2003 Spring Midterm for Information Theory

1. Suppose sources Z_1 and Z_2 are independent to each other, and have the same distribution as Z with

$$\begin{cases} \Pr[Z = e_1] = 0.4; \\ \Pr[Z = e_2] = 0.3; \\ \Pr[Z = e_3] = 0.2; \\ \Pr[Z = e_4] = 0.1 \end{cases}$$

   (a) Design the Huffman code for Z. (Requirement: The codeword of event e_1 must be the single bit, 0.)

   (b) Applying the Huffman code in (a) to the two sources in sequence yields the codeword $U_1, U_2, \ldots, U_k$, where k ranges from 2 to 6, depending on the outcomes of Z_1 and Z_2. Is U_1 and U_2 independent? Justify your answer. (Hint: Examine the value of $\Pr[U_2 = 0|U_1 = u_1]$ for different u_1.)

   (c) Is the average per-letter codeword length equal to the per-letter source entropy

$$0.4 \log_2 \frac{1}{0.4} + 0.3 \log_2 \frac{1}{0.3} + 0.2 \log_2 \frac{1}{0.2} + 0.1 \log_2 \frac{1}{0.1} = 1.84644 \text{ bits/letter?}$$

   Justify your answer.

   (d) Now if we apply the Huffman code in (a) sequentially to the i.i.d. sequence $Z_1, Z_2, Z_3, \ldots$ with marginal distribution the same as Z, and yield the output $U_1, U_2, U_3, \ldots$, can $U_1, U_2, U_3, \ldots$ be further compressed?

   If your answer to this question is NO, prove the i.i.d. uniformality of $U_1, U_2, U_3, \ldots$.
   If your answer to this question is YES, then explain why the optimal Huffman code does not give an i.i.d. uniform output. (Hint: Achievability of per-letter average codeword length to per-letter source entropy.)

Solution:

(a)

$$\begin{cases} e_1 & : & 0 \\ e_2 & : & 10 \\ e_3 & : & 110 \\ e_4 & : & 111 \end{cases}$$

(b) It can be derived that:

$$\Pr[U_2 = 0|U_1 = 0] = \Pr[Z_2 = e_1|Z_1 = e_1] = \Pr[Z_2 = e_1] = 0.4,$$

and

$$\begin{aligned} \Pr[U_2 = 0|U_1 = 1] &= \Pr[Z_1 = e_2|Z_1 \neq e_1] = \frac{\Pr[Z_1 = e_2 \wedge Z_1 \neq e_1]}{\Pr[Z_1 \neq e_1]} \\ &= \frac{\Pr[Z_1 = e_2]}{\Pr[Z_1 \neq e_1]} = 0.5. \end{aligned}$$

Since $\Pr[U_2 = 0|U_1 = 0] \neq \Pr[U_2 = 0|U_1 = 1]$, U_1 and U_2 are surely dependent.

(c) No, since $0.4 \times 1 + 0.3 \times 2 + 0.2 \times 3 + 0.1 \times 3 = 1.9 \neq 1.84644$.

(d) The answer is "YES, $U_1, U_2, \ldots$ can be further compressed" which has already been justified by (c). This result does not violate what have been stated in the first lecture:

> The output of an *optimal* source encoder in the sense of minimizing the average per-letter codeword length, which asymptotically achieves the per-letter source entropy, should be asymptotically i.i.d. with uniform marginal distribution.
> In case the average per-letter codeword length of the optimal source code equals the per-letter source entropy, its output becomes exactly i.i.d. with equally probable marginal.

since the average per-letter codeword length of the applied code, although it is optimal for each single letter (but not for all), is strictly larger than the per-letter source entropy.

2. Let the relation between the channel input $\{X_n\}_{n\geq 1}$ and channel output $\{Y_n\}_{n\geq 1}$ be:

$$Y_n = (\boldsymbol{\alpha}_n \times X_n) \oplus N_n \text{ for each } n,$$

where $\boldsymbol{\alpha}_n$, X_n, Y_n and N_n all take values from $\{0, 1\}$, and "$\oplus$" represents XOR operation. Assume that the attenuation $\{\boldsymbol{\alpha}_n\}_{n=1}^{\infty}$, channel input $\{X_n\}_{n=1}^{\infty}$ and additive noise $\{N_n\}_{n=1}^{\infty}$ are independent. Also, $\{\boldsymbol{\alpha}_n\}_{n=1}^{\infty}$ and $\{N_n\}_{n=1}^{\infty}$ are i.i.d. random sequences with

$$\Pr[\boldsymbol{\alpha}_n = 1] = \Pr[\boldsymbol{\alpha}_n = 0] = \frac{1}{2} \quad \text{and} \quad \Pr[N_n = 1] = 1 - \Pr[N_n = 0] = \varepsilon \in (0, 1/2).$$

(a) Derive the channel transition probability matrix

$$\begin{bmatrix} P_{Y_j|X_j}(0|0) & P_{Y_j|X_j}(1|0) \\ P_{Y_j|X_j}(0|1) & P_{Y_j|X_j}(1|1) \end{bmatrix}.$$

(b) The channel is apparently a discrete memoryless channel. Determine its channel capacity C. (Hint: $I(x;Y) = C$ if $P_X(x) > 0$ and $I(x;Y) \leq C$ if $P_X(x) = 0$, where P_X is the channel input distribution that achieves the channel capacity.)

(c) Suppose that α^n is known, and consists of k 1's. Find the maximum $I(X^n;Y^n)$ for the same channel with known α^n. (Hint: For known α^n, $\{(X_j, Y_j)\}_{j=1}^n$ are independent. So $I(X^n;Y^n) \leq \sum_{j=1}^n I(X_j;Y_j)$. You can treat that "the capacity, namely maximum mutual information, for a binary symmetric channel is $\log(2) - H_b(\varepsilon)$ as a known fact," where $H_b(\cdot)$ is the binary entropy function.)

(d) Some researchers attempt to derive the capacity of the channel in (b) in terms of the following steps:

- Derive the maximum mutual information between channel input X^n and output Y^n for a given α^n (namely the solution in (c));
- Calculate the expectation value of the maximum mutual information obtained from the previous step according to the statistics of α^n.
- Then the capacity of the channel is equal to this "expectation value" divided by n.

Does this "capacity" $\bar{C}$ coincide with that in (b)?

Answer:

(a)

$$\begin{aligned} P_{Y_j|X_j}(y_j|x_j) &= \sum_{\alpha_j\in\{0,1\}} P_{\boldsymbol{\alpha}_j}(\alpha_j) P_{Y_j|X_j,\boldsymbol{\alpha}_j}(y_j|x_j,\alpha_j) \\ &= \sum_{\alpha_j\in\{0,1\}} P_{\boldsymbol{\alpha}_j}(\alpha_j)\Pr[N_j = y_j \oplus (\alpha_j \times x_j)] \\ &= \frac{1}{2}\Pr[N_j = y_j] + \frac{1}{2}\Pr[N_j = y_j \oplus x_j] \\ &= \begin{cases} \Pr[N_j = y_j], & \text{if } x_j = 0; \\ \dfrac{1}{2}, & \text{if } x_j = 1. \end{cases} \end{aligned}$$

Hence, the channel transition probability matrix for $P_{Y_j|X_j}$ is:

$$\begin{bmatrix} P_{Y_j|X_j}(0|0) & P_{Y_j|X_j}(1|0) \\ P_{Y_j|X_j}(0|1) & P_{Y_j|X_j}(1|1) \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ 1/2 & 1/2 \end{bmatrix}.$$

(b)

$$\begin{aligned} I(x=0;Y) &= \sum_{y=0}^{1} P_{Y|X}(y|0)\log\frac{P_{Y|X}(y|0)}{P_Y(y)} \\ &= (1-\varepsilon)\log\frac{(1-\varepsilon)}{P_Y(0)} + \varepsilon\log\frac{\varepsilon}{P_Y(1)} \\ &= -H_b(\varepsilon) - (1-\varepsilon)\log P_Y(0) - \varepsilon\log P_Y(1), \end{aligned}$$

and

$$\begin{aligned} I(x=1;Y) &= \sum_{y=0}^{1} P_{Y|X}(y|1)\log\frac{P_{Y|X}(y|1)}{P_Y(y)} \\ &= \frac{1}{2}\log\frac{1/2}{P_Y(0)} + \frac{1}{2}\log\frac{1/2}{P_Y(1)} \\ &= -\log(2) - \frac{1}{2}\log P_Y(0) - \frac{1}{2}\log P_Y(1), \end{aligned}$$

where $H_b(\cdot)$ is the binary entropy function. Now if the channel input that achieves the channel capacity satisfies either $P_X(0) = 1$ or $P_X(1) = 1$, then $C = 0$ because $I(X;Y) \leq H(X) = 0$. Hence, the only possible case that results in $C > 0$ is that the channel input that achieves the channel capacity satisfies $\min\{\Pr[X_j = 0], \Pr[X_j = 1]\} > 0$. As a consequence, $C = I(x = 0; Y) = I(x = 1; Y)$, which implies that

$$\log \frac{P_Y(0)}{P_Y(1)} = \frac{\log(2) - H_b(\varepsilon)}{(1/2) - \varepsilon},$$

which in turns implies that

$$P_Y(0) = \frac{\exp\left\{\frac{\log(2)-H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}}{1+\exp\left\{\frac{\log(2)-H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}} \quad \text{and} \quad P_Y(1) = \frac{1}{1+\exp\left\{\frac{\log(2)-H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}}.$$

Consequently,

$$\begin{aligned}
C &= -\log(2) - \frac{1}{2}\log P_Y(0) - \frac{1}{2}\log P_Y(1) \\
&= -\log(2) - \frac{1}{2}\log[P_Y(0)P_Y(1)] \\
&= -\log(2) - \frac{1}{2}\log\left[\frac{\exp\left\{\frac{\log(2)-H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}}{\left(1+\exp\left\{\frac{\log(2)-H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}\right)^2}\right] \\
&= -\log(2) - \frac{\log(2) - H_b(\varepsilon)}{1-2\varepsilon} + \log\left(1+\exp\left\{\frac{\log(2) - H_b(\varepsilon)}{(1/2)-\varepsilon}\right\}\right) \\
&= -2\log(2)\frac{(1-\varepsilon)}{(1-2\varepsilon)} + \frac{1}{1-2\varepsilon}H_b(\varepsilon) + \log\left(1+\exp\left\{\frac{\log(2) - H_b(\varepsilon)}{1/2-\varepsilon}\right\}\right).
\end{aligned}$$

(c) $I(X^n;Y^n) \leq \sum_{j=1}^n I(X_j;Y_j)$ with equality holds if $\{X_j\}_{j=1}^n$ are independent. When $\alpha_j = 0$, $\max_{X_j} I(X_j;Y_j) = 0$ because X_j and Y_j are independent. When $\alpha_j = 1$, $\max_{X_j} I(X_j;Y_j) = \log(2) - H_b(\varepsilon)$. Hence, take $\{X_j\}_{j=1}^n$ to be i.i.d. with equal probable marginal, we have:

$$\max_{X^n} I(X^n;Y^n) = k[\log(2) - H_b(\varepsilon)].$$

(d) For a given α^n,

$$\max_{X^n} I(X^n;Y^n) = [\log(2) - H_b(\varepsilon)](\alpha_1 + \alpha_2 + \cdots + \alpha_n).$$

Notably, it is always the i.i.d. $\{X_j\}_{j=1}^n$ with equal probable marginal that achieves this maximum mutual information. So

$$E_{\boldsymbol{\alpha}^n}\left[\max_{X^n} I(X^n;Y^n)\right] = \frac{n}{2}[\log(2) - H_b(\varepsilon)],$$

which immediately gives that

$$\bar{C} = \frac{1}{2}[\log(2) - H_b(\varepsilon)].$$

Hence, $\bar{C}$ is not equal to C in (b).

In fact, $\bar{C} \geq C$. One can view the capacity $\bar{C}$ as a system in which the transmitter and the receiver have equipped with a circuit that can accurately estimate $\boldsymbol{\alpha}^n$ before the transmission begins. Hence, the transmitter and the receiver can jointly use the best scheme corresponding to the accurately estimated $\boldsymbol{\alpha}^n$. Hence, $\bar{C}$ for which a perfect channel identifier is assumed should be no less than C for which only coding technique is employed.

3. **Statement 1:** From the proof of the converse to Shannon's channel coding theorem, the error probability shall be ultimately larger than $1 - (C/R)$, where R is the transmission rate above channel capacity C. This converse theorem is applicable to all discrete memoryless channels, including the binary symmetric channels. So if taking, say, $R = 4C$, Shannon said that the error rate will be ultimately larger than 0.75.

   Statement 2: For the binary symmetric channel with equal probable input from $\{0, 1\}$, a pure random guess at the receiver side straightforwardly gives **Bit-Error-Rate** $= 0.5$, no matter how large the transmission rate R is.

Is there a conflict between the above two statements? Justify your answer. (Hint: The definition of average probability of error in Shannon's channel coding theorem.)

Definition (average probability of error) The average probability of error for a $\mathscr{C}_n = (n, M)$ code with encoder $f(\cdot)$ and decoder $g(\cdot)$ transmitted over channel $Q_{Y^n|X^n}$ is defined as

$$P_e(\mathscr{C}_n) = \frac{1}{M}\sum_{i=1}^{M} \lambda_i,$$

where

$$\lambda_i \stackrel{\triangle}{=} \sum_{\left\{y^n \in \mathcal{Y}^n \,:\, g(y^n) \neq i\right\}} Q_{Y^n|X^n}(y^n|f(i)).$$

Answer: As an example of binary symmetric channels, what Shannon concerned in his theorem is the n-block error rate (or symbol error rate, where a symbol consists of n concatenated channel inputs from $\{0, 1\}^n$), not the bit error rate. Notably, the symbol error rate can be very high (> 0.75), while the bit error rate remains small (≤ 0.5). For example, a symbol decision based upon the received signal y^n may have only one bit error; however, it is still considered as an error decision from the symbol error rate standpoint. □

4. Suppose that blocklength $n = 2$ and code size $M = 2$. Assume each code bit is either 0 or 1.

   (a) What is the number of all possible codebook designs? (Note: This number includes those lousy code designs, such as $\{00, 00\}$.)

   (b) Suppose that one randomly draws one of these possible code designs according to uniform distribution, and applies the selected code to *Binary Symmetric Channel* with crossover probability ε. Then what is the expected error probability, if the decoder simply selects the codeword whose Hamming distance to the received vector is the smallest? (When both codewords have the same Hamming distance to the received vector, the decoder makes an equal-probable guess on the transmitted codeword.)

   (c) Explain why the error in (b) does not vanish as $\varepsilon \downarrow 0$.

   (d) Based on the explanation in (c), show that the random error of a uniformly-drawn binary (n, M) random code with fixed ultimate code rate (namely, $\log_2(M)/n \to R$ fixed) cannot approach zero faster than "exponentially" with respect to blocklength n in memoryless BSC. In other words, there exists $p > 0$ and a constant $A > 0$ such that $\mathrm{BER}_{\mathrm{random}}(n) \geq Ap^n$. (Hint: The error of random (n, M) code is lower bounded by the error of random $(n, 2)$ code for $M \geq 2$.)

Answer:

(a) 2^4.

(b) Let $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be the two randomly drawn codewords. Then the decoder error given that $\boldsymbol{c}_1$ is transmitted is:

$$\begin{aligned}\lambda_1 \;=\; & \sum_{\{y^2 \in \{0,1\}^2 \;:\; d_H(y^2, \boldsymbol{c}_2) < d_H(y^2, \boldsymbol{c}_1)\}} P_{Y^2|X^2}(y^2|\boldsymbol{c}_1) \\ & + \frac{1}{2} \sum_{\{y^2 \in \{0,1\}^2 \;:\; d_H(y^2, \boldsymbol{c}_2) = d_H(y^2, \boldsymbol{c}_1)\}} P_{Y^2|X^2}(y^2|\boldsymbol{c}_1),\end{aligned}$$

where $d_H(\cdot, \cdot)$ is the Hamming distance, and X^2 and Y^2 denote the channel input

and output for memoryless BSC. Therefore, the expected value of λ_1 is equal to:

$$
\begin{aligned}
E[\lambda_1] &= \frac{1}{2^4} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) < d_H(y^2,\boldsymbol{c}_1)\}} P_{Y^n|X^2}(y^2|\boldsymbol{c}_1) \\
&\quad + \frac{1}{2^5} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) = d_H(y^2,\boldsymbol{c}_1)\}} P_{Y^2|X^2}(y^2|\boldsymbol{c}_1) \\
&= \frac{1}{2^4} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) < d_H(y^2,\boldsymbol{c}_1)\}} \varepsilon^{d_H(y^2,\boldsymbol{c}_1)} (1-\varepsilon)^{2-d_H(y^2,\boldsymbol{c}_1)} \\
&\quad + \frac{1}{2^5} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) = d_H(y^2,\boldsymbol{c}_1)\}} \varepsilon^{d_H(y^2,\boldsymbol{c}_1)} (1-\varepsilon)^{2-d_H(y^2,\boldsymbol{c}_1)} \\
&= \frac{(1-\varepsilon)^2}{2^4} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) < d_H(y^2,\boldsymbol{c}_1)\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(y^2,\boldsymbol{c}_1)} \\
&\quad + \frac{(1-\varepsilon)^2}{2^5} \sum_{\boldsymbol{c}_1 \in \{0,1\}^2} \sum_{\boldsymbol{c}_2 \in \{0,1\}^2} \sum_{\{y^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) = d_H(y^2,\boldsymbol{c}_1)\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(y^2,\boldsymbol{c}_1)} \\
&= \frac{(1-\varepsilon)^2}{2^4} \sum_{(\boldsymbol{c}_1,y^2) \in \{0,1\}^2 \times \{0,1\}^2} \sum_{\{\boldsymbol{c}_2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) < d_H(y^2,\boldsymbol{c}_1)\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(y^2,\boldsymbol{c}_1)} \\
&\quad + \frac{(1-\varepsilon)^2}{2^5} \sum_{(\boldsymbol{c}_1,y^2) \in \{0,1\}^2 \times \{0,1\}^2} \sum_{\{\boldsymbol{c}_2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) = d_H(y^2,\boldsymbol{c}_1)\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(y^2,\boldsymbol{c}_1)} \\
&= \frac{(1-\varepsilon)^2}{2^4} \Bigg(\sum_{\{(\boldsymbol{c}_1,y^2) \in \{0,1\}^2 \times \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_1)=1\}} \sum_{\{\boldsymbol{c}_2^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2)<1\}} \left(\frac{\varepsilon}{1-\varepsilon}\right) \\
&\quad + \sum_{\{(\boldsymbol{c}_1,y^2) \in \{0,1\}^2 \times \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_1)=2\}} \sum_{\{\boldsymbol{c}_2^2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2)<2\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^2 \Bigg) \\
&\quad + \frac{(1-\varepsilon)^2}{2^5} \sum_{(\boldsymbol{c}_1,y^2) \in \{0,1\}^2 \times \{0,1\}^2} \sum_{\{\boldsymbol{c}_2 \in \{0,1\}^2 \,:\, d_H(y^2,\boldsymbol{c}_2) = d_H(y^2,\boldsymbol{c}_1)\}} \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(y^2,\boldsymbol{c}_1)} \\
&= \frac{(1-\varepsilon)^2}{2^4} \Bigg(\underbrace{2^2}_{\#\text{ of } \boldsymbol{c}_1} \times \underbrace{2}_{\#\text{ of } y^2 \text{ for given } \boldsymbol{c}_1} \times \underbrace{1}_{\#\text{ of } \boldsymbol{c}_2 \text{ for given } \boldsymbol{c}_1, y^2} \times \left(\frac{\varepsilon}{1-\varepsilon}\right) \\
&\quad + 2^2 \times 1 \times 3 \times \left(\frac{\varepsilon}{1-\varepsilon}\right)^2 \Bigg) + \frac{(1-\varepsilon)^2}{2^5} \Bigg(2^2 \times 1 \times 1 \times \left(\frac{\varepsilon}{1-\varepsilon}\right)^0 \\
&\quad + 2^2 \times 2 \times 2 \times \left(\frac{\varepsilon}{1-\varepsilon}\right)^1 + 2^2 \times 1 \times 1 \times \left(\frac{\varepsilon}{1-\varepsilon}\right)^2 \Bigg) \\
&= \frac{1+6\varepsilon}{8}.
\end{aligned}
$$

By symmetry, $E[\lambda_2]$, the expectation value of error given $\boldsymbol{c}_2$ is transmitted, is equal to $E[\lambda_1]$. This concludes that the expected error probability is $(1+6\varepsilon)/8$.

(c) There are 4 lousy code designs, which are $\{00,00\}, \{01,01\}, \{10,10\}$ and $\{11,11\}$. In these four cases, only random guess is possible which results in $1/2$ error probability. Since the probability of adopting these four lousy codes are $4/16 = 1/4$, their contribution to the random code error is

$$\frac{1}{2} \times \frac{1}{4} = \frac{1}{8} = \lim_{\varepsilon \downarrow 0} \frac{1+6\varepsilon}{8}.$$

(d) From (c), it is trivial that the decoding error of (n, M) code is lower-bounded by that of $(n, 2)$ code for $M \geq 2$. Hence,

$$\mathrm{BER}_{\mathrm{random}}(n, M) \geq \mathrm{BER}_{\mathrm{random}}(n, 2) \geq \frac{2^n}{2^{2n}} \frac{1}{2} = \frac{1}{2} \left(\frac{1}{2} \right)^n.$$

□

5. Prove that the binary divergence $D(p\|q) = p \log(p/q) + (1-p) \log[(1-p)/(1-q)]$ is upper bounded by $\dfrac{(p-q)^2}{q(1-q)}$ for $0 < p < 1$ and $0 < q < 1$.

(Hint: $\log(1+u) \leq u$ for $u > -1$ and $\dfrac{p}{q} = 1 + \dfrac{p-q}{q}$.)

Answer:

$$\begin{aligned}
D(p\|q) &= p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)} \\
&= p \log \left(1 + \frac{p-q}{q} \right) + (1-p) \log \left(1 + \frac{(1-p)-(1-q)}{(1-q)} \right) \\
&= p \log \left(1 + \frac{p-q}{q} \right) + (1-p) \log \left(1 + \frac{q-p}{1-q} \right) \\
&\leq p \frac{p-q}{q} + (1-p) \frac{q-p}{1-q} \\
&= \frac{(p-q)^2}{q(1-q)},
\end{aligned}$$

with equality holds if, and only if, $p = q$.

6. Give a channel with a vectored output for a scalar input as follows.

$$X \rightarrow \boxed{\text{Channel}} \rightarrow Y_1, Y_2$$

Suppose that $P_{Y_1,Y_2|X}(y_1, y_2|x) = P_{Y_1|X}(y_1|x) P_{Y_2|X}(y_2|x)$ for every y_1, y_2 and x.

(a) Show that $I(X;Y_1,Y_2) = \sum_{i=1}^{2} I(X;Y_i) - I(Y_1;Y_2)$. (Hint: $I(X;Y_1,Y_2) = H(Y_1,Y_2) - H(Y_1,Y_2|X)$ and $H(Y_1,Y_2|X) = H(Y_1|X) + H(Y_2|X)$)

(b) Prove that the channel capacity C_{two} of using two outputs (Y_1,Y_2) is less than $C_1 + C_2$, where C_j is the channel capacity of using one output Y_j and ignoring the other output.

(c) Further assume that $P_{Y_i|X}$ is Gaussian with mean x and variance σ_j^2. In fact, this channel can be expressed as $Y_1 = X + N_1$ and $Y_2 = X + N_2$, where (N_1, N_2) are independent Gaussian distributed with mean zero and covariance matrix $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. Using the fact that $h(Y_1,Y_2) \leq \frac{1}{2}\log(2\pi e)^2 |\boldsymbol{K}_{Y_1,Y_2}|$ with equality holds when (Y_1,Y_2) are joint Gaussian, where $\boldsymbol{K}_{Y_1,Y_2}$ is the covariance matrix of (Y_1,Y_2), derive $C_{\text{two}}(S)$ for the two-output channel under the power constraint $E[X^2] \leq S$. (Hint: $I(X;Y_1,Y_2) = h(Y_1,Y_2) - h(N_1,N_2) = h(Y_1,Y_2) - h(N_1) - h(N_2)$.)

Proof:

(a)

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) \\ &= (H(Y_1) + H(Y_2) - I(Y_1;Y_2)) - (H(Y_1|X) + H(Y_2|X)) \\ &= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2). \end{aligned}$$

(b)

$$\begin{aligned} C_{\text{two}} &= \max_X I(X;Y_1,Y_2) \\ &= \max_X [I(X;Y_1) + I(X;Y_2)] - I(Y_1;Y_2) \\ &\leq \max_X I(X;Y_1) + \max_X I(X;Y_2) - I(Y_1;Y_2) \\ &\leq C_1 + C_2. \end{aligned}$$

(c) By $\text{Var}[Y_j^2] = \text{Var}[X^2] + \sigma_j^2$ and $\text{Cov}[Y_1,Y_2] = E[(X - E[X] + N_1)(X - E[X] + N_2)] = \text{Var}[X^2]$,

$$\begin{aligned} |\boldsymbol{K}_{Y_1,Y_2}| &= \text{Var}[Y_1]\text{Var}[Y_2] - \text{Cov}[Y_1,Y_2] \\ &= (\text{Var}[X] + \sigma_1^2)(\text{Var}[X] + \sigma_2^2) - \text{Var}^2[X^2] \\ &= \text{Var}[X](\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2 \end{aligned}$$

$$
\begin{aligned}
C_{\text{two}}(S) &= \max_{\{X:E[X^2]\le S\}} I(X;Y_1,Y_2) \\
&= \max_{\{X:E[X^2]\le S\}} h(Y_1;Y_2) - h(N_1) - h(N_2) \\
&\le \max_{\{X:E[X^2]\le S\}} \frac{1}{2}\log(2\pi e)^2\left(\mathrm{Var}[X](\sigma_1^2+\sigma_2^2)+\sigma_1^2\sigma_2^2\right) \\
&\quad -\frac{1}{2}\log(2\pi e)\sigma_1^2 - \frac{1}{2}\log(2\pi e)\sigma_2^2 \\
&\le \frac{1}{2}\log(2\pi e)^2\left(S(\sigma_1^2+\sigma_2^2)+\sigma_1^2\sigma_2^2\right) \\
&\quad -\frac{1}{2}\log(2\pi e)\sigma_1^2 - \frac{1}{2}\log(2\pi e)\sigma_2^2 \\
&= \frac{1}{2}\log\left(1+S\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)\right),
\end{aligned}
$$

where the second inequality equates when $\mathrm{Var}[X] = S$, and the first inequality equates if X is zero-mean Gaussian with variance S.