Barcelona Graduate School of Mathematics, Fall 2013

Random Structures and the Probabilistic Method

Entropy and Shannon theorem

---

1. Entropy and set systems

The notion of *entropy* is central in information theory and it has found many applications in other areas. For a discrete random variable X taking values in a finite set S with probability function $p(i) = \Pr(X = i)$, the entropy of X is (here and in the sequel log stands for $\log_2$):

$$H(X) = -\mathbb{E}(\log p(X)) = -\sum_{i \in S} p(i) \log p(i).$$

In particular, for a Bernouilli random variable $X \sim B(p)$ we write

$$H(p) := H(X) = -p \log p - (1-p) \log(1-p).$$

For a pair X, Y of discrete random variables taking values in S and T with joint distribution $p(i,j) = \Pr(X = i, Y = j)$,

$$H(X,Y) = -\sum_{i,j} p(i,j) \log p(i,j),$$

and

$$H(X|Y) = H(X,Y) - H(Y).$$

Lemma 1.1. *Let X, Y be discrete random variables taking values in the finite sets S and T respectively. We have*

$$H(X) \leq \log |S|$$

with equality if and only if X is uniform on S. Moreover,

$$H(X) \leq H(X,Y) \leq H(X) + H(Y).$$

By the concavity of log and $\sum_i p_i = 1$ we have $H(X) = \sum_i p_i \log(1/p_i) \leq \log(\sum_i p_i(1/p_i)) = \log |S|$. By monotonicity of log we have $H(X,Y) = -\sum_{i,j} p_{ij} \log(p_{ij}) \geq -\sum_{i,j} p_{ij} \log p_i = -\sum_i p_i \log p_i = H(X)$. Finally,

$$H(X,Y) - H(X) - H(Y) = -\sum_{i,j} p_{ij} \log(p_{ij}/p_i p_j) = -\sum_{ij} p_i p_j z_{ij} \log z_{ij},$$

and from the concavity of $f(z) = z \log z$ the last term is at least $f(\sum_{ij} p_i p_j z_{ij}) = f(1) = 0$.

Some applications of the entropy function to set systems and geometry rely in the following result by Shearer. We use the following notation. For a subset $I \subset [n]$ we denote by $X(I) = (X_i : i \in I)$. By convention $H(X(\emptyset)) = 0$. We note the following facts.

Lemma 1.2. *It holds*

(1) *If $I \subset I'$ then $H(X(I)) \subset H(X(I'))$.*

(2) *For two subsets I, J, $H(X(I \cup J)) \leq H(X(I)) + H(X(J)) - H(X(I \cap J))$.*

Theorem 1.3 (Shearer, 1986)**.** *Let $X_1, \ldots, X_n$ be random variables, each taking values in a finite set S_i. Let $\mathcal{I}$ be a family of subsets of $[n] = \{1, 2, \ldots, n\}$. If each i belongs to at least k sets in $\mathcal{I}$ then*

$$H(X_1, \ldots, X_n) \leq \frac{1}{k} \sum_{I \in \mathcal{I}} H(X(I)),$$

When $k = 1$ we can reduce our family $\mathcal{G}$ of sets to a partition $\mathcal{G}'$ of $\{1, \ldots, n\}$. Since $H(X(G)) \geq H(X(G'))$ when $G \supset G'$, we get $\sum_{G \in \mathcal{G}} X(G) \geq \sum_{G' \in \mathcal{G}'} X(G') \geq H(X)$, where the last inequality follows from the above Lemma. Suppose now the result true for $k - 1$. If one of the sets is $G = \{1, \ldots, n\}$ then we are set.

One application is a proof of the following geometric result.

Theorem 1.4 (Loomis and Whitney, 1949)**.** *Let B be a measurable set in $\mathbb{R}^n$. For each i denote by $\pi_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$ the projection on the coordinates different from i. Then*

$$(vol(B))^{n-1} \leq \prod_{i=1}^{n} vol(\pi_i(B)).$$

The above theorem follows from the following discrete version.

Theorem 1.5. *Let $\mathcal{F}$ be a family of vectors in $S_1 \times \cdots \times S_n$. Let $\mathcal{I}$ be a family of subsets of $\{1, 2, \ldots, n\}$ such that each i belongs to at least k sets in $\mathcal{I}$. Then*

$$|\mathcal{F}|^k \leq \prod_{I \in \mathcal{I}} |\mathcal{F}_I|,$$

where $\mathcal{F}_I = \{\pi_I(v) : v \in \mathcal{F}\}$ and π_I denotes the projection on the coordinates $I \in \mathcal{I}$.

Consider the random variable $X = (X_1, \ldots, X_n)$ with the probability distribution $\Pr(X = \mathbf{1}_F) = 1/|\mathcal{F}|$ for each $F \in \mathcal{F}$, where $\mathbf{1}_F$ denotes the characteristic vector of F in $\{0, 1\}^n$. Each component X_i follows a Bernouilli distribution $B(p_i)$, and $H(X) = |\mathcal{F}|$, because X is uniform on $\mathcal{F}$. The inequality follows from $H(X) \leq \sum_{i=1}^{n} H(X_i)$.

---

2. Shannon theorem

A *coding scheme* is a pair of functions $f : \mathbb{Z}_2^m \to \mathbb{Z}_2^n$ and $g : \mathbb{Z}_2^n \to \mathbb{Z}_2^m$. The *rate of transmission* of the code scheme is $r(f, g) = m/n$. Let $e \in \mathbb{Z}_2^n$ be chosen randomly, each coordinate being 1 with probability p independently. Then, the *probability of correctness* of the coding scheme is $p(f, g) = \Pr(g(f(x) + e) = x)$ where x is randomly chosen in $\mathbb{Z}_2^m$ with the uniform distribution and independent of e. Shannon's celebrated theorem states that there are coding schemes with arbitrarily small probability of incorrect transmission and still with bounded rate of transmission.

The use of the entropy function arises from the following asymptotic estimations:

$$\binom{n}{np} = 2^{n(H(p)+o(1))},$$

and, for $0 < p < 1/2$,

$$\sum_{i \leq np} \binom{n}{i} \leq (1 + pn)\binom{n}{pn} = 2^{n(H(p)+o(1))}.$$

Theorem 2.1 (Shannon, 1948)**.** *Let $p \in (0, 1/2)$. For every fixed $\epsilon > 0$ there is a coding scheme with rate of ransmission at least $1 - H(p) - \epsilon$ and probability of incorrect transmission at most ϵ. Moreover there is a coding scheme in which f can be chosen to be linear.*

Actually Shannon theorem is complemented by the result that a larger rate of transmission forces a lower bound on the probability of incorrect transmission, so that $1 - H(p)$ can be seen as the optimal rate of a code with arbitrary small error probability.

We first fix $\delta > 0$ such that $p + \delta < 1/2$ and $H(p + delta) < H(p) + \epsilon/2$. For n large we set $m = n(1 - H(p) - \epsilon)$. This will effectively set the rate of transmission. Choose a random function $f : \mathbb{Z}_2^m \to \mathbb{Z}_2^n$. Let S be a ball of radius $n(p + \delta)$ (with the Hamming distance). For $x \in \mathbb{Z}_2^m$ let A_x denote the event that $f(x) \in S$. We have $\Pr(A_x) = |S|2^{-n}$ and

$$\Pr(\cup_{x\in\mathbb{Z}_2^m} A_x) \leq \sum_{x\in Z_2^m} \Pr(A_x) = 2^{m-n}|S| = 2^{-n(H(p)+\epsilon)} \sum_{i\leq n(p+\delta)} \binom{n}{i} \leq 2^{n(H(p+\delta)-H(p)-\epsilon+o(1))} = 2^{-n(\epsilon/2+o(1))}.$$

We define the decoding function $g : \mathbb{Z}_2^n \to \mathbb{Z}_2^m$ as follows. For each $y = f(x) + e$ we choose x' such that $f(x')$ is in the ball of radius $n(p + \delta)$ centered at y if such an x' exists and is unique. Otherwise we define $g(y) = 0$. The probability of error is either that such an x' exists but is different from x, which means that e has weight larger than $n(p+\delta)$, and this occurs with an exponentially small probability (by Chernoff bound, for instance), or that there are more than one image of f within the ball S, but his occurs with probability at most $2^{-n(\epsilon/2+o(1))}$, also exponentially small with n. It follows that the probability of error can be made smaller than ϵ for each y if n is sufficiently large. If this happens with a random choice of f, it follows that there must be a specific choice of f for which the same bound for the probability error applies. In order to see that f can be chosen to be linear, we now choose f randomly on a base $u_1, \ldots, u_m$ and extend its values linearly. The probability that a ball of radius $n(p + \delta)$ contains two images of f, $|f(x) - f(x')| = |f(x - x')| \leq n(p + \delta)$ is the same as the probability that $z = x - x'$ lies in the ball of the same radius on the origin. But $f(z)$ is also distributed uniformly in$\mathbb{Z}_2^n$ and the above argument applies.

---

3. Exercises

(1) Let $\mathcal{F}$ and $\mathcal{G}$ be two families of subsets of $[n]$. Suppose that each i belongs to at least k members of $\mathcal{G}$. For each $G \in \mathcal{G}$ define $\mathcal{F}_G = \{F \cap G : F \in \mathcal{F}\}$. Prove that

$$|\mathcal{F}|^k \leq \prod_{G\in\mathcal{G}} |\mathcal{F}_G|.$$

(2) Prove the following result of Chung, Frankl, Graham and Shearer. Let $\mathcal{G}$ be a family of labelled graphs on $2n$ vertices. If every pair of graphs in $\mathcal{G}$ have a triangle in common then

$$|\mathcal{G}| \leq \frac{1}{4} 2^{\binom{2n}{2}}.$$

In order to prove the result, consider the graphs as sets of elements in $\binom{[n]}{2}$ (edges). Consider the family $\mathcal{I}$ consisting of the edges of two disjoint complete subgraphs with n vertices each one., for all partitions of the $2n$ vertices into two parts. Now note that, since every pair of graphs in $\mathcal{G}$ have a triangle in common (and the complement of the union of two complete graphs is bipartite) they must share one edge in one of the two complete graphs. Now apply the above exercise.