Homework due: Monday 3/19/12 at noon. Submit by email or by bringing the HW to my office, MSB 2218 (slide it under the door if I'm not there).

Problems

1. Prove that entropy is a concave function of probability vectors, i.e., satisfies the inequality

$$H(\alpha\mathbf{p} + (1-\alpha)\mathbf{q}) \geq \alpha H(\mathbf{p}) + (1-\alpha)H(\mathbf{q}).$$

2. (a) Prove that the entropy of a geometric random variable $X \sim \text{Geom}(p)$ is given by the formula

$$H(X) = \frac{1}{p}H(p, 1-p).$$

Can you think of an intuitive explanation for this identity? (See also problem 5 below for a possible clue.)

(b) Show that $H(X)$ for the above geometric r.v. maximizes the entropy $H(Z)$ among all random variables taking values in the positive integers and satisfying $\mathbf{E}(Z) = \mathbf{E}(X)$.

Hint. Use Gibbs's inequality.

3. (a) Let $f : \mathbb{R} \to [0,\infty)$ be a probability density function, and let $g : \mathbb{R} \to [0,\infty)$ be a *sub-probability* density function, i.e., a nonnegative function satisfying $\int g(x)\,dx \leq 1$. Prove that

$$-\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \leq -\int_{-\infty}^{\infty} f(x)\log g(x)\,dx$$

(note that both sides of the inequality may be infinite).

(b) The quantity $-\int_{-\infty}^{\infty} f(x)\log f(x)\,dx$ is a continuous version of the entropy of a probability distribution, and is denoted by $H(f)$, or $H(X)$ if X is a random variable such that $f = f_X$. (In this context, the convention is to use the natural logarithm rather than the logarithm to base 2, since the entropy no longer has the meaning of measuring bits.)

Let $X \sim N(\mu, \sigma^2)$ be a normal r.v. with mean μ and variance σ^2. In analogy with the result of problem 2(b) above, prove that $H(X) \geq H(Z)$ for any absolutely continuous random variable Z satisfying $\mathbf{E}(Z) = \mu$, $\mathbf{V}(Z) = \sigma^2$.

Note. The **maximum entropy principle** is a principle in statistics that says that if one has partial information about a probability distribution, it is natural (in some senses) to assume that the distribution is the one that has maximal entropy subject to the known information. For example, given a distribution on d symbols with no additional information, it makes sense to assume that it is the uniform distribution. Problems 2(b) and 3(b) show that the geometric and normal distributions are both natural entropy-maximizing distributions subject to simple constraints and therefore are likely to appear in many applications—as they indeed do.

4. The goal of this problem is to show that any uniquely decodable code can be replaced by a prefix code with the same word lengths. Let $C = \{w_1, \ldots, w_d\} \subset \{0,1\}^* = \cup_{m=1}^{\infty}\{0,1\}^m$ be a uniquely decodable code, with $\text{length}(w_j) = \ell_j$, $j = 1, \ldots, d$.

(a) Consider the polynomial $G(x) = \sum_{j=1}^{d} x^{\ell_j}$. Explain why for each $k \geq 1$, the coefficients of the polynomial $G(x)^k$ give for each power x^m the number of binary strings of length m that can be formed by concatenating k words from the code C.

(b) Explain why this implies that for $x > 0$, $G(x)^k \leq \sum_{i=1}^{kL} 2^i x^i$, where $L = \max(\ell_1, \ldots, \ell_d)$ is the maximal length of a code word. Set $x = 1/2$ and let $k \to \infty$ to deduce that the word lengths $\ell_1, \ldots, \ell_d$ satisfy Kraft's inequality

$$\sum_{j=1}^{d} 2^{-\ell_j} \leq 1.$$

(c) Infer (no need to write this) using the converse to Kraft's inequality that we proved in class that a prefix code with word lengths $\ell_1, \ldots, \ell_d$ exists.

5. (Optional problem) Let $\mathbf{p} = (p_1, \ldots, p_d)$ be a probability vector. A *simulation method for the discrete distribution* $\mathbf{p}$ *using unbiased coin tosses* is a (possibly infinite) prefix code

$$C \subset \{0,1\}^*,$$

together with a function $f : C \to \{1, \ldots, d\}$, such that for any $1 \leq j \leq d$ we have

$$\sum_{w \in C \,:\, f(w)=j} 2^{-\text{length}(w)} = p_j,$$

(so, in particular, $\sum_{w \in C} 2^{-\text{length}(w)} = \sum_j p_j = 1$). The idea is that the code defines an almost surely finite stopping time T on a sequence $X_1, X_2, \ldots$ of i.i.d. Bernoulli(1/2) random variables by

$$T = \min\{k \geq 1 \,:\, (X_1, \ldots, X_k) \in C\},$$

and then the random variable $Y = f((X_1, \ldots, X_T))$ (the output of the simulation) is distributed according to $\mathbf{p}$.

Prove the following analogue of the noiseless coding theorem to simulations, which gives an alternative interpretation of the meaning of entropy:

Theorem (Simulation theorem, Knuth-Yao 1976)**.** *(i) For any simulation method of* $\mathbf{p}$ *using unbiased coin tosses, the expected stopping time (i.e., the average number of coin tosses needed to simulate* $\mathbf{p}$*) satisfies*

$$\mathbf{E}(T) = \sum_{w \in C} \text{length}(w) 2^{-\text{length}(w)} \geq H(\mathbf{p}).$$

(ii) It is possible to find a simulation method of $\mathbf{p}$ *for which the expected stopping time satisfies*

$$\mathbf{E}(T) \leq H(\mathbf{p}) + 2.$$

Note. If you don't feel like figuring out the proof yourself, you can read about it in Section 5.11 of the book *Elements of Information Theory, 2nd ed.*, by Cover and Thomas. A generalization of the theorem that deals with simulation of $\mathbf{p}$ that uses i.i.d. samples from an arbitrary discrete distribution $\mathbf{q}$, instead of unbiased coin tosses, is proved in my paper *Sharp entropy bounds for discrete statistical simulation.*