ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

**Handout 15** Information Theory and Coding

Homework 7 November 1, 2011

PROBLEM 1.

(a) We have $\rho(X_1^\infty) = 0$. We show this by showing that $\rho(X_1^\infty) \leq \delta$ for any $\delta > 0$. To see the last statement, build an invertible FSM which "recognizes" a string of type "ab...ab" for a particular even length, call it $L$, and outputs lets say "0" at the end of this string and returns to the starting state. Hence this machine will output an infinite string of "0" when the input is $X_1^\infty$. From each state (including the starting state) of the chain which recognizes the special string make an edge back to the starting state in the case the next input is not the correct one. The output for each such edge is $1 + \lceil \log L \rceil$ bits long, the first bit is 1 to indicate that it is not the special path and on the next $\lceil \log L \rceil$ bits we give the index of the state (in binary representation) from which the return edge is drawn. This machine is clearly lossless and has a compressibility of $1/L$ for the desired sequence.

(b) A machine as described above will have $\rho_M(X_1^\infty) = 1/4$. In fact, one cannot do better than this. Consider a cycle, when from a given state we get back to the same state. During such a cycle we have to output at least one symbol, because the machine has to be information lossless. In an $L$ state machine we eventually create such a cycle within at most $L$ steps. This means that we output at least one symbol for every $L$ input symbols, so $\rho_M(X_1^\infty) \geq 1/L$.

(b) We have $\rho_{\text{LZ}} = 0$ since compressibility is non-negative and we know that the compressibility of LZ is at least as good as that of any FSM, i.e., we know that $\rho_{\text{LZ}}(X_1^\infty) \leq \rho(X_1^\infty)$.

(c) The dictionary increases by 1 every time and has size 2 in the beginning. Hence, if we look at lets say $c$ steps of the algorithm then we need in total

$$\sum_{i=1}^{c} \lceil \log(1+i) \rceil \leq c \log(2(c+1))$$

bits to describe the output.

What are the words which we are using. Note that the parsing is *a*, *b*, *ab*, *aba*, *ba*, *bab*,... Note that in average at most every second step the length of the used dictionary word increases by 1, i.e., we have a linear increase in the used dictionary words. Therefore, if we compute the total length which we have parsed after $c$ steps, this length increases like the squre of $c$.

It follows that the ratio of the total number of bits used divided by the total length described behaves like $1/c$, i.e., it tends to 0.

PROBLEM 2. (a) By the chain rule

$$I(U,T;V) = I(U;V) + I(T;V|U) = I(U;V),$$

since $I(T;V|U) = 0$ from the Markov property. Also,

$$I(U,T;V) = I(T;V) + I(U;V|T) \geq I(U;V|T),$$

from the non-negativity of the mutual information. These together imply that $I(U;V) \geq I(U;V|T)$.

(b)

$$I(X;Y|W) = \Pr\{W=1\}I(X;Y|W=1) + \Pr\{W=2\}I(X;Y|W=2)$$

Conditional on $W = k$, the distribution of $(X,Y)$ is $p_k(x)p(y|x)$, thus

$$I(X;Y|W=1) = \lambda I_1 + (1-\lambda)I_2.$$

(c) We obtain $p(x)$ by summing $p(w,x,y)$ over $y$ and $w$. This gives

$$p(x) = \lambda p_1(x) + (1-\lambda)p_2(x).$$

(d) Note that

$$p(w,x,y) = p(w)p(x|w)p(y|x),$$

that is $Y$ is independent of $W$ when $X$ is given. Thus from (a)

$$I(X;Y) \geq I(X;Y|W). \tag{1}$$

Letting $f(p_X)$ denote the value of $I(X;Y)$ as a function of the distribution of $X$ we can rewrite (1) as

$$f(\lambda p_1 + (1-\lambda)p_2) \geq \lambda f(p_1) + (1-\lambda)f(p_2),$$

which says that $f$ is concave.

Problem 3.

(a) By Bayes rule, for any events $A$ and $B$,

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}.$$

In this case, we wish to calculate the conditional probability of $a_1$ given the channel output. Thus we take the event $A$ to the event that the source produced $a_1$, and $B$ to be the event corresponding to one of the 8 possible output sequences. Thus $\Pr(A) = 1/2$, and $\Pr(B|A) = \epsilon^i(1-\epsilon)^{3-i}$, where $i$ is the number of ones in the received sequence. $\Pr(B)$ can then be calculated as $\Pr(B) = \Pr(a_1)\Pr(B|a_1) + \Pr(a_2)\Pr(B|a_2)$. Thus we can calculate

$$\begin{aligned}
\Pr(a_1|000) &= \frac{\frac{1}{2}(1-\epsilon)^3}{\frac{1}{2}(1-\epsilon)^3 + \frac{1}{2}\epsilon^3} \\
\Pr(a_1|100) = \Pr(a_1|010) = \Pr(a_1|001) &= \frac{\frac{1}{2}(1-\epsilon)^2\epsilon}{\frac{1}{2}(1-\epsilon)^2\epsilon + \frac{1}{2}\epsilon^2(1-\epsilon)} \\
\Pr(a_1|110) = \Pr(a_1|011) = \Pr(a_1|101) &= \frac{\frac{1}{2}(1-\epsilon)\epsilon^2}{\frac{1}{2}(1-\epsilon)\epsilon^2 + \frac{1}{2}\epsilon(1-\epsilon)^2} \\
\Pr(a_1|111) &= \frac{\frac{1}{2}\epsilon^3}{\frac{1}{2}\epsilon^3 + \frac{1}{2}(1-\epsilon)^3}
\end{aligned}$$

(b) If $\epsilon < 1/2$, then the probability of $a_1$ given 000,001,010 or 100 is greater than 1/2, and the probability of $a_2$ given 110,011,101 or 111 is greater than 1/2. Therefore, the decoding rule above chooses the source symbol that has maximum probability given the observed output. This is the *maximum a posteriori* decoding rule, and is optimal in that it minimizes the probability of error. To see that this is true, let the input source symbol be $X$, let the output of the channel be denoted by $Y$ and the decoded symbol be $\hat{X}(Y)$. Then

$$\begin{aligned}
\Pr(E) &= \Pr(X \neq \hat{X}) \\
&= \sum_y \Pr(Y = y) \Pr(X \neq \hat{X} | Y = y) \\
&= \sum_y \Pr(Y = y) \sum_{x \neq \hat{x}(y)} \Pr(x | Y = y) \\
&= \sum_y \Pr(Y = y) \left(1 - \Pr(\hat{x}(y) | Y = y)\right) \\
&= \sum_y \Pr(Y = y) - \sum_y Pr(Y = y) \Pr(\hat{x}(y) | Y = y) \\
&= 1 - \sum_y Pr(Y = y) \Pr(\hat{x}(y) | Y = y)
\end{aligned}$$

and thus to minimize the probability of error, we have to maximize the second term, which is maximized by choosing $\hat{x}(y)$ to the the symbol that maximizes the conditional probability of the source symbol given the output.

(c) The probability of error can also be expanded

$$\begin{aligned}
\Pr(E) &= \Pr(X \neq \hat{X}) \\
&= \sum_x \Pr(x) \Pr(\hat{X} \neq x) \\
&= \Pr(a_1) \Pr(Y = 011, 110, 101, \text{ or } 111) \\
&\quad + \Pr(a_2) \Pr(Y = 000, 001, 010 \text{ or } 100) \\
&= \frac{1}{2}\left(3\epsilon^2(1-\epsilon) + \epsilon^3\right) + \frac{1}{2}\left(3\epsilon^2(1-\epsilon) + \epsilon^3\right) \\
&= 3\epsilon^2(1-\epsilon) + \epsilon^3.
\end{aligned}$$

(d) By extending the same arguments, it is easy to see that the decoding rule that minimizes the probability of error is the maximum a posteriori decoding rule, which in this case is the same as the maximum likelihood decoding rule (since the two input symbols are equally likely). So we choose the source symbol that is most likely to have produced the given output. This corresponds to choosing $a_1$ if the number of 1's in the received sequence is $n$ or less, and choosing $a_2$ otherwise. The probability of error is then equal to (by symmetry) the probability of error given that $a_1$ was sent, which is the probability that $n+1$ or more 0's have been changed to 1's by the channel. This probability is

$$\Pr(E) = \sum_{i=n+1}^{2n+1} \binom{2n+1}{i} \epsilon^i (1-\epsilon)^{2n+1-i}$$

This probability goes to 0 as $n \to \infty$, since this is the probability that the number of 1's is $n+1$ or more, and since the expected proportion of 1's is $n\epsilon < n+1$, by the weak law of large numbers the above probability goes to 0 as $n \to \infty$.

Problem 4.

(a) The statistician calculates $\tilde{Y} = g(Y)$. Since $X \to Y \to \tilde{Y}$ forms a Markov chain, we can apply the data processing inequality. Hence for every distribution on $X$,

$$I(X;Y) \geq I(X;\tilde{Y}).$$

Let $\tilde{p}(x)$ be the distribution on $x$ that maximizes $I(X;\tilde{Y})$. Then

$$C = \max_{p(x)} I(X;Y) \geq I(X;Y)_{p(x)=\tilde{p}(x)} \geq I(X;\tilde{Y})_{p(x)=\tilde{p}(x)} = \max_{p(x)} I(X;\tilde{Y}) = \tilde{C}.$$

Thus, the statistician is wrong and processing the output does not increase capacity.

(b) We have equality (no decrease in capacity) in the above sequence of inequalities only if we have equality in data processing inequality, i.e., for the distribution that maximizes $I(X;\tilde{Y})$, we have $X \to \tilde{Y} \to Y$ forming a Markov chain, in other words if given $\tilde{Y}$, $X$ and $Y$ are independent.

Problem 5.

First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $x = \Pr(X=1)$:

$$\begin{aligned} H(Y|X) &= x h_2(\varepsilon) \\ H(Y) &= h_2(\Pr(Y=1)) = h_2((1-\varepsilon)x) \\ I(X;Y) &= H(Y) - H(Y|X) = h_2((1-\varepsilon)x) - x h_2(\varepsilon) \end{aligned} \tag{2}$$

We deduce that if $\varepsilon = 0$, the capacity equals 1 bit/symbol and is attained for $x = 1/2$. If $\varepsilon = 1$, then $I(X;Y) = 0$ for every $0 \leq x \leq 1$. Hence, the capacity is equal to zero and any value of $x$ achieves it. From now on we assume $\varepsilon \neq 0, 1$.

Using elementary calculus, we have that

$$\frac{d}{dx} I(X;Y) = (1-\varepsilon)\log\left(\frac{1-(1-\varepsilon)x}{(1-\varepsilon)x}\right) - h_2(\varepsilon)\ .$$

Imposing the condition $\frac{d}{dx} I(X;Y) = 0$ yields to the unique solution

$$x^*(\varepsilon) = \left((1-\varepsilon)(2^{\frac{h_2(\varepsilon)}{1-\varepsilon}} + 1)\right)^{-1}\ .$$

From (2) we have $I(X;Y) = 0$ for $x = 0$ and $x = 1$, and therefore the maximum of the mutual information is achieved for $x = x^*(\varepsilon)$. The capacity $C(\varepsilon)$ is given by

$$C(\varepsilon) = h_2((1-\varepsilon)x^*(\varepsilon)) - x^*(\varepsilon)h_2(\varepsilon) \text{ bits/symbol.}$$