Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 3 of October 5, 2016

http://www.isi.ee.ethz.ch/teaching/courses/it1/

Problem 1 *Csiszár's Identity*

By the chain rule for mutual information, we have

$$I\big(A_{i+1}^n; B^i\big) = I\big(A_{i+1}^n; B^{i-1}\big) + I\big(A_{i+1}^n; B_i \big| B^{i-1}\big),$$
$$I\big(B^{i-1}; A_i^n\big) = I\big(B^{i-1}; A_{i+1}^n\big) + I\big(B^{i-1}; A_i \big| A_{i+1}^n\big).$$

Applying these identities, we obtain

$$\begin{aligned}
&\sum_{i=1}^n \Big(I\big(A_{i+1}^n; B_i \big| B^{i-1}\big) - I\big(B^{i-1}; A_i \big| A_{i+1}^n\big)\Big) \\
&\quad = \sum_{i=1}^n \Big[\Big(I\big(A_{i+1}^n; B^i\big) - I\big(A_{i+1}^n; B^{i-1}\big)\Big) - \Big(I\big(B^{i-1}; A_i^n\big) - I\big(B^{i-1}; A_{i+1}^n\big)\Big)\Big] \\
&\quad = \sum_{i=1}^n \Big(I\big(A_{i+1}^n; B^i\big) - I\big(B^{i-1}; A_i^n\big)\Big) \\
&\quad \overset{(i)}{=} I\big(A_{n+1}^n; B^n\big) - I\big(B^0; A^n\big) \\
&\quad = 0 - 0 \\
&\quad = 0,
\end{aligned}$$

where (i) follows because we have a *telescoping sum*, i.e., because $\sum_{i=1}^n (c_i - c_{i-1}) = c_n - c_0$.

Problem 2 *Entropy is Submodular*

For $\mathcal{S}, \mathcal{T} \in 2^\Omega$, define the chance variable X as the tuple of all chance variables from $\mathcal{S} \setminus \mathcal{T}$; Y as the tuple of all chance variables from $\mathcal{S} \cap \mathcal{T}$; and Z as the tuple of all chance variables from $\mathcal{T} \setminus \mathcal{S}$. The proof is accomplished by the following chain of (in)equalities:

$$\begin{aligned}
H(\mathcal{S} \cup \mathcal{T}) + H(\mathcal{S} \cap \mathcal{T}) &\overset{(i)}{=} H(X, Y, Z) + H(Y) \\
&\overset{(ii)}{=} H(X, Y) + H(Z|X, Y) + H(Y) \\
&\overset{(iii)}{\leq} H(X, Y) + H(Z|Y) + H(Y) \\
&\overset{(iv)}{=} H(X, Y) + H(Y, Z) \\
&\overset{(v)}{=} H(\mathcal{S}) + H(\mathcal{T}),
\end{aligned}$$

where (i) and (v) follow because we partitioned the set $\mathcal{S} \cup \mathcal{T}$ into the sets $\mathcal{S} \setminus \mathcal{T}$, $\mathcal{S} \cap \mathcal{T}$, and $\mathcal{T} \setminus \mathcal{S}$ and because grouping chance variables together does not change the entropy; (ii) and (iv) follow from the chain rule; and (iii) follows because conditioning does not increase entropy.

Problem 3 ***Pure Randomness and Bent Coins***

a) i) We make use of the chain rule:

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

Since in our case the random variables $X_1, \ldots, X_n$ are IID, we have

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) = nH_{\mathrm{b}}(p),$$

where the last equation follows from the fact that the X_i's are binary random variables with probability p.

ii) The random vector $\mathbf{W} \triangleq (Z_1, Z_2, \ldots, Z_K, K)$ is a function of the random vector $\mathbf{Y} \triangleq (X_1, X_2, \ldots, X_n)$, i.e., $\mathbf{W} = f(\mathbf{Y})$. However, we know that applying a function to a random vector never increases the uncertainty. To prove this, consider the following sequence of (in)equalities:

$$\begin{aligned} H(\mathbf{Y}, f(\mathbf{Y})) &= H(\mathbf{Y}) + \overbrace{H(f(\mathbf{Y})|\mathbf{Y})}^{=0} &&= H(\mathbf{Y}); \\ H(\mathbf{Y}, f(\mathbf{Y})) &= H(f(\mathbf{Y})) + \underbrace{H(\mathbf{Y}|f(\mathbf{Y}))}_{\geq 0} &&\geq H(f(\mathbf{Y})). \end{aligned}$$

Thus, $H(X_1, X_2, \ldots, X_n) \geq H(Z_1, Z_2, \ldots, Z_K, K)$.

iii) This follows from the *chain rule*.

iv) For a given K, $Z_1, \ldots, Z_K$ are pure random bits, i.e., $H(Z_i|K = k) = 1$ bit and all Z_i's are independent of each other. Therefore,

$$\begin{aligned} H(Z_1, Z_2, \ldots, Z_K|K) &= \sum_k P_K(k) \cdot H(Z_1, Z_2, \ldots, Z_K|K = k) \\ &= \sum_k P_K(k) \sum_{i=1}^{k} H(Z_i|K = k) \\ &= \sum_k P_K(k) \sum_{i=1}^{k} 1 \text{ bit} \\ &= \sum_k P_K(k)\, k \\ &= \mathsf{E}[K] \text{ bits.} \end{aligned}$$

v) This follows from the nonnegativity of entropy.

b) Since we do not know p, the only way to generate pure random bits is to use the fact that all sequences with the same number of ones are equally likely. For example, the sequences 0001, 0010, 0100 and 1000 are equally likely and can be used to generate two pure random bits. The sequences 0011, 0101, 1001, 0110, 1010, and 1100 are also equally likely, however one cannot produce $\log_2 6$ pure random bits as six is not a power of 2. Here, we can choose four out of the six sequences and produce two pure random bits, and then take the remaining

two sequences for producing one random bit. An example of a mapping to generate random bits is

$$\begin{array}{llll} 0000 \to \Lambda & & & \\ 0001 \to 00 & 0010 \to 01 & 0100 \to 10 & 1000 \to 11 \\ 0011 \to 00 & 0110 \to 01 & 1100 \to 10 & 1001 \to 11 \\ 1010 \to 0 & 0101 \to 1 & & \\ 1110 \to 11 & 1101 \to 10 & 1011 \to 01 & 0111 \to 00 \\ 1111 \to \Lambda & & & \end{array}$$

The resulting expected number of produced random bits is

$$\begin{aligned} \mathsf{E}[K] &= (1-p)^4 \cdot 0 + 4p(1-p)^3 \cdot 2 + 4p^2(1-p)^2 \cdot 2 + 2p^2(1-p)^2 \cdot 1 \\ &\quad + 4p^3(1-p) \cdot 2 + p^4 \cdot 0 \\ &= 8p(1-p)^3 + 10p^2(1-p)^2 + 8p^3(1-p). \end{aligned}$$

For example, for $p \approx \frac{1}{2}$, the expected number of pure random bits is close to 1.625. This is substantially less than the 4 pure random bits that could be generated if p were exactly $\frac{1}{2}$.

Problem 4 *Conditional vs. Unconditional Mutual Information*

a) We want to find a situation where the information "about X", which you "get from Y", is greater than the information gained from Y about X when we already know Z. A simple example of such a situation is if X is the result of a fair coin flip, and $Y = X$ and $Z = Y$. In this case,

$$I(X;Y) = H(X) - H(X|Y) = H(X) = 1 \text{ bit}$$

and

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 0 \text{ bits},$$

so that $I(X;Y|Z) < I(X;Y)$.

More generally we can say that if X and Y are *dependent* and either $Z = X$ or $Z = Y$, then

$$I(X;Y) = H(X) - \underbrace{H(X|Y)}_{<H(X)} > 0.$$

Also,

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 0 < I(X;Y).$$

The last equation follows from the fact that either the two addends are both zero (if $Z = X$), or that $H(X|Z) = H(X|Y,Z)$ (if $Z = Y$).

b) In this example Z is supposed to increase the information you get from Y about X. Let X and Y be two independent binary random variables with $P_X(x{=}0) = P_X(x{=}1) = P_Y(y{=}0) = P_Y(y{=}1) = \frac{1}{2}$. And let $Z = X + Y$. Then

$$I(X;Y) = H(X) - H(X|Y) = H(X) - H(X) = 0 \text{ bits}$$

and

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = \frac{1}{2} \text{ bits},$$

so that $I(X;Y|Z) > I(X;Y)$. More generally, the inequality is still valid if X and Y are any *independent* random variables and $X = g(Y,Z)$ is a function that allows to uniquely determine X given Z and Y but not so if only Z is provided. In this case, $H(X|Z) > 0$ and $H(X|Y,Z) = 0$.

Problem 5 ***Classes of Codes***

a) No, the code is not instantaneous, since the first codeword, 0, is a prefix of the second codeword, 01.

b) Yes, the code is uniquely decodable. Given a sequence of codewords, first isolate occurrences of 01 (i.e., find all the 1's) and then parse the rest into 0's.

c) Yes, all uniquely decodable codes are nonsingular.