ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 19 Information Theory and Coding
Solutions to Homework 7 Nov. 9, 2015

Problem 1.

(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

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

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

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

(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 2.

(a) Observe that with P_3 defined as in the problem, whatever distribution we choose for X, the random variables X, Y, Z form a Markov chain, i.e., given Y, the random variables X and Z are independent. The data processing theorem then yields:

$$
\begin{aligned}
I(X;Z) \leq I(X;Y) \leq C_1 \\
I(X;Z) \leq I(Y;Z) \leq C_2
\end{aligned}
$$

and thus $I(X;Z) \leq \min\{C_1, C_2\}$ for any distribution on X. We then conclude that $C_3 = \max_{p_X} I(X;Z) \leq \min\{C_1, C_2\}$.

(b) The statistician calculates $\tilde{Y} = g(Y)$.

(b1) 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.

(b2) 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 3. Observe that $H(Y) - H(Y|X) = I(X;Y) = I(X;Z) = H(Z) - H(Z|X)$.

(a) Consider a channel with binary input alphabet $\mathcal{X} = \{0,1\}$ with X uniformly distributed over $\mathcal{X}$, output alphabet $\mathcal{Y} = \{0,1,2,3\}$, and probability law

$$P_{Y|X}(y|x) = \begin{cases} \frac{1}{2}, & \text{if } x = 0 \text{ and } y = 0 \\ \frac{1}{2}, & \text{if } x = 0 \text{ and } y = 1 \\ \frac{1}{2}, & \text{if } x = 1 \text{ and } y = 2 \\ \frac{1}{2}, & \text{if } x = 1 \text{ and } y = 3 \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to verify $H(Y|X) = 1$. Since Y takes any value in $\mathcal{Y}$ with equal probability, its entropy is $H(Y) = 2$. Therefore $I(X;Y) = 1$. Define the processor output to be in alphabet $\mathcal{Z}$ and construct a deterministic processor $g : y \mapsto z = g(y)$ such that,

$$\begin{aligned} g: \quad & \mathcal{Y} \to \mathcal{Z} = \{0,1\} \\ & 0 \mapsto 0 \\ & 1 \mapsto 0 \\ & 2 \mapsto 1 \\ & 3 \mapsto 1. \end{aligned}$$

Clearly, $H(Z|X) = 0$ and $H(Z) = 1$. Therefore $I(X;Z) = 1$. We conclude that $I(X;Z) = I(X;Y)$ and $H(Z) < H(Y)$.

(b) Consider an error-free channel with binary input alphabet $\mathcal{X} = \{0,1\}$ with X uniformly distributed over $\mathcal{X}$, binary output alphabet $\mathcal{Y} = \{0,1\}$, and probability law

$$P_{Y|X}(y|x) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases}$$

Choose now $\mathcal{Z} = \{0,1,2,3\}$ an construct a probabilistic processor G such that

$$\begin{aligned} G: \quad & \mathcal{Y} \to \mathcal{Z} \\ & 0 \mapsto 0 \;\text{ with probability } \frac{1}{2} \;\text{ or }\; 1 \;\text{ with probability } \frac{1}{2} \\ & 1 \mapsto 2 \;\text{ with probability } \frac{1}{2} \;\text{ or }\; 3 \;\text{ with probability } \frac{1}{2}. \end{aligned}$$

Clearly, $I(X;Y) = 1 = I(X;Z)$ and $H(Y) = 1 < 2 = H(Z)$.

Problem 4. Since given X, one can determine Y from Z and vice versa, $H(Y|X) = H(Z|X) = H(Z) = \log 3$, regardless of the distribution of X. Hence the capacity of the channel is

$$\begin{aligned} C &= \max_{p_X} I(X;Y) \\ &= \max_{p_X} H(Y) - H(Y|X) \\ &= \log 11 - \log 3 \end{aligned}$$

which is attained when X has uniform distribution. The same result can also be seen by observing that this channel is symmetric.

Problem 5.

(a) Since the channel is symmetric, the input distribution that maximizes the mutual information is the uniform one. Therefore, $C = 1 + \epsilon \log_2(\epsilon) + (1-\epsilon)\log_2(\epsilon)$ which is equal to 0 when $\epsilon = \frac{1}{2}$.

(b) We have

- $I(X^n;Y^n) = I(X_2^n;Y^{n-1}) + I(X_2^n;Y_n|Y^{n-1}) + I(X_1;Y^n|X_2^n)$.
- $X_2^n = Y^{n-1}$ and $Y_1,\ldots,Y_n$ are i.i.d. and uniform in $\{0,1\}$, so $I(X_2^n;Y^{n-1}) = H(Y^{n-1}) = n-1$.
- Y_n is independent of (X_2^n, Y^{n-1}), so $I(X_2^n;Y_n|Y^{n-1}) = 0$.
- X_1 is independent of (Y^n, X_2^n), so $I(X_1;Y^n|X_2^n) = 0$.

Therefore, $I(X^n;Y^n) = n-1$.

(c) W is independent of Y^n, so $I(W;Y^n) = 0 = nC$.

Problem 6.

(a) Chain rule for mutual information.

(b) $I(W,Y^{i-1};Y_i) = I(Y^{i-1};Y_i) + I(W;Y_i|Y^{i-1}) \geq I(W;Y_i|Y^{i-1})$.

(c) $I(W,X_i,X^{i-1},Y^{i-1};Y_i) = I(W,Y^{i-1};Y_i) + I(X_i,X^{i-1};Y_i|W,Y^{i-1}) \geq I(W,Y^{i-1};Y_i)$. Note that this inequality is in fact equality, unless the mapping f_i is randomized.

(d) $W \to (X_i, X^{i-1}, Y^{i-1}) \to Y_i$ is a Markov chain. This follows from the following facts:

- For all $1 \leq j \leq i$, X_j is a function of (W, Y^{j-1}).
- For all $1 \leq j \leq i$, Y_j depends on (W, X^j, Y^{j-1}) only through X_j since the channel is memoryless.

This means that the joint probability distribution of (W, X^i, Y^i) can be written as follows:

$$\begin{aligned} P_{W,X^i,Y^i}(w,x^i,y^i) &= P_W(w) \times P_{X_1|W}(x_1|w)P_{Y_1|X_1}(y_1|x_1) \\ &\quad \times P_{X_2|W,Y_1}(x_2|w,y_1)P_{Y_2|X_2}(y_2|x_2) \times \ldots \times P_{X_i|W,Y^{i-1}}(x_i|w,x^{i-1})P_{Y_i|X_i}(y_i|x_i), \end{aligned}$$

which can be rewritten as

$$P_{W,X^i,Y^i}(w,x^i,y^i) = P_W(w)P_{X_i,X^{i-1},Y^{i-1}|W}(x_i,x^{i-1},y^{i-1}|w)P_{Y_i|X_i}(y_i|x_i).$$

(e) Since the channel is stationary and memoryless, $(X^{i-1}, Y^{i-1}) \rightarrow X_i \rightarrow Y_i$ is a Markov chain.

(f) From the definition of the capacity.

This proof still works even when the mappings f_i are randomized. We conclude that feedback does not increase the capacity even if we are allowed to use a randomized encoder.