ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 24 Information Theory and Coding
Midterm Solutions January 13, 2004

PROBLEM 1.

(a) $E[X] = \sum_m p(m)m = \sum_m q(m)\frac{p(m)m}{q(m)} = \sum_m q(m)\exp\Big[\ln\frac{p(m)m}{q(m)}\Big]$.

(b) Jensen's inequality says for a convex function ϕ, $E[\phi(Y)] \geq \phi(E[Y])$ for any random variable Y. Letting $\phi : z \mapsto \exp(z)$, and Y be the random variable that takes the value $\ln\frac{p(m)m}{q(m)}$ with probability $q(m)$, we see that

$$E[\phi(Y)] = \sum_m q(m)\exp\Big[\ln\frac{p(m)m}{q(m)}\Big] = E[X],$$

and

$$\phi(E[Y]) = \exp(\sum_m q(m)\ln\frac{p(m)m}{q(m)} = \exp\Big[-D(q\|p) + \sum_{m=1}^{M} q(m)\ln m\Big],$$

from which the desired result follows.

(c) $\sum_m q(m)\ln m = -\sum_m q(m)\ln\frac{q(m)}{mq(m)} = H(q) - \sum_m q(m)\ln\frac{1}{mq(m)}$. Letting Y to be the random variable that takes the value $1/(mq(m))$ with probability $q(m)$, and $\phi : z \mapsto -\ln z$, we see that

$$E[\phi(Y)] = -\sum_m q(m)\ln\frac{1}{mq(m)} \quad \text{and} \quad \phi(E[Y]) = -\ln\Big[\sum_{m=1}^{M} 1/m\Big].$$

Jensen's inequality now gives us the desired conclusion.

(d) From parts (b) and (c), $E[X] \geq \frac{1}{\sum_{m=1}^{M} 1/m}\exp\Big[-D(q\|p) + H(q)\Big]$. Using the fact that $\sum_{m=1}^{M} 1/m \leq 1 + \ln M$ gives us the required inequality.

(e) Setting $q(m) = \sqrt{p(m)}$ we have

$$H(q) - D(q\|p) = \sum_m q(m)\ln\frac{p(m)}{q(m)^2} = \sum_m q(m)\ln[1/C^2] = \ln[1/C^2],$$

and thus (d) yields to $E[X] \geq \frac{1}{1+\ln M}\frac{1}{C^2}$. But since $\sum_m q(m) = 1$, we see that $1/C = \sum_m \sqrt{p(m)}$, and the required inequality follows.

(f) Observe that with the given questioning strategy we ask m questions when $X = m$. Thus, the expected number of questions is simply $E[X]$. The result then follows from part (e).

(g) If $p(1) \geq p(2) \dots p(M)$, then

$$m = \sum_{i=1}^{m} 1 \leq \sum_{i=1}^{m} \sqrt{p(i)/p(m)} \leq \sum_{i} \sqrt{p(i)/p(m)}.$$

Thus,

$$E[X] = \sum_m p(m)m \leq \sum_m \sum_i \sqrt{p(m)}\sqrt{p(i)} = \Big[\sum_m \sqrt{p(m)}\Big]^2.$$

This result is relevant to 'guessing:' Observe that since the bound in (f) is independent of the ordering of the set $\mathcal{X}$, it says that any questioning strategy to guess X, as long as it can only ask questions of the form "is $X = x$" will ask $(1+\ln|\mathcal{X}|)^{-1}\big[\sum_x \sqrt{p(x)}\big]^2$ questions on the average. Part (g) then tells us that the bound is tight within the initial factor. If one now considers the case when the variable to be guessed is a random vector $\mathbf{X} = (X_1, \dots, X_n)$, with iid components then the bounds on the expected number of questions $E[G]$ become

$$\frac{1}{1 + n \ln M}\Big[\sum_x \sqrt{p(x)}\Big]^{2n} \leq E[G(\mathbf{X})] \leq \Big[\sum_x \sqrt{p(x)}\Big]^{2n},$$

and thus capture the exponential dependence of $E[G]$ on n.

Guessing strategies that are limited to asking questions of the form "is $X = x$" are encountered, for example, in the guessing of passwords.

Problem 2.

(a) We have (among many other possible solutions)

$$\begin{aligned} I(K, X; K, Y) &= H(K, X) - H(K, X|K, Y) \\ &= H(K) + H(X|K) - H(X|K, Y) \\ &= H(K) + I(X; Y|K), \end{aligned}$$

where the second line is obtained via the chain rule and by observing that $H(K, X|K, Y) = H(X|K, Y)$.

(b) Since $I(X; Y|K) = \sum_k \Pr(K = k) I(X; Y|K = k)$, and since $I(X; Y|K = k)$, being the mutual information over channel k, is less than C_k, it follows that

$$I(K, X; K, Y) \leq H(K) + E[C_K].$$

The equality will hold if $p(x|k) = p_k^*(x)$ where p_k^* is the distribution that maximizes the mutual information for channel k.

(c) Observe that

$$\frac{\partial}{\partial p_k} H(K) + E[C_K] = \ln \frac{1}{p_k} - 1 + C_k.$$

With $p_k = \exp(C_k) / \sum_i \exp(C_i)$, we see that the above derivative becomes

$$-1 + \ln \sum_i \exp(C_i)$$

which does not depend on k. Thus, this choice of p_k satisfies the Kuhn-Tucker conditions. As $H(K)$ is a concave function of $(p_1, \dots, p_M)$ and $E[C_K]$ a linear function of $(p_1, \dots, p_M)$, the sum $H(K) + E[C_K]$ is a concave function of $(p_1, \dots, p_M)$. Thus the Kuhn-Tucker conditions are necessary and sufficient for optimality.

(d) With p_k as above, we see that

$$\begin{aligned} H(K) + E[C_K] &= \sum_k p_k \ln \frac{1}{p_k} + \sum_k p_k C_k \\ &= \sum_k p_k \ln \frac{\exp(C_k)}{p_k} \\ &= \sum_k p_k \ln \Big[\sum_i \exp(C_i) \Big] \\ &= \ln \sum_i \exp(C_i). \end{aligned}$$

Problem 3.

(a) We have

$$\begin{aligned} E[-\log_2 q(X)] &= -\sum_x p(x) \log_2 q(x) \\ &= \sum_x p(x) \log_2 \frac{p(x)}{p(x) q(x)} \\ &= \sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \\ &= H(p) + D(p \| q). \end{aligned}$$

(b) When $q(x)$ is an integer power of $1/2$, the code which minimizes $\sum_x q(x)[\text{length}[C(x)]]$ will choose $\text{length}[C(x)] = -\log_2 q(x)$.

(c) From part (b) and (c) we see that

$$E[\text{length}[C(x)] - H(p) = H(p) + D(p \| q) - H(p) = D(p \| q).$$

(d) Observe that $-\frac{1}{n} \log_2 q(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (-\log_2 q(X_i))$. Since X_i are i.i.d., so are $-\log_2 q(X_i)$. Thus, the weak law of large numbers tells us that for any $\delta > 0$

$$\lim_{n \to \infty} \Pr \Big\{ \Big| -\frac{1}{n} \log_2 q(X_1, \dots, X_n) - E[-\log_2 q(X_1)) \Big| > \delta \Big\} = 0.$$

But from (a), $E[-\log_2 q(X_1)] = H(p) + H(p \| q)$, and the desired result follows.

(e) We see from part (d) that as n gets large $-\frac{1}{n}\log_2 q(X_1, \ldots, X_n)$ gets close to $H(p) + D(p\|q)$ with probability 1. Under the assumption that $H(p) + D(p\|q) \notin [H(q) - \epsilon, H(q) + \epsilon]$, we then conclude that as n gets large $(X_1, \ldots, X_n)$ will not be in $A(n, \epsilon)$ with probability 1, and thus the source output will not be assigned a codeword with probability 1.

(f) By the same reasoning as in part (e), under the assumption that $H(p) + D(p\|q) \in [H(q) - \epsilon, H(q) + \epsilon]$, as n gets large, the source output will be assigned a codeword with probability 1. Now, since

$$|A(n, \epsilon)| \geq (1 - \epsilon)2^{n[H(q) - \epsilon])},$$

we see that as n gets large

$$\frac{1}{n}\lceil \log_2 |A(n, \epsilon)| \rceil \geq H(q) - \epsilon + O(1/n).$$

Under the assumption, $H(q) \geq H(p)D(p\|q) - \epsilon$, and thus for large n, the codeword length per source letter exceeds

$$H(p) + D(p\|q) - 2\epsilon.$$