Homework Set No. 7 – Probability Theory (235A), Fall 2013

Due: 11/18/13 at discussion section

1. Aliens on the planet Mars communicate in a binary language with two symbols, 0 and 1. A text of length n symbols written in the Martian language looks like a sequence $X_1, X_2, \ldots, X_n$ of i.i.d. random symbols with the Bernoulli distribution $\mathrm{Ber}(p)$. Here, $p \in (0,1)$ is a parameter (the "Martian bias").

Define the **entropy function** $H(p)$ by

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

The graph of $H(p)$ is shown in the figure below.

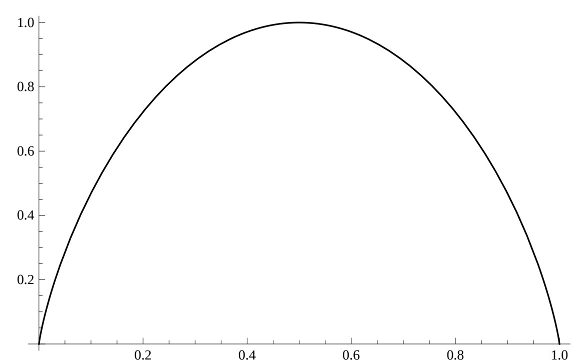

Figure 1: Graph of the entropy function $h(p)$

The goal of this problem is to prove the following result, which states loosely that if n is large, then with high probability a Martian text of length n can be encoded into an ordinary (human-made) computer file of length approximately $n \cdot H(p)$ computer bits (note that if $p \neq 1/2$ then this is smaller than n, meaning that the text can be compressed by a linear factor $H(p)$; for example in the case $p = 0.3$ we have $H(p) \approx 0.881$, giving a compression ratio of around 88%).

Theorem. *Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. Martian symbols (i.e., Bernoulli variables with bias p). Denote by $\boldsymbol{T}_n = (X_1, \ldots, X_n)$ the Martian text comprising the first n*

symbols. For any $\epsilon > 0$, if n is sufficiently large, the set $\{0,1\}^n$ of possible texts of length n can be partitioned into two disjoint sets,

$$\{0,1\}^n = A_n \cup B_n,$$

such that the following statements hold:

1. $\mathbf{P}(\boldsymbol{T}_n \in B_n) < \epsilon$
2. $2^{n(H(p)-\epsilon)} \le |A_n| \le 2^{n(H(p)+\epsilon)}$.

Notes: The texts in B_n can be thought of as the "exceptional sequences" – they are the Martian texts of length n that are observed only rarely (with probability less than ϵ). The texts in A_n are called "typical sequences". Because of the upper bound the theorem gives on the number of typical sequences, it follows that we can encode them in a computer file of size approximately $n(H(p)+\epsilon)$ bits, provided we prepare in advance a "code" that translates the typical sequences to computer files of the appropriate size (this can be done algorithmically, for example by making a list of typical sequences sorted in lexicographic order, and matching them to successive binary strings of length $(H(p)+\epsilon)n$). Conversely, the lower bound on $|A_n|$ implies that we cannot encode the typical sequences using less than $n(H(p)-\epsilon)$ bits.

To prove the theorem, let P_n be the random variable given by

$$P_n = \prod_{k=1}^{n} \left(p^{X_k}(1-p)^{1-X_k}\right).$$

Note that P_n measures the probability of the sequence that was observed up to time n. (Somewhat unusually, in this problem the probability itself is thought of as a random variable). Then proceed as follows:

(a) Represent P_n in terms of cumulative sums of a sequence of i.i.d. random variables.

(b) Apply the Weak Law of Large Numbers to that sequence, and see where that gets you.

2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with the exponential distribution $\mathrm{Exp}(1)$, and denote $S_n = \sum_{k=1}^{n} X_k$. For each $0 < p < 1$, let T_p be a random variable with the geometric distribution $T_p \sim \mathrm{Geom}(p)$, chosen independently from the X_k's.

(a) Compute explicitly the distribution of the random variable $S_{T_p} = \sum_{k=1}^{T_p} X_k$, a sum of a random number of random variables.

(b) (Optional) Prove the convergence in probability $\frac{S_{T_p}}{T_p} \xrightarrow[p\to 0]{\text{prob.}} 1$. (Note that this is a version of the weak law of large numbers for a sum of a randomly varying number of i.i.d. components.)