Grafy a počty - NDMI078 April 2009

Lecture 5

M. Loebl *J.-S. Sereni*

The goal of this lecture is to present how the entropy of random variables can be used to obtain bounds on the number of combinatorial objects. This is illustrated by a proof of Brègman's theorem found by Radhakrishnan in the late nineties [5].

1 Basics on the entropy

We only present some basics concepts about entropy. We refer to the books by McEliece [3, 4] for a nice exposition of the topic. Simonyi wrote a survey on graph entropy [6], and another one devoted to the links between graph entropy and perfect graphs [7].

We consider only finite discrete probabilistic spaces. A *discrete probabilistic space* is a pair $(\mathscr{U}, p)$ where $\mathscr{U}$ is a finite set and $p : \mathscr{U} \to [0,1]$ satisfies $\sum_{u\in\mathscr{U}} p(u) = 1$. An *event* is a subset A of $\mathscr{U}$, and its *probability* is $\mathbf{Pr}(A) := \sum_{u\in A} p(u)$. A *random variable* is a mapping from $\mathscr{U}$ to some set.

Let X be a random variable taking values in a set $\mathscr{X}$. The *entropy* of X is

$$H(X) := \sum_{x\in\mathscr{X}} \mathbf{Pr}(X = x) \log \frac{1}{\mathbf{Pr}(X = x)}\,.$$

We let $0\cdot\log(1/0) := 0$ in the previous definition (or, alternately, we implicitly assume that the sum is taken only over the elements $x \in \mathscr{X}$ such that $P(X = x) > 0$).

The entropy can be sought as the amount of uncertainty the observer of a system is left with once (s)he knows that X has distribution $\mathbf{Pr}$. This can be explained as follows. Let $A \subseteq \mathscr{X}$. We want to associate to A a real number I_A that can be interpreted as the amount of information in the claim "$X \in A$". If one requires that I_A is a continuous function of the probability $\mathbf{Pr}(X \in A) = \sum_{x\in A} \mathbf{Pr}(X = x)$ and that $I_{A\cap B} = I_A + I_B$ for any two independent events A and B (i.e. such that $p(X \in A \cap B) = p(X \in A) \cdot p(X \in B)$), the only possible choice is $I_A = -\log \mathbf{Pr}(X \in A)$, where the logarithm can be taken to any base. The entropy of X thus models the average amount of information of the elementary claims $X = x$ for $x \in \mathscr{X}$.

Note that the *values* taken by X are not relevant, only the *probabilities* with which X takes those values are. Moreover, the image of X is a finite set (since $\mathscr{U}$ is).

If X is a 0-1 random variable being 0 with a fixed probability $p \in (0,1)$, then $\mathbf{E}(X)$ is the *binary entropy function*, i.e.

$$\mathbf{E}(X) = H(p) := -p\log p - (1-p)\log(1-p)\,.$$

Let Y be a random variable taking values in a set $\mathscr{Y}$. The *joint entropy* of the two random variables X and Y is

$$H(X,Y) = \sum_{\substack{x\in\mathscr{X}\\ y\in\mathscr{Y}}} \mathbf{Pr}(X=x,Y=y)\log\left(\frac{1}{\mathbf{Pr}(X=x,Y=y)}\right).$$

We can condition the entropy of a random variable on a particular observation, or more generally on the outcome of another random variable. The *conditional entropy of X given that $Y=y$* is

$$H(X|Y=y) = \sum_{x\in\mathscr{X}} \mathbf{Pr}\left(X=x|Y=y\right)\log\left(\frac{1}{\mathbf{Pr}\left(X=x|Y=y\right)}\right).$$

The *conditional entropy of X given Y* is the average of the preceding, i.e.

$$\begin{aligned} H(X|Y) &:= \sum_{y\in\mathscr{Y}} \mathbf{Pr}(Y=y)H(X|Y=y) \\ &= \sum_{\substack{x\in\mathscr{X}\\ y\in\mathscr{Y}}} \mathbf{Pr}\left(X=x,Y=y\right)\log\left(\frac{1}{\mathbf{Pr}\left(X=x|Y=y\right)}\right). \end{aligned}$$

Let us see some relations between those quantities.

Proposition 1. *Let X and Y be two random variables taking values in $\mathscr{X}$ and $\mathscr{Y}$, respectively.*

(i) $H(X) \le \log(|\mathscr{X}|)$ *with equality if and only if X is uniformly distributed.*

(ii) $H(X,Y) = H(X) + H(Y|X)$.

(iii) $H(X,Y) \le H(X)+H(Y)$ *with equality if and only if X and Y are independent.*

(iv) $H(X|Y) \le H(X)$ *with equality if and only if X and Y are independent.*

Before starting the proof, we recall that by Jensen's equality for concave functions,

$$\sum_i \alpha_i \log(\beta_i) \le \log\left(\sum_i \alpha_i\beta_i\right) \tag{1}$$

for all positive reals such that $\sum_i \alpha_i = 1$. Moreover, there is equality if and only if $\beta_i = \beta_j$ for any i,j.

Proof of Proposition 1:
(i) Jensen's inequality implies that

$$\begin{aligned} H(X) &= \sum_{x\in\mathscr{X}} \mathbf{Pr}(X=x)\log\left(\frac{1}{\mathbf{Pr}(X=x)}\right) \\ &\leq \log\left(\sum_{x\in\mathscr{X}} \mathbf{Pr}(X=x)/\,\mathbf{Pr}(X=x)\right) \\ &= \log(|\mathscr{X}|)\,, \end{aligned}$$

with equality if and only if $\mathbf{Pr}(X=x) = \mathbf{Pr}(X=x')$ for all $x, x' \in \mathscr{X}$, i.e. if and only if X is uniformly distributed.
(ii) Since $\mathbf{Pr}(X=x, Y=y) = \mathbf{Pr}(X=x)\cdot\mathbf{Pr}(Y=y|X=x)$ and $\mathbf{Pr}(X=x) = \sum_{y\in Y}\mathbf{Pr}(X=x,Y=y)$, we deduce that

$$\begin{aligned} H(X,Y) - H(Y|X) &= \sum_{x,y} \mathbf{Pr}(X=x,Y=y)\log\left(\frac{1}{\mathbf{Pr}(X=x,Y=y)}\right) \\ &\quad - \sum_{x,y} \mathbf{Pr}(X=x,Y=y)\log\left(\frac{1}{\mathbf{Pr}(Y=y|X=x)}\right) \\ &= \sum_{x,y} \mathbf{Pr}(X=x,Y=y)\log\left(\frac{\mathbf{Pr}(X=x,Y=y)}{\mathbf{Pr}(X=x)\cdot\mathbf{Pr}(X=x,Y=y)}\right) \\ &= \sum_{x\in\mathscr{X}} \log\left(\frac{1}{\mathbf{Pr}(X=x)}\right)\cdot\sum_{y\in\mathscr{Y}}\mathbf{Pr}(X=x,Y=y) \\ &= H(X)\,. \end{aligned}$$

(iii) Since $\mathbf{Pr}(X=x) = \sum_{y\in\mathscr{Y}}\mathbf{Pr}(X=x,Y=y)$,

$$H(X) + X(Y) = -\sum_{x,y}\mathbf{Pr}(X=x,Y=y)\log\left(\mathbf{Pr}(X=x)\cdot\mathbf{Pr}(Y=y)\right)\,.$$

Consequently, using Jensen's inequality we obtain

$$\begin{aligned} H(X,Y) - (H(X)+H(Y)) &= \sum_{x,y}\mathbf{Pr}(X=x,Y=y)\log\left(\frac{\mathbf{Pr}(X=x)\cdot\mathbf{Pr}(Y=y)}{\mathbf{Pr}(X=x,Y=y)}\right) \\ &\leq \log\left(\sum_{x,y}\mathbf{Pr}(X=x)\cdot\mathbf{Pr}(Y=y)\right) \\ &= \log 1 = 0\,, \end{aligned}$$

with equality if and only if X and Y are independent.

(iv) By (ii) and (iii)

$$H(X|Y) - H(X) = H(X,Y) - H(Y) - H(X) \leq 0\,,$$

with equality if and only if X and Y are independent. □

By induction, (1) generalises to the so-called *chain rule*, i.e.

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})\,. \tag{2}$$

We end by presenting a useful lemma [1] with a small application. If $X = (X_i)_{i\in\mathscr{I}}$ is a vector and A a subset of $\mathscr{I}$, we set $X_A := (X_i)_{i\in A}$.

Lemma 2 (Shearer, 1986). *Let $X = (X_1, X_2, \ldots, X_n)$ be a random variable and let $\mathscr{A} = \{A_i\}_{i\in\mathscr{I}}$ be a collection of subsets of $\{1, 2, \ldots, n\}$ such that each integer $i \in \{1, 2, \ldots, n\}$ belongs to at least k sets of $\mathscr{A}$. Then*

$$H(X) \leq \frac{1}{k} \sum_{i\in\mathscr{I}} H\left(X_{A_i}\right)\,.$$

Proof. By the chain rule, $H(X) = \sum_{i=1}^{n} H(X_i|X_j : j < i)$. On the other hand, for each $i \in \mathscr{I}$,

$$\begin{aligned} H(X_{A_i}) &= \sum_{j\in A_i} H(X_j|X_s : s < j \text{ and } s \in A_i) \\ &\geq \sum_{j\in A_i} H(X_j|X_s : s < j)\,, \end{aligned}$$

by Proposition 1(iv). Summing the last inequality over all indices $i \in \mathscr{I}$, we obtain

$$\begin{aligned} \sum_{i\in\mathscr{I}} H\left(X_{A_i}\right) &\geq \sum_{i\in\mathscr{I}} \sum_{j\in A_i} H(X_j|X_s : s < j) \\ &\geq k \cdot \sum_{j=1}^{n} H(X_j|X_s : s < j) \\ &= k \cdot H(X)\,. \end{aligned}$$

□

The following geometric proposition illustrates the use of entropy to obtain bounds via Shearer's lemma.

Proposition 3. *Let $\mathscr{P}_1, \mathscr{P}_2$ and $\mathscr{P}_3$ be the hyperplanes (x, y), (x, z) and (y, z) of $\mathbf{R}^3$, respectively. If n points of $\mathbf{R}^3$ are have exactly n_i different projections on $\mathscr{P}_i$ for $i \in \{1, 2, 3\}$, then*

$$n_1 n_2 n_3 \geq n^2\,.$$

Proof. Let us choose uniformly at random a point among the n points given. We consider the random variable $P = (X, Y, Z)$ corresponding to the three coordinates of the chosen point. By Proposition 1(i), it holds that $H(P) = \log n$. Let us consider the sets $A_i := \{i, i+1\}$ for $i \in \{1, 2\}$ and the set $A_3 := \{1, 3\}$. Every index is in two of the three sets, thus Shearer's lemma implies that

$$2 \cdot H(P) \le H(X) + H(Y) + H(Z) \le \log n_1 + \log n_2 + \log n_3 \,.$$

Therefore, $2 \cdot \log n \le \log n_1 + \log n_2 + \log n_3$, i.e. $n^2 \le n_1 n_2 n_3$. □

2 Radhakrishnan's proof of Brègman's theorem

Let us state Brègman's theorem in terms of the number of perfect matchings in a bipartite graph.

Theorem 4 (Brègman, 1973)**.** *Let G be a bipartite graph with parts A and B. The number of perfect matchings of G is at most*

$$\prod_{v \in A} (\deg(v)!)^{1/\deg(v)} \,.$$

Proof. Let G be a bipartite graph with parts A and B. We define $\mathscr{M}$ to be the set of all the perfect matchings of G, and we suppose that $\mathscr{M} \neq \emptyset$, otherwise the statement of the theorem holds trivially. In particular, $|A| = |B|$; let us set $n := |A|$. For a perfect matching M and a vertex $a \in A$, we let $M(a)$ be the vertex of B that is adjacent to a in M. Further, for every vertex $b \in B$, we let $M^{-1}(b)$ be the vertex of A that is adjacent to b in M.

We choose a perfect matching $M \in \mathscr{M}$ uniformly at random. Thus, $\log |\mathscr{M}| = H(M)$. Let $a_1, a_2, \ldots, a_n$ be an ordering of the vertices of A. Then, by the chain rule (2),

$$\begin{aligned} H(M) =& H(M(a_1)) + H(M(a_2)|M(a_1)) \\ &+ \ldots + H(M(a_n)|M(a_1), M(a_2), \ldots, M(a_{n-1})) \,. \end{aligned} \tag{3}$$

Note that this equation yields the trivial upper bound $|\mathscr{M}| \le \prod_{a \in A} \deg(a)$. Indeed, the conditional entropy of $M(a_i)$ given $M(a_1), M(a_2), \ldots, M(a_{i-1})$ is at most the entropy of $M(a_i)$ (by Proposition 1(iv)), which in turn is at most $\log \deg(a_i)$ (by Proposition 1(i)). We would obtain a better upper bound on $|\mathscr{M}|$ if we manage to infer a better upper bound on $H(M(a_i)|M(a_1), M(a_2), \ldots, M(a_{i-1}))$.

To this end, note that the range of $M(a_i)$ given $M(a_j)$ for $j \in \{1, 2, \ldots, i-1\}$ is actually contained in $N_G(a_i) \setminus \{M(a_1), M(a_2), \ldots, M(a_{i-1})\}$. So, it may well be smaller than $\deg(a_i)$. Moreover, its size depends on the ordering chosen for the vertices of A.

To exploit this remark, let σ be a permutation of $\{1, 2, \ldots, n\}$, chosen uniformly at random. For each index $i \in \{1, 2, \ldots, n\}$, we set

$$R_i(M, \sigma) := |N_G(a_i) \setminus \{M(a_{\sigma(1)}), \ldots, M(a_{\sigma(k-1)})\}| \,,$$

with $k := \sigma^{-1}(i)$. Observe that, for every integer $j \in \{1, 2, \ldots, \deg(a_i)\}$,

$$\mathop{\mathbf{Pr}}_{M,\sigma}\left(R_i(M,\sigma) = j\right) = \frac{1}{\deg(a_i)} \,. \tag{4}$$

Indeed, for any fixed matching M,

$$\mathop{\mathbf{Pr}}_{\sigma}\left(R_i(M,\sigma) = j|M\right) = \frac{1}{\deg(a_i)} \,, \tag{5}$$

since σ is chosen uniformly at random. In fact, (5) can also be proved, for instance, by counting directly: the number of permutations such that $\alpha = \deg(a_i) - j$ vertices of $M^{-1}(N_G(a_i))$ occur before a_i is

$$\begin{aligned}
&\sum_{k=1}^{n} \binom{\deg(a_i) - 1}{\alpha} \binom{n - \deg(a_i)}{k - \alpha - 1} (k-1)!(n-k)! \\
=&(\deg(a_i) - 1)!(n - \deg(a_i))! \cdot \sum_{k=1}^{n} \binom{k-1}{\alpha} \binom{n-k}{\deg(a_i) - \alpha - 1} \\
=&\frac{n!}{\deg(a_i) \cdot \binom{n}{\deg(a_i)}} \cdot \sum_{k=0}^{n-1} \binom{k}{\alpha} \binom{n-1-k}{\deg(a_i) - \alpha - 1} \\
=&\frac{n!}{\deg(a_i)} \,,
\end{aligned}$$

where the last line follows from the following classical binomial identity [2, p. 129].

$$\sum_{k=0}^{n-1} \binom{k}{j} \binom{n-1-k}{d-j-1} = \binom{n}{d} \,.$$

Now, (5) implies (4) by averaging over all $M \in \mathscr{M}$, i.e.

$$\mathop{\mathbf{Pr}}_{M,\sigma}\left(R_i(M,\sigma) = j\right) = \sum_{M} \mathbf{Pr}(M) \cdot \mathop{\mathbf{Pr}}_{\sigma}\left(R_i(M,\sigma) = j|M\right) = \frac{1}{\deg(a_i)} \,.$$

On the other hand, applying Proposition 1(i), we obtain

$$H(M(a_i)|M(a_{\sigma(1)}), \ldots, M(a_{\sigma(\sigma^{-1}(i)-1)})) \leq \sum_{j=1}^{\deg(a_i)} \mathop{\mathbf{Pr}}_{M}\left(R_i(M,\sigma) = j\right) \cdot \log j \,. \tag{6}$$

Furthermore, (3) translates to

$$
\begin{aligned}
H(M) =& H(M(a_{\sigma(1)})) + H(M(a_{\sigma(2)})|M(a_{\sigma(1)})) \\
&+ \ldots + H(M(a_{\sigma(n)})|M(a_{\sigma(1)}), M(a_{\sigma(2)}), \ldots, M(a_{\sigma(n-1)})) \,. \qquad (7)
\end{aligned}
$$

Summing (7) over all the permutations σ, we obtain

$$
n! H(M) = \sum_{\sigma} \sum_{i=1}^{n} H\left(M(a_{\sigma(i)})|M(a_{\sigma(1)}), \ldots, M(a_{\sigma(i-1)})\right) \,,
$$

i.e.

$$
H(M) = \mathop{\mathbf{E}}_{\sigma}\left[\sum_{i=1}^{n} H\left(M(a_{\sigma(i)})|M(a_{\sigma(1)}), \ldots, M(a_{\sigma(i-1)})\right)\right] \,.
$$

We write the terms of the sum in a different order, and use the linearity of Expectation.

$$
\begin{aligned}
H(M) &= \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\sigma}\left[H\left(M(a_i)|M(a_{\sigma(1)}), \ldots, M(a_{\sigma(\sigma^{-1}(i)-1)})\right)\right] \\
&\leq \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\sigma}\left[\sum_{j=1}^{\deg(a_i)} \mathop{\mathbf{Pr}}_{M}\left(R_i(M,\sigma) = j\right) \cdot \log j\right] && \text{by (6)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{\deg(a_i)} \sum_{\sigma} \mathbf{Pr}(\sigma) \mathop{\mathbf{Pr}}_{M}\left(R_i(M,\sigma) = j\right) \cdot \log j \,.
\end{aligned}
$$

Observe that

$$
\sum_{\sigma} \mathbf{Pr}(\sigma) \mathop{\mathbf{Pr}}_{M}\left(R_i(M,\sigma) = j\right) = \mathop{\mathbf{Pr}}_{M,\sigma}(R_i(M,\sigma) = j) \,.
$$

Thus, (4) implies that

$$
\begin{aligned}
H(M) &\leq \sum_{i=1}^{n} \sum_{j=1}^{\deg(a_i)} \frac{1}{\deg(a_i)} \cdot \log j \\
&= \sum_{i=1}^{n} \log\left(\deg(a_i)!\right)^{1/\deg(a_i)} \,,
\end{aligned}
$$

which concludes the proof. □

We conclude by explicitly stating some key points when trying to bound the size of a set $\mathscr{M}$ using entropy. We choose an element M of $\mathscr{M}$ uniformly at random, so that $H(M) = \log|\mathscr{M}|$. The goal is then to bound the entropy. To this end, the chain rule and Shearer's lemma are crucial tools.

References

[1] F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Combin. Theory Ser. A*, 43(1):23–37, 1986.

[2] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics*. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1989. A foundation for computer science.

[3] R. J. McEliece. *The theory of information and coding*, volume 86 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 2002.

[4] R. J. McEliece. *The theory of information and coding*, volume 86 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, student edition, 2004. With a foreword by Mark Kac.

[5] J. Radhakrishnan. An entropy proof of Bregman's theorem. *J. Combin. Theory Ser. A*, 77(1):161–164, 1997.

[6] G. Simonyi. Graph entropy: a survey. In *Combinatorial optimization (New Brunswick, NJ, 1992–1993)*, volume 20 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 399–441. Amer. Math. Soc., Providence, RI, 1995.

[7] G. Simonyi. Perfect graphs and graph entropy. An updated survey. In *Perfect graphs*, Wiley-Intersci. Ser. Discrete Math. Optim., pages 293–328. Wiley, Chichester, 2001.