. \end{aligned}$$

Hence $H_{A \otimes B}(p \times q) \leq H_A(p) + H_B(q)$. By Corollary 2.4,

$$H(p \times q) = H_{A \otimes B}(p \times q) + H_{(A \otimes B)^*}(p \times q).$$

Since obviously $A^* \otimes B^* \subseteq (A \otimes B)^*$, we obtain:

$$\begin{aligned} H(p \times q) &\leq H_{A \otimes B}(p \times q) + H_{A^* \otimes B^*}(p \times q) \\ &\leq H_A(p) + H_B(q) + H_{A^*}(p) + H_{B^*}(q) H(p) + H(q) \\ &= H(p \times q). \end{aligned} \tag{10}$$

We must have equality everywhere in (10), proving

$$H_{A \otimes B}(p \times q) = H_A(p) + H_B(q)$$

and

$$H_{(A \otimes B)^*}(p \times q) = H_{A^* \otimes B^*}(p \times q) = H_{A^*}(p) + H_{B^*}(q).$$

This proves (i). Statement (ii) is obvious. ∎

References

[1] C. Berge: *Graphs and hypergraphs*, 2nd edition. North Holland, Amsterdam, 1976.

[2] V. Chvátal: On certain polytopes associated with graphs. *J. Comb. Theory* B **18** (1975) pp. 138–154.

[3] I. Csiszár, J. Körner: *Information Theory. Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York 1982.

[4] D.R.Fulkerson: Blocking and anti-blocking pairs of polyhedra. *Math. Programming* **1** (1971), pp. 168–194.

[5] D.R. Fulkerson: On the perfect graph theorem. In: *Mathematical Programming* (T.C.Hu, S.M. Robinson, Eds), Academic Press, New York 1973, pp. 69–77.

[6] M. Grötschel, L.Lovász, A.Schrijver: Relaxations of vertex packing. *J.Comb.Theory* B **40** (1986) pp. 330–343.

[7] J. Körner: Coding of an information source having ambiguous alphabet and the entropy of graphs. In: *Transactions of the 6th Prague Conference on Information Theory, etc.*, Academia, Prague, 1973, pp. 411–425.

[8] J.Körner: An extension of the class of perfect graphs, *Studia Sci. Math. Hung.* **8** (1973), pp. 405–409.

[9] J.Körner: Fredman–Komlós bounds and information theory. *SIAM J. of Algebraic and Discrete Methods* **7** (1986), pp. 560–570.

[10] J. Körner, G. Longo: Two-step encoding of finite memoryless sources. *IEEE Trans. on Inform. Theory* **19** (1973), pp. 778–782.

[11] J.Körner, K.Marton: New bounds for perfect hashing via information theory. *European J. of Combinatorics*, to appear.

[12] J. Körner, K. Marton: Graphs that split entropy, *SIAM J. on Discrete Mathematics*, to appear.

[13] J. Körner, A. Sgarro: A new approach to rate distorsion theory, *Rend. Inst. Matem. Univ. di Trieste* **18** (1986), pp. 117–187.

[14] L. Lovász: On the Shannon capacity of a graph. *IEEE Trans. on Inform. Theory* **25** (1979) pp. 1–7.

[15] L. Lovász: Perfect graphs. In: *More Selected Topics in Graph Theory* (L.W. Beineke, R. J. Wilson, eds), Academic Press, New York–London, 1983, pp. 55–87.

[16] K. Marton: On the Shannon capacity of graphs within a given type; in preparation.

[17] R.J. McEliece: *The theory of information and coding.* Addison–Wesley, Reading, Mass. 1977.