ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

**Handout 18** Information Theory and Coding
Solutions to Homework 7 Nov. 11, 2019

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

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

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

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

(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. The assertion is clearly true with $n = 1$. To complete the proof by induction we need to show that the cascade of a BSC with parameter $q = \frac{1}{2}(1 - (1-2p)^n)$ with a BSC with parameter $p$ is equivalent to a BSC with parameter $\frac{1}{2}(1 - (1-2p)^{n+1})$. To do so, observe that for a cascade of a BSC with parameter $q$ and a BSC with parameter $p$, when a bit is sent, the opposite bit will be received if exactly one of the channels makes a flip, and this happens with probability $(1-q)p + (1-p)q$. Thus, the cascade is equivalent to a BSC with this parameter. For $q = \frac{1}{2}(1 - (1-2p)^n)$,

$$(1-q)p + (1-p)q = \frac{1}{2}(1 + (1-2p)^n)p + \frac{1}{2}(1 - (1-2p)^n)(1-p) = \frac{1}{2}(1 - (1-2p)^{n+1}),$$

and the assertion is proved.

Alternate proof: the cascade makes flips the incoming bit if an odd number of the elements of the cascade flip. Thus the cascade is equivalent to a BSC with parameter

$$a = \sum_{k:k \text{ odd}} \binom{n}{k} p^k (1-p)^{n-k}.$$

Let $b = \sum_{k:k \text{ even}} \binom{n}{k} p^k (1-p)^{n-k}$. Observe that

$$a + b = \sum_k \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1,$$

and

$$-a + b = \sum_k \binom{n}{k} (-p)^k (1-p)^{n-k} = (-p + 1 - p)^n = (1-2p)^n.$$

Subtracting the two equalities and dividing by two, we get $a = \frac{1}{2}(1 + (1-2p)^n)$.

Problem 3.

(a) Suppose $p_X(0) = p$, $p_X(1) = 1 - p = \bar{p}$ and $H(\mathbf{p}) = -\sum p_i \log(p_i)$. Then,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(\epsilon, (1-\alpha-\epsilon)p + \bar{p}\alpha, (1-\alpha-\epsilon)\bar{p} + p\alpha) - H(\alpha, \epsilon, 1-\alpha-\epsilon) \\ &= h_2(\epsilon) + (1-\epsilon) h_2\left(\frac{(1-\alpha-\epsilon)p + \bar{p}\alpha}{1-\epsilon}\right) - H(\alpha, \epsilon, 1-\alpha-\epsilon) \end{aligned}$$

Since $h_2\left(\frac{(1-\alpha-\epsilon)p+\bar{p}\alpha}{1-\epsilon}\right)$ is maximized when $p = \bar{p} = 1/2$, we obtain

$$\begin{aligned} C &= h_2(\epsilon) + (1-\epsilon) - H(\alpha, \epsilon, 1-\alpha-\epsilon) \\ &= h_2(\epsilon) + (1-\epsilon) - (h_2(\epsilon) + (1-\epsilon) h_2(\alpha/(1-\epsilon))) \\ &= (1-\epsilon)(1 - h_2(\alpha/(1-\epsilon))) \end{aligned}$$

(b) For $\alpha = 0$ and $\epsilon \neq 0$, we have a binary erasure channel and $C = (1-\epsilon)$. For $\alpha \neq 0$ and $\epsilon = 0$, we have a binary symmetric channel and $C = 1 - h_2(\alpha)$. When $\alpha + \epsilon = 1$, again we have a binary erasure channel with $C = 1 - \epsilon$.

(c)

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i|X_i) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i). \end{aligned}$$

where equality is achieved when $X_i$'s are independent. From (a), we know that $p = \bar{p} = 1/2$ is maximizes the mutual information for each channel use. Hence,

$$\max_{p(x_1^n)} I(X^n;Y^n) = \sum_{i=1}^n (1-\epsilon_i)(1 - h_2(\alpha_i/(1-\epsilon_i)))$$

Problem 4.

(a) Suppose on the contrary there is no $c \in \mathcal{X}$ such that $\Pr(E_c|X = c) \geq 1 - \epsilon$. This implies for all $c \in \mathcal{X}$, $\Pr(E_c|X = c) < 1 - \epsilon$. Taking the expectation over $p_x$, we have

$$\Pr(E_c) = \Pr\big(W(Y|X) \geq \gamma W(Y)\big) < 1 - \epsilon.$$

This also implies

$$\Pr\big(W(Y|X) < \gamma W(Y)\big) > \epsilon$$

which is a contradiction. Therefore we conclude that at least for one $c \in \mathcal{X}$, $\Pr(E_c|X = c) \geq 1 - \epsilon$.

(b) The termination condition is equivalent to the statement that

$$\Pr\big(E_c \setminus \cup_{m=0}^{M} D_m | X = c\big) \leq 1 - \epsilon, \ \forall c \in \mathcal{X}.$$

For removing the condition, we take the expectation of the above expression with respect to $p_x$, i.e.

$$\begin{aligned} &\sum_{c \in \mathcal{X}} \Pr\big(E_c \setminus \cup_{m=0}^{M} D_m | X = c\big) p_x(c) \\ &= \Pr\big(\{x, y : \ W(y|x) \geq \gamma W(y)\} \setminus \{x, y : \ y \in \cup_{m=1}^{M} D_m\}\big) \\ &< 1 - \epsilon \end{aligned}$$

where the last inequality follows from the fact that each term in the summation is smaller than $1 - \epsilon$.

To show the second part, we use the following inequalities. For any sets $A$ and $B = \cup_{m=1}^{M} B_m$, we have

$$\begin{aligned} \Pr\big(A|B\big) &= \Pr(A) - \Pr(A \cap B) \geq \Pr(A) - \Pr(B) \\ &\geq \Pr(A) - \sum_{m=1}^{M} \Pr(B_m). \end{aligned}$$

Substituting $A = \{x, y : \ W(y|x) \geq \gamma W(y)\}$ and $B_m = D_m$ we obtain the result.

(c) Since $D_m \subseteq E_m$ for all $m$ in $\{1, \ldots, M\}$, we have

$$\Pr\big(y \in D_m\big) \leq \Pr\big(y \in E_m\big)$$

and for all $y \in E_m$, we know $W(y|c_m) \geq \gamma W(y)$. Hence we obtain

$$\Pr\big(y \in E_m\big) = \sum_{y \in E_m} W(y) \leq \frac{1}{\gamma} \sum_{y \in E_m} W(y|c_m) = \frac{1}{\gamma} \Pr\big(y \in E_m | X = c_m\big) \leq \frac{1}{\gamma}.$$

In part (b), we had

$$\Pr\big(\{x, y : \ W(y|x) \geq \gamma W(y)\}\big) - \sum_{m=1}^{M} \Pr\big(y \in D_m\big) < 1 - \epsilon.$$

Upper bounding the summation term will decrease the value of the left hand side, so we upper bound each summation term by $\frac{1}{\gamma}$ and obtain

$$\Pr\big(\{x, y : \ W(y|x) \geq \gamma W(y)\}\big) - \frac{M}{\gamma} < 1 - \epsilon.$$

Rearranging the terms, we get the result as

$$\epsilon < \frac{M}{\gamma} + \Pr\big(\{x, y : \ W(y|x) < \gamma W(y)\}\big).$$

(d) From the result of part (c), we know

$$\begin{aligned}\epsilon_n &< \frac{M}{\gamma} + \Pr\big(W(Y_1^n|X_1^n) < \gamma W(Y_1^n)\big) \\ &= \frac{M}{\gamma} + \Pr\big(i(X_1^n;Y_1^n) < \log\gamma\big).\end{aligned}$$

Choosing $\gamma = 2^{n\delta}M = 2^{n(\delta+R)}$ we obtain the intended result.

(e) Since $X_1, \ldots, X_n$ are i.i.d and the channel is memoryless, we have $i(X_1^n;Y_1^n) = \sum_{i=1}^n i(X_i;Y_i)$. Hence, the inequality found in part (d) becomes

$$\epsilon_n < 2^{-n\delta} + \Pr\left(\frac{1}{n}\sum_{i=1}^n i(X_i;Y_i) < R + \delta\right).$$

Assume $I(X;Y) = R + \delta + \epsilon'$ for some $\epsilon' > 0$. Note that $E[i(X_i;Y_i)] = I(X;Y)$ for all $i \in \{1, \ldots, n\}$. Then the inequality becomes

$$\epsilon_n < 2^{-n\delta} + \Pr\left(\frac{1}{n}\sum_{i=1}^n i(X_i;Y_i) < I(X;Y) - \epsilon'\right).$$

We see from the above inequality that $\epsilon_n$ can be made arbitrarily small by law of large numbers.

*Note:* In contrary with Shannon's proof, which shows the existence of codes with small average error probability among a random ensemble of codebooks, this construction is deterministic. Moreover, in Shannon's proof, recall that expurgation of the half of the codes was needed to obtain the result that there are codes which have small maximal error probabilities. This proof directly shows this fact without any additional arguments.