CS258 Information Theory Homework

Zhou Litao 518030910407 F1803016

May 31, 2020

1 Introduction

Exercise 1 (Coin flips) A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

1. Find the entropy $H(X)$ in bits. The following expressions may be useful:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \quad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2} \tag{1}$$

2. A random variable X is drawn according to this distribution. Find an "efficient" sequence of yes–no questions of the form, "Is X contained in the set S?" Compare $H(X)$ to the expected number of questions required to determine X.

Solution.

1. The distribution of X is

$$P(X = k) = (\frac{1}{2})^k, k = 1, 2, \cdots$$

Hence the entropy in bits is

$$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^k \log_2 \left(\frac{1}{2}\right)^k \\ &= \sum_{n=1}^{\infty} k \left(\frac{1}{2}\right)^k = 2. \end{aligned} \tag{2}$$

2. Since the probability of X to be a number x will decrease as x grows, an "efficient" way of asking questions about X will be like 'Is X contained in the set S_i?', where $S_i = \{1, 2, \cdots, i\}$ for $i \in \mathbf{R}$.

   If the answer is yes, then the question is done by concluding X is the current i. Otherwise, the next question will be asked. Let Y denote the number required to determine X. Clearly, $Y = X$.

   Thus the expected number of questions is

$$E(Y) = \sum_{k=1}^{\infty} k(\frac{1}{2})^k = 2.$$

   The result is equal to $H(X)$.

□

Exercise 2 (Zero conditional entropy) Show that if $H(Y|X) = 0$, then Y is a function of X [i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$].

Proof. By condition we have

$$H(Y|X) = \sum_x p(x)H(Y|X = x) = 0$$

Note that $p(x) > 0$, thus for any x, we have $H(Y|X = x) = 0$.
It follows that when x is determined, the distribution of Y is a single value.
That is to say, Y is a function of X. □

Exercise 3 (Coin weighing) Suppose that one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance.

1. Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter.
2. (Difficult) What is the coin-weighing strategy for $k = 3$ weighings and 12 coins?

Solution.

1. For a single weighing, there are three cases of results, namely left, right or euqal. Therefore k weighings can be represented by at most 3^k situations.

   For n coins, there are at most $2n + 1$ different cases, namely, one of the coins is heavier or lighter, or there exist no counterfeit coins.

   Theoretically, if we can find an injection from $2n+1$ cases to 3^k test results, we can determine the counterfeit directly from the information given by k pieces of facts. Hence, a necessary condition for this problem will be

   $$2n + 1 \leq 3^k.$$

   It follows that $\frac{3^k-1}{2}$ is an upper bond of n.

2. In this part we should construct a mapping from $2 \times 12 + 1 = 25$ cases to 3^3 test results. A natural idea is to use the ternary number.

   Let $\{-1, 0, 1\}$ be the test result for every weighing, where -1 indicates the left side is heavier, and 1 indicating the heavier right side. Then the test result can be represented a ternary number.

   In order to maximize the information this ternary number represents, we should arrange the order of every weighing carefully. Here we propose several principles to follow.

   - The numbers of coins on both side of the weighing should be equal. Otherwise no extra information will be gained from this weighing.
   - All coins should be weighed at least once. Otherwise no information about the coin will be gained from three weighings.
   - The weighing arrangement for every coin should be unique. That is to say, if coin A is weighed three times, no other coins should be weighed also three times. Otherwise we can't disguise between these two coins, indicating that the information we gain from the three weighings is not sufficient.

   It follows naturally from the third rule to construct a ternary mapping from digits $\{-1, 0, 1\}$ to 12 coins in the following table, where -1 represents that the coin is weighed on the left side, 1 represents the coin should be weighed on the right side.

Coin	1	2	3	4	5	6	7	8	9	10	11	12
3^0	1	-1	0	1	-1	0	1	-1	0	1	-1	0
3^1	0	1	1	1	-1	-1	-1	0	0	0	1	1
3^2	0	0	0	0	1	1	1	1	1	1	1	1

Table 1: First Try in Building Weighing Strategy

In this way, a unique weighing strategy is generated for each coin. However, this schedule violates the first rule. In the second weighing, for example, there are two extra coins on the right. To address this problem, we can simply reverse some of the coins.

Coin	1	2	3	4	5	6	**7**	8	**9**	10	**11**	**12**
3^0	1	-1	0	1	-1	0	**-1**	-1	**0**	1	**1**	**0**
3^1	0	1	1	1	-1	-1	**1**	0	**0**	0	**-1**	**-1**
3^2	0	0	0	0	1	1	**-1**	1	**-1**	1	**-1**	**-1**

Table 2: Actual Weighing Strategy

Now we can build the weighing roadmap.

- For every weighing, put the -1-labelled coins on the left and 1-labelled to the right.
- Record the weighing result as -1 to be leftward-sloping, 0 to be balanced and 1 to be rightward-sloping.
- Combine the number and check the table 2. $(0, 0, 0)$ indicates that there are no counterfeits. If the vector matches the i-th column, then coin i is heavier. Or if the component-wise negation of the vector matches the i-th column, then coin i is lighter.

□

2 Entropy Measures

Exercise 4 Show that $D(p\|q) = 0$ if and only if $p(x) = q(x)$.

Proof.

$$\begin{aligned} -D(p(x)\|q(x)) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{q(x)}{p(x)} \\ &\le \log(\sum p(x)\frac{q(x)}{p(x)}) \qquad \text{(By Concavity of log(x))} \\ &= \log(\sum q(x)) \le \log 1 = 0 \end{aligned} \tag{3}$$

The first equality holds if and only if

$$\frac{q(x)}{p(x)} = k, \text{ for every } x \in \mathcal{X} \text{ such that } p(x) > 0$$

The second equality holds if and only if there exists no x such that $p(x) = 0$ while $q(x) > 0$. Since we have $\sum p(x) = 1$, we know that

$$\frac{q(x_1)}{p(x_1)} = \frac{q(x_2)}{p(x_2)} = \cdots = \frac{\sum q(x)}{\sum p(x)}.$$

By the second condition we know that $\sum q(x) = \sum p(x) = 1$. Hence $p(x) = q(x)$ for all $x \in \mathcal{X}$.

□

Exercise 5 Show that $I(X;Y) \ge 0$, with equality if and only if X and Y are independent.

Proof.

$$\begin{aligned} -I(X;Y) &= \sum_{(x,y)\in\mathcal{X}\star\mathcal{Y}} p(x,y)\log\frac{p(x)p(y)}{p(x,y)} \\ &\le \log\left(\sum_{(x,y)\in\mathcal{X}\star\mathcal{Y}} p(x,y)\frac{p(x)p(y)}{p(x,y)}\right) \\ &= \log\sum_{(x,y)\in\mathcal{X}\star\mathcal{Y}} p(x)p(y) = \log\left(\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y)\right) = \log 1 = 0 \end{aligned} \tag{4}$$

The equality holds if and only if

$$\frac{p(x,y)}{p(x)p(y)} = k, \text{ for every } (x,y)\in\mathcal{X}\star\mathcal{Y} \text{ such that } p(x,y)>0$$

Since $\sum p(x,y)=1$, we know that $k=1$. That is to say, $p(x,y)=p(x)p(y)$ for every possible x,y. X and Y are independent.

□

Exercise 6 Show that $D(p(y|x)\|q(y|x))\ge 0$ with equality if and only if $p(y|x)=q(y|x)$ for all x and y such that $p(x)>0$.

Proof.

$$\begin{aligned} -D(p(y|x)\|q(y|x)) &= \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{q(y|x)}{p(y|x)} \\ &\le \sum_{x\in\mathcal{X}} p(x)\log\sum_{y\in\mathcal{Y}} q(y|x) \\ &\le \sum_{x\in\mathcal{X}} p(x)\log 1 = 0 \end{aligned} \tag{5}$$

The first equality holds if and only if

$$\frac{q(y|x)}{p(y|x)} = k, \text{ for every } y\in\mathcal{Y} \text{ such that } p(y|x)>0 \text{ given } p(x)>0 \text{ with } x\in\mathcal{X}$$

The second equality holds if and only if given $p(x)>0$, there exists no y such that $p(y|x)=0$ while $q(y|x)>0$. Since we have $\sum p(y|x)=1$, we know that

$$\frac{q(y_1|x)}{p(y_1|x)} = \frac{q(y_2|x)}{p(y_2|x)} = \cdots = \frac{\sum q(y|x)}{\sum p(y|x)}.$$

It follows from the second condition that $\sum q(y|x) = \sum p(y|x) = 1$. Hence the equality holds if and only if $p(y|x)=q(y|x)$ for all $p(x)>0$.

□

Exercise 7 Show that $I(X;Y|Z)\ge 0$ with equality if and only if X and Y are conditionally independent given Z.

Proof.

$$\begin{aligned} -I(X;Y|Z) &= \sum_{(x,y,z)\in\mathcal{X}\star\mathcal{Y}\star\mathcal{Z}} p(x,y,z)\log\frac{p(x|z)p(y|z)}{p(x,y|z)} \\ &\leq \sum_{z\in\mathcal{Z}} p(z)\log \sum_{(x,y)\in\mathcal{X}\star\mathcal{Y}} p(x,y|z)\frac{p(x|z)p(y|z)}{p(x,y|z)} \\ &= \sum_{z\in\mathcal{Z}} p(z)\log\left(\sum_{x\in\mathcal{X}} p(x|z)\sum_{y\in\mathcal{Y}} p(y|z)\right) \\ &= \sum_{z\in\mathcal{Z}} p(z)\log 1 = 0 \end{aligned} \tag{6}$$

The equality holds if and only if

$$\frac{p(x,y|z)}{p(x|z)p(y|z)} = k, \text{ for every } (x,y)\in\mathcal{X}\star\mathcal{Y} \text{ such that } p(x,y)>0 \text{ given } p(z)>0 \text{ with } z\in\mathcal{Z}$$

Since $\sum p(x,y|z)=1$, we know that $k=1$. That is to say, $p(x,y|z)=p(x|z)p(y|z)$ for every possible x,y given $p(z)>0$. Therefore the equality holds if and only if X and Y are independent given Z.

□

Exercise 8 Let $u(x)=\frac{1}{|\mathcal{X}|}$ be the uniform probability mass fuction over X, and let $p(x)$ be the probability mass function for X, Then

$$0 \leq D(p\|u) = \log|\mathcal{X}| - H(X)$$

Proof. From Exercise 4 we know $D(p\|u)\geq 0$. By definition of mutual entropy we have

$$\begin{aligned} D(p\|u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} = \sum_{x\in\mathcal{X}} p(x)\log|\mathcal{X}|p(x) \\ &= \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) - \sum_{x\in\mathcal{X}} p(x)\log p(x) = \log|\mathcal{X}| - H(X) \end{aligned} \tag{7}$$

□

Exercise 9 (Conditioning reduces entropy) Show that

$$H(X|Y)\leq H(X)$$

with equality if and only if X and Y are independent.

Proof. We know that $I(X;Y)=H(X)-H(X|Y)$. From Exercise 5 we know that $I(X;Y)\geq 0$. It follows that $H(X|Y)\leq H(X)$ with equality if and only if X and Y are independent. □

3 Entropies

Exercise 10 (Run-length coding) Let $X_1, X_2, \ldots, X_n$ be (possibly dependent) binary random variables. Suppose that one calculates the run lengths $\mathbf{R}=(R_1,R_2,\ldots)$ of this sequence (in order as they occur). For example, the sequence $\mathbf{X}=0001100100$ yields run lengths $\mathbf{R}=(3,2,2,1,2)$. Compare $H(X_1,X_2,\ldots,X_n), H(\mathbf{R})$ and $H(X_n,\mathbf{R})$. Show all equalities and inequalities, and bound all the differences..

Solution. When $X_1, X_2, \ldots, X_n$ is determined, their running length is determined. $H(\mathbf{R}|X_1, X_2, \ldots, X_n) = 0$, which implie that

$$H(\mathbf{R}, X_1, X_2, \ldots, X_n) = H(X_1, X_2, \ldots, X_n)$$

When one element X_i is determined, given the running length, the whole sequence will be determined. That is to say $H(X_1, X_2, \ldots, X_n|X_i, \mathbf{R}) = 0$, which implies that

$$H(\mathbf{R}, X_1, X_2, \ldots, X_i \ldots, X_n, X_i) = H(\mathbf{R}, X_1, X_2, \ldots, X_n) = H(X_i, \mathbf{R})$$

Hence we have

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_i, \mathbf{R}) && (\star) \\ &= H(\mathbf{R}) + H(X_i|\mathbf{R}) \\ &\leq H(\mathbf{R}) + H(X_i) \\ &\leq H(\mathbf{R}) + \log 2 = H(\mathbf{R}) + 1 && (\star) \end{aligned} \tag{8}$$

On the other hand, since $H(X_i|R) \geq 0$, we have

$$H(X_1, X_2, \ldots, X_n) = H(X_i, \mathbf{R}) = H(\mathbf{R}) + H(X_i|\mathbf{R}) \geq H(\mathbf{R}) \quad (\star) \tag{9}$$

The starred lines make up all the equalities and inequalities required by the problem. □

Exercise 11 (Grouping rule for entropy) Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ be a probability distribution on m elements (i.e., $p_i \geq 0$ and $\sum_{i=1}^{m} p_i = 1$). Define a new distribution q on $m-1$ elements as $q_1 = p_1, q_2 = p_2 \cdots, q_{m-2} = p_{m-2}$, and $q_{m-1} = p_{m-1} + p_m$ [i.e., the distribution $\mathbf{q}$ is the same as $\mathbf{p}$ on $\{1, 2, \ldots, m-2\}$, and the probability of the last element in $\mathbf{q}$ is the sum of the last two probabilities of $\mathbf{p}$] Show that

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1} + p_m)\, H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right) \tag{10}$$

.

Proof. By unfolding the definiton of entropy we have

$$\begin{aligned} H(\mathbf{p}) &= -\sum_{i=1}^{m} p_i \log p_i = -\sum_{i=1}^{m-2} p_i \log p_i - p_{m-1} \log p_{m-1} - p_m \log p_m \\ &= -\sum_{i=1}^{m-2} q_i \log q_i - q_{m-1} \log q_{m-1} + q_{m-1} \log q_{m-1} - p_{m-1} \log p_{m-1} - p_m \log p_m \\ &= H(\mathbf{q}) + (p_{m-1} + p_m) \log(p_{m-1} + p_m) - p_{m-1} \log p_{m-1} - p_m \log p_m \\ &= H(\mathbf{q}) + (p_{m-1} + p_m) \left(-\frac{p_{m-1}}{p_{m-1} + p_m} \log \frac{p_{m-1}}{p_{m-1} + p_m} - \frac{p_m}{p_{m-1} + p_m} \log \frac{p_m}{p_{m-1} + p_m}\right) \\ &= H(\mathbf{q}) + (p_{m-1} + p_m)\, H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right) \end{aligned} \tag{11}$$

□

Exercise 12 (Fano) We are given the following joint distribution on (X, Y) :

X \ Y	a	b	c
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

Let $\hat{X}(Y)$ be an estimator for X (based on Y) and let $P_e = \Pr\{\hat{X}(Y) \neq X\}$

1. Find the minimum probability of error estimator $\hat{X}(Y)$ and the associated P_e
2. Evaluate Fano's inequality for this problem and compare.

Solution. 1. By observation, a feasible deterministic estimator for X can be defined as

$$\hat{X}(Y) = \begin{cases} 1 & Y = a \\ 2 & Y = b \\ 3 & Y = c \end{cases} \tag{12}$$

In this case, the error probability is

$$P_e = \sum_{(x,y)\in X\star Y, x\neq y} p(x,y) = 6 \times \frac{1}{12} = \frac{1}{2}$$

2. The general Fano's inequality implies that

$$P_e \geqslant \frac{H(X|Y) - 1}{\log |\mathcal{X}|} \tag{13}$$

We can calculate the conditional entropy

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y = y) \\ &= 3 \cdot \frac{1}{3} H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2} \end{aligned} \tag{14}$$

By substituting Equation 14 into Equation 13 we know

$$P_e \geq \frac{1.5 - 1}{\log_2 3} \approx 0.3155$$

If we assume $\hat{X} : y \to x$, then by the stronger Fano's inequality we have

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} \geq \frac{1.5 - 1}{\log_2 2} = 0.5$$

Hence, the estimator we have found is the best under condition that $\hat{X} : y \to x$. It may be improved by introducing randomness. However, the P_e will not be less than 0.3155.

□

Exercise 13 (Discrete entropies) Let X and Y be two independent integervalued random variables. Let X be uniformly distributed over $\{1, 2, \ldots, 8\}$, and let $\Pr\{Y = k\} = 2^{-k}, k = 1, 2, 3, \ldots$

1. Find $H(X)$.
2. Find $H(Y)$.
3. Find $H(X + Y, X - Y)$

Solution.

1. For uniform distribution of X, $H(X) = \log |\mathcal{X}| = \log 8 = 3$
2. By definiton $H(Y) = \sum_{k=1}^{\infty} 2^{-k} \log 2^k = \sum_{k=1}^{\infty} k 2^{-k} = 2$.

3. Since $(X, Y) \Leftrightarrow (X+Y, X-Y)$, we have $H(X+Y, X-Y|X, Y) = 0$ and $H(X, Y|X+Y, X-Y) = 0$. It follows that

$$\begin{aligned} H(X,Y) &= H(X+Y, X-Y|X,Y) + H(X,Y) \\ &= H(X+Y, X-Y, X, Y) \\ &= H(X+Y, X-Y) + H(X,Y|X+Y, X-Y) \\ &= H(X+Y, X-Y) \end{aligned} \tag{15}$$

Since X and Y are independent,

$$H(X+Y, X-Y) = H(X,Y) = H(X) + H(Y) = 3 + 2 = 5$$

□

4 Advanced Entropies

Exercise 14 Prove that under the constraint that $X \to Y \to Z$ forms a Markov Chain, $X \perp Y|Z$ and $X \perp Z$ imply $X \perp Y$.

Proof. From $X \perp Y|Z$, we have $I(X;Y|Z) = 0$. From $X \perp Z$, we have $I(X;Z) = 0$. It follows that

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && \text{(Unfold by definition of mutual information)} \\ &= H(X) - H(X|Z) + H(X|Z) - H(X|Y) \\ &= H(X) - H(X|Z) + H(X|Z) - H(X|Y,Z) && \text{(Markov Chain: } p(x|y) = p(x|y,z)\text{)} \\ &= I(X;Z) + I(X;Y|Z) = 0 && \text{(Fold by definition of mutual information)} \end{aligned} \tag{16}$$

,which implies that $X \perp Y$. □

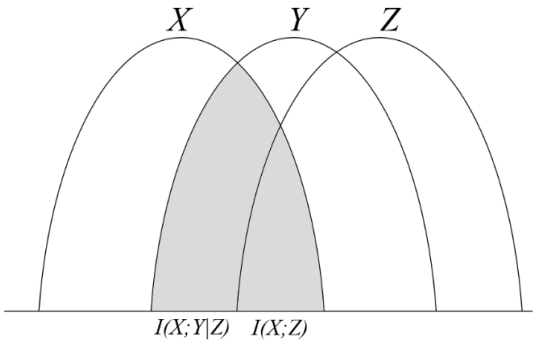


Figure 1: Venn Diagram of Exercise 1

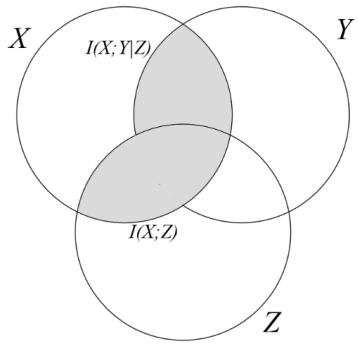


Figure 2: Venn Diagram of Exercise 2

Exercise 15 Prove that the implication in Exercise 14 continues to be valid without the Markov Chain constraint

Proof.

$$\begin{aligned} I(X;Y) &= I(X;Y|Z) + (I(X;Y) - I(X;Y|Z)) && \text{(Note } X \perp Y|Z \to I(X;Y|Z) = 0\text{)} \\ &= H(X) - H(X|Y) - (H(X|Z) - H(X|Y,Z)) && \text{(Fold by definition of mutual information)} \\ &= (H(X) - H(X|Z)) - (H(X|Y) - H(X|Y,Z)) && \text{(Unfold by definition of mutual information)} \\ &= I(X;Z) - I(X;Z|Y) && \text{(Note } X \perp Y \to I(X;Y) = 0\text{)} \\ &= -I(X;Z|Y) \leq 0 && \text{(Nonnegative conditional mutual information)} \end{aligned} \tag{17}$$

On the other hand, $I(X;Y) \geq 0$. Hence $I(X;Y)$ must be zero. That is to say, $X \perp Y$. □

Exercise 16 Prove that $Y \perp Z|T$ implies $Y \perp Z|(X,T)$ conditioning on $X \to Y \to Z \to T$.

Proof.

$$\begin{aligned}
I(Y;Z|X,T) &= H(Y|X,T) - H(Y|Z,X,T) && \text{(Unfold mutual information)} \\
&= H(X,Y,T) - H(X,T) - H(X,Y,Z,T) + H(X,Z,T) && \text{(Unfold conditional entropy)} \\
&= (H(X,Y,T) - H(X,Y,Z,T)) - (H(X,T) - H(T)) \\
&\quad +(H(X,Z,T) - H(Z,T)) - H(T) + H(Z,T) \\
&= -H(Z|X,Y,T) - H(X|T) + H(X|Z,T) + H(Z|T) && \text{(Fold conditional entropy)} \\
&= (H(Z|T) - H(Z|Y,T)) - (H(X|T) - H(X|Z,T)) && \text{(Markov Chain: } p(z|x,y,t) = p(z|y,t)) \\
&= I(Y;Z|T) - I(X;Z|T) && \text{(Note } Y \perp Z|T \to I(Y;Z|T) = 0) \\
&= -I(X;Z|T) \leq 0
\end{aligned} \tag{18}$$

On the other hand, $I(Y;Z|X,T) \geq 0$ can be proved by unfolding the definiton of conditional mutual information and the convexity property. Hence $I(Y;Z|X,T)$ must be zero. That is to say, $Y \perp Z|(X,T)$. □

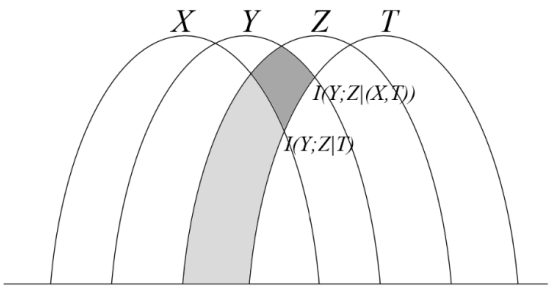


Figure 3: Venn Diagram of Exercise 3

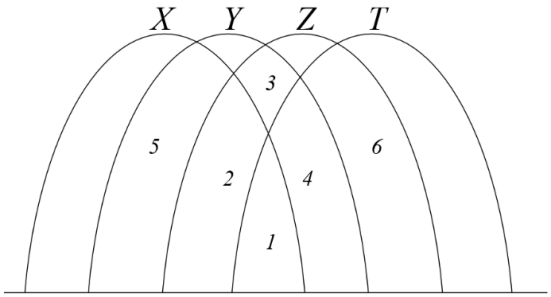


Figure 4: Venn Diagram of Exercise 4

Exercise 17 Let $X \to Y \to Z \to T$ form a Markov Chain. Determine which of the following always hold:

1. $I(X;T) + I(Y;Z) \geq I(X;Z) + I(Y;T)$
2. $I(X;T) + I(Y;Z) \geq I(X;Y) + I(Z;T)$
3. $I(X;Y) + I(Z;T) \geq I(X;Z) + I(Y;T)$

Solution. Inequality (1) and (3) always hold. We illustrate the answer through the venn diagram shown in Figure 4, where area $1 \sim 6$ is respectively represented by $I(X;T), I(X;Z|T), I(Y;Z|(X,T)), I(Y;T|X), I(X;Y|Z), I(Z;T|Y)$.

1. The inequality can be rewritten in form of areas as

$$1 + (1+2+3+4) \geq (1+2) + (1+4).$$

   Since $I(Y;Z|(X,T)) \geq 0$, the inequality holds.

2. The inequality can be rewritten in form of areas as

$$1 + (1+2+3+4) \geq (1+2+5) + (1+4+6).$$

   We can't determine the relation between $I(Y;Z|(X,T))$ (Area 3) and $I(X;Y|Z) + I(Z;T|Y)$ (Area 5 and 6) except that they are both nonnegative. The inequality will not always hold.

3. The inequality can be rewritten in form of areas as

$$(1+2+5)+(1+4+6) \geq (1+2)+(1+4).$$

Since $I(X;Y|Z)+I(Z;T|Y) \geq 0$ can be proved by the nonnegativity of conditional mutual information, the inequality holds. Furthermore, the conclusion can also be derived from the data-processing inequality of Markov Chain with $I(X;Y) \geq I(X;Z)$ and $I(Z;T) \geq I(Y;T)$

□

Exercise 18 (Drawing with and without replacement) An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a difficult way and a relatively simple way to do this.)

Solution. We use $X_i \in \{\text{red, white, black}\}$ to identify the result of the i-th drawing. No matter with replacement or without replacement, the distributions of a single arbitrary variable X_i are the same.

$$\begin{array}{cccc} X_i & red & white & black \\ p(x) & \dfrac{r}{r+w+b} & \dfrac{w}{r+w+b} & \dfrac{b}{r+w+b} \end{array} \tag{19}$$

With replacement, the previous result won't interfere with the present drawing. Hence we have

$$H\left(X_i|X_{i-1},\ldots,X_1\right) = H\left(X_i\right)$$

. It follows that

$$H(X_1,X_2,...,X_k) = \sum_{i=1}^{k} H\left(X_i|X_{i-1},\ldots,X_1\right) = \sum_{i=1}^{k} H\left(X_i\right) \quad \text{with replacement} \tag{20}$$

Without replacement, we only have

$$H(X_1,X_2,...,X_k) = \sum_{i=1}^{k} H\left(X_i|X_{i-1},\ldots,X_1\right) \quad \text{without replacement} \tag{21}$$

Note that in Equation 20 and Equation 21, all the single-variable entropies are of the same value. By condition-reduce-entropy theorem we know that

$$H\left(X_i|X_{i-1},\ldots,X_1\right) \leq H(X_i) \quad \text{for any } i$$

Since the equality holds if and only if X_i are mutually independent, which is not true in this problem, it follows that the entropy will be larger with replacement. □

Exercise 19 (Metric) A function $\rho(x,y)$ is a metric if for all x,y,

- $\rho(x,y) \geq 0$.
- $\rho(x,y) = \rho(y,x)$.
- $\rho(x,y) = 0$ if and only if $x = y$.
- $\rho(x,y)+\rho(y,z) \geq \rho(x,z)$.

1. Show that $\rho(X,Y) = H(X|Y)+H(Y|X)$ satisfies the first, second, and fourth properties above. If we say that $X = Y$ if there is a one-to-one function mapping from X to Y, the third property is also satisfied, and $\rho(X,Y)$ is a metric.

2. Verify that $\rho(X,Y)$ can also be expressed as

$$\begin{aligned} \rho(X,Y) &= H(X)+H(Y)-2I(X;Y) \\ &= H(X,Y)-I(X;Y) \\ &= 2H(X,Y)-H(X)-H(Y) \end{aligned} \tag{22}$$

Proof.

1. - Note that $H(X|Y) \geqslant 0, \quad H(Y(X) \geqslant 0 \Rightarrow \quad \rho(X,Y) \geqslant 0$
   - By unfolding the definition it's easy to see $H(X|Y) \neq H(Y|X) = H(Y|X) + H(X|Y)$
   - If $X = Y$, there exists a one-to-one mapping, i.e. $\rho(X,Y) = H(X|Y) + H(Y|X) = 0$
     On the other hand, if $\rho(X,Y) = H(X|Y) + H(Y|X) = 0$ Since $H(X|Y) \geqslant 0, \quad H(Y(X) \geqslant 0$, we have $H(X|Y) = 0, \quad H(Y(X) = 0$. By the conclusion in Exercise 2, Assignment 1, X and Y are mutually each other's function, i.e.there exists a one-to-one mapping between X and Y.
   - The conclusion can be derived using condition-reduce-entropy and nonnegativity of conditional entropy.

$$\begin{aligned} H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) &\geqslant H(X|Y,Z) + H(Y|X) + H(Y|Z) + H(Z|Y,X) \\ &= H(X,Y|Z) + H(Z,Y|X) \\ &= H(X|Z) + H(Y|X,Z) + H(Z|X) + H(Y|Z,X) \\ &\geqslant H(X|Z) + H(Z|X) = \rho(X,Z) \end{aligned} \tag{23}$$

2.

$$\begin{aligned} \rho(X,Y) &= H(X|Y) + H(Y|X) \\ &= H(X) - I(X;Y) + H(Y) - I(X;Y) \\ &= H(X) + H(Y) - 2I(X;Y) \qquad (\star) \\ &= (H(X) - I(X;Y) + H(Y)) - I(X;Y) \\ &= H(X,Y) - I(X;Y) \qquad (\star) \\ &= H(X,Y) - (H(X) + H(Y) - H(X,Y)) \\ &= 2H(X,Y) - H(X) - H(Y) \qquad (\star) \end{aligned} \tag{24}$$

The expressions required by the problem have been labeled with $(\star)$ in the derivation.

□

Exercise 20 (Entropy of a disjoint mixture) Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $X_1 = \{1, 2, \cdots, m\}$ and $X_2 = \{m+1, \cdots, n\}$. Let

$$X = \begin{cases} X_1 & \text{with probability } \alpha \\ X_2 & \text{with probability } 1 - \alpha \end{cases} \tag{25}$$

1. Find $H(X)$ in terms of $H(X_1), H(X_2)$ and α.
2. Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.

Solution.

1. We calculate $H(X)$ by unfolding the definition of entropy.

$$\begin{aligned} H(X) &= -\sum_{x \in X_1} \alpha p_1(x) \log \alpha p_1(x) - \sum_{x \in X_2} (1-\alpha) p_2(x) \log (1-\alpha) p_2(x) \\ &= -\alpha \log \alpha \sum_{x \in X_1} p_1(x) - (1-\alpha) \log(1-\alpha) \sum_{x \in X_2} p_2(x) + \alpha H(X_1) + (1-\alpha) H(X_2) \\ &= -\alpha \log \alpha - (1-\alpha) \log(1-\alpha) + \alpha H(X_1) + (1-\alpha) H(X_2) \end{aligned} \tag{26}$$

2. We consider $H(X)$ to be a function over α. Note that $g(\alpha) = -\alpha \log(\alpha)$ is a concave function, and some affine transformation over α and linear components won't interfere with the concavity. The function of $H(x)$ is a concave function.

We can get the maximal value by calculating the derivative of $H(X)$ over α.

$$\begin{aligned}\frac{dH(X)}{d\alpha} &= -\frac{1}{d\alpha}\left(\frac{\alpha\ln\alpha}{\ln 2}+\frac{(1-\alpha)\ln(1-\alpha)}{\ln 2}-\alpha H(X_1)-(1-\alpha)H(X_2)\right)\\ &= -\frac{1+\ln\alpha}{\ln 2}-\frac{-1-\ln(1-\alpha)}{\ln 2}+H(X_1)-H(X_2):=0\end{aligned} \tag{27}$$

The maximal value is obtained at the derivative to be 0.

$$\begin{aligned}-\ln\alpha+\ln(1-\alpha) &= \ln 2\,(H(X_2)-H(X_1))\\ \ln\frac{1-\alpha}{\alpha} &= \ln 2\,(H(X_2)-H(X_1))\\ \frac{1-\alpha}{\alpha} &= 2^{H(X_2)-H(X_1)}\\ \alpha &= \frac{2^{H(X_1)}}{2^{H(X_2)}+2^{H(X_1)}}\end{aligned} \tag{28}$$

The optimal solution is in the domain, so the maximal value can be obtained. By substituting the α value into $2^{H(X)}$ we can obtain its upper bond.

$$\begin{aligned}2^{H(X)} &= 2^{-\alpha\log\alpha-(1-\alpha)\log(1-\alpha)+\alpha H(X_1)+(1-\alpha)H(X_2)}\\ &= \alpha^{-\alpha}\cdot(1-\alpha)^{\alpha-1}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha}\\ &\le \left(\frac{2^{H(X_1)}}{2^{H(X_2)}+2^{H(X_1)}}\right)^{-\alpha}\cdot\left(\frac{2^{H(X_2)}}{2^{H(X_2)}+2^{H(X_1)}}\right)^{\alpha-1}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha}\\ &= \left(2^{H(X_1)}+2^{H(X_2)}\right)\cdot 2^{-\alpha H(X_1)}\cdot 2^{-(1-\alpha)H(X_2)}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha}\\ &= 2^{H(X_1)}+2^{H(X_2)}\end{aligned} \tag{29}$$

An interpretation of this conclusion is that $2^{H(X)}$ is the effective alphabet size of X, while $2^{H(X_1)}+2^{H(X_2)}$ is the sum sizes of the effective alphabets X_1, X_2. The alphabets of X_1 and X_2 do not overlap, with independent distribution, and they add up exactly to the alphabet of X.

If our probability of choice between X_1 and X_2 is in proportion to their effective alphabet size, as the third line in Equation 28 shows, the resulting X will have the effective alphabet size equivalent to the sum of X_1 and X_2.

Otherwise, the unbalanced weight of X_1 and X_2 will reduce the actual effective alphabet size in X, since one variable's excessive occurence will reduce the occurence of the other, so that the latter's effective alphabet size will be less than what it really is.

□

Exercise 21 (Entropy of a sum) Let X and Y be random variables that take on values $x_1, x_2, \cdots, x_r$ and $y_1, \cdots, y_s$, respectively. Let $Z = X + Y$.

- Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \le H(Z)$ and $H(X) \le H(Z)$. Thus, the addition of independent random variables adds uncertainty.
- Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.
- Under what conditions does $H(Z) = H(X) + H(Y)$?

Solution.

1. $Z = X + Y$ indicates that any of the two variable can determine the third variable. That is to say,

$$H(X|Y,Z) = H(Y|Z,X) = H(Z|X,Y) = 0$$

. By observing $I(Y;Z|X)$ we have

$$\begin{aligned} I(Y;Z|X) &= H(Y|X) - H(Y|X,Z) \\ &= H(Z|X) - H(Z|X,Y) \end{aligned} \tag{30}$$

,which implies that $H(Y|X) = H(Z|X)$.

If X and Y are independent, $H(X,Y) = H(X) + H(Y)$.

$$\begin{aligned} H(X,Y,Z) &= H(Z|X,Y) + H(X,Y) = H(X) + H(Y) \\ &= H(X|Y,Z) + H(Y,Z) = H(Z|X) + H(X) \\ &= H(Y|X,Z) + H(X,Z) = H(Z|Y) + H(Y) \end{aligned} \tag{31}$$

Equation 31 indicates that $H(Z|X) = H(Y)$ and that $H(Z|Y) = H(X)$. By condition-reduce-entropy theorem we have $H(Z) \geq H(Z|X)$ and $H(Z) \geq H(Z|Y)$. It follows that $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$.

2. An exampling distribution of X and Y can be

Prob		x	
		0	1
y	0	$\frac{1}{2}$	0
	-1	0	$\frac{1}{2}$

The entropy of X and Y are

$$H(X) = H(Y) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$$

The distribution of $Z = X + Y$ is $\Pr(Z = 0) = 1$, which results in the entropy $H(Z) = 0 < H(X) = H(Y)$.

3. From $Z = X + Y$ we know $H(Z) = H(Z) - H(Z|X,Y) = I(X,Y;Z)$.

$I(X,Y;Z) = H(X,Y) - H(X,Y|Z)$ indicates that $H(Z) \leq H(X,Y)$. The equality holds if and only if $H(X,Y|Z) = 0$.

Furthermore, $H(X,Y) = H(X)+H(Y)-I(X;Y)$, which implies that $H(X,Y) \leq H(X)+H(Y)$. The equality holds if and only if $I(X;Y) = 0$, i.e. X and Y are independent.

The second equality constraint and the propositions that $H(X|Y,Z) = H(Y|Z,X) = 0$ can ensure the first equality constraint. Therefore, under the condition that X and Y are independent will $H(Z) = H(X)+H(Y)$ hold.

□

Exercise 22 (Data processing) Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; that is, let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_{n-1})$$

Reduce $I(X_1; X_2, \cdots, X_n)$ to its simplest form.

Solution. By the chain rule of mutual information we have

$$I(X_1; X_2, \ldots, X_n) = \sum_{i=2}^{n} I(X_i; X_1 | X_{i-1}, X_{i-2}, \ldots, X_2)$$

Note that for $i > 2$, we have

$$\begin{aligned} I\left(X_i; X_1 | X_{i-1}, X_{i-2}, \ldots, X_2\right) &= H\left(X_i | X_{i-1}, X_{i-2}, \ldots, X_2\right) - H\left(X_i | X_{i-1}, X_{i-2}, \ldots, X_2, X_1\right) \\ &= H(X_i|X_{i-1}) - H(X_i|X_{i-1}) = 0 \quad \text{(Markov Chain)} \end{aligned} \tag{32}$$

It follows that $I\left(X_1; X_2, \ldots, X_n\right) = I\left(X_1; X_2\right)$.

□

Exercise 23 (Infinite entropy) This problem shows that the entropy of a discrete random variable can be infinite. Let $A = \sum_{n=2}^{\infty} (n \log^2 n)^{-1}$. [It is easy to show that A is finite by bounding the infinite sum by the integral of $(x \log^2 x)^{-1}$.] Show that the integer-valued random variable X defined by $\Pr(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \cdots$, has $H(X) = +\infty$.

Proof. By definition of entropy we can calculate that

$$\begin{aligned} H(X) &= -\sum_{n=2}^{\infty} p(n) \log p(n) \\ &= \sum_{n=2}^{\infty} (An \log^2 n)^{-1} \log(An \log^2 n) \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + \log^2 n}{An \log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{\log^2 n}{An \log^2 n} \end{aligned} \tag{33}$$

As has been indicated by the condition, the first component is finite. The last component will be nonnegative with sufficiently large n. We show that the second component is infinite. Note that

$$0 < \sum_{n=2}^{\infty} \frac{1}{An \log n} < \int_2^{\infty} \frac{\ln 2 dx}{Ax \ln x} = \int_2^{\infty} \frac{\ln 2 d(\ln x)}{A \ln x} = \frac{\ln 2}{A} \ln(\ln x)\Big|_2^{\infty} \to \infty$$

It follows that $H(X) = +\infty$ □

5 Entropy Rate

Exercise 24 (Monotonicity of entropy per element) For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

$$\begin{aligned} \frac{H\left(X_1, X_2, \ldots, X_n\right)}{n} &\leq \frac{H\left(X_1, X_2, \ldots, X_{n-1}\right)}{n-1} \\ \frac{H\left(X_1, X_2, \ldots, X_n\right)}{n} &\geq H\left(X_n | X_{n-1}, \ldots, X_1\right) \end{aligned}$$

Proof.

We first prove the second statement by properties of stationary process.

$$\begin{aligned} H\left(X_1, \cdots, X_n\right) &= \sum_{i=1}^{n} H\left(X_i | X_{i-1}, \cdots X_2, X_1\right) \\ &\geqslant \sum_{i=1}^{n} H\left(X_i | X_{i-1}, X_{i-2}, \cdots, X_{n-i+1}\right) \\ &= \sum_{i=1}^{n} H\left(X_n | X_{n-1}, \cdots, X_2, X_1\right) \\ &= nH\left(X_n | X_1, X_2, \cdots, X_{n-1}\right) \end{aligned} \tag{34}$$

By Equation 34 and entropy equalities we can prove the first statement.

$$\begin{aligned}(n-1)H\left(X_1,\cdots,X_n\right) &\leqslant nH\left(X_1,\cdots,X_n\right)-H\left(X_1,\cdots,X_n\right)\\ &\leqslant n\left[H\left(X_1,\cdots,X_n\right)-H\left(X_n|X_1,\cdots,X_{n-1}\right)\right]\\ &= nH\left(X_1,\cdots,X_{n-1}\right)\end{aligned} \tag{35}$$

□

Exercise 25 (Initial conditions) Show, for a Markov chain, that

$$H\left(X_0|X_n\right)\geq H\left(X_0|X_{n-1}\right)$$

Thus, initial conditions X_0 become more difficult to recover as the future X_n unfolds.

Proof. For Markov Chain, we have $p(x_0|x_n,x_{n-1})=p(x_0|x_{n-1})$. Hence

$$\begin{aligned}H(X_0|X_n,X_{n-1}) &= -\sum_{x_n,x_{n-1}}p(x_n,x_{n-1})\sum_{x_0}p(x_0|x_n,x_{n-1})\log p(x_0|x_n,x_{n-1})\\ &= -\sum_{x_{n-1}}\left(\sum_{x_n}p(x_n,x_{n-1})\right)\sum_{x_0}p(x_0|x_{n-1})\log p(x_0|x_{n-1})\\ &= -\sum_{x_{n-1}}p(x_{n-1})\sum_{x_0}p(x_0|x_{n-1})\log p(x_0|x_{n-1})=H(X_0|X_{n-1})\end{aligned} \tag{36}$$

Since condition reduce entropy, we have $H(X_0|X_{n-1})=H(X_0|X_n,X_{n-1})\leq H(X_0|X_n)$ □

Exercise 26 (The past has little to say about the future) For a stationary stochastic process $X_1,X_2,\ldots,X_n,\ldots$, show that

$$\lim_{n\to\infty}\frac{1}{2n}I\left(X_1,X_2,\ldots,X_n;X_{n+1},X_{n+2},\ldots,X_{2n}\right)=0$$

Proof. First Note that

$$I\left(X_1,X_2,\ldots,X_n;X_{n+1},X_{n+2},\ldots,X_{2n}\right)=H(X_1,\ldots,X_n)-H(X_{n+1},\ldots,X_{2n}|X_1,\ldots,X_n) \tag{37}$$

By definition of entropy rate we know that

$$\frac{1}{2}\sum_{i=1}^{n}H(X_1,\ldots,X_n)\to H(\mathcal{X})$$

□

Exercise 27 (Entropy rate) Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathcal{X})$. Show that

$$\frac{1}{n}H\left(X_n,\ldots,X_1|X_0,X_{-1},\ldots,X_{-k}\right)\to H(\mathcal{X})$$

for $k=1,2,\ldots$

Proof.

$$H\left(X_n,\ldots,X_1|X_0,X_{-1},\ldots,X_{-k}\right)=\sum_{i=1}^{n}H(X_i|X_{i-1},\ldots,X_{-k}) \tag{38}$$

Note that

$$H'(X)=\lim_{n\to\infty}H(X_n|X_{n-1},ldots,X_1)=\lim_{n\to\infty}H(X_n|X_{n-1},\ldots,X_{-k})$$

We already know that $H'(X) \to H(\mathcal{X})$.

By Cesaro Mean,

$$\frac{1}{n} H\left(X_n, \ldots, X_1 | X_0, X_{-1}, \ldots, X_{-k}\right) \to H(\mathcal{X})$$

□

Exercise 28 (Markov's inequality and Chebyshev's inequality)

1. (Markov's inequality) For any nonnegative random variable X and any $t > 0$, show that

$$\Pr\{X \geq t\} \leq \frac{EX}{t} \tag{39}$$

   Exhibit a random variable that achieves this inequality with equality.

2. (Chebyshev's inequality) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y - \mu)^2$, show that for any $\epsilon > 0$

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \tag{40}$$

3. (Weak law of large numbers) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ be the sample mean. Show that

$$\Pr\left\{\left|\bar{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2} \tag{41}$$

   Thus, $\Pr\left\{\left|\bar{Z}_n - \mu\right| > \epsilon\right\} \to 0$ as $n \to \infty$. This is known as the weak law of large numbers.

Proof.

1. Note that for any t,

$$t \cdot 1_{\{X \geq t\}} \leq X \tag{42}$$

   By taking expectation at both sides we have

$$E(1_{\{X \geq t\}}) = \Pr(X \geq t) \leq \frac{EX}{t} \tag{43}$$

2. Note $EX = DY = \sigma^2$. By letting $X = (Y - \mu)^2$ and $t = \epsilon^2$ in the Markov inequality, we can derive that

$$\Pr\left\{\left|\bar{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2} \tag{44}$$

3. Note $E\bar{Z}_n = \mu$ and $D\bar{Z}_n = \frac{1}{n^2}\sum_{i=1}^n DZ_i = \frac{\sigma^2}{n}$. By applying Chebyshev's inequality on $\bar{Z}_n$ we have

$$\Pr\left\{\left|\bar{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2} \tag{45}$$

□

Exercise 29 (Piece of cake) A cake is sliced roughly in half, the largest piece being chosen each time, the other pieces discarded. We will assume that a random cut creates pieces of proportions

$$P = \begin{cases} \left(\frac{2}{3}, \frac{1}{3}\right) & \text{with probability } \frac{3}{4} \\ \left(\frac{2}{5}, \frac{3}{5}\right) & \text{with probability } \frac{1}{4} \end{cases}$$

Thus, for example, the first cut (and choice of largest piece) may result in a piece of size $\frac{3}{5}$. Cutting and choosing from this piece might reduce it to size $\left(\frac{3}{5}\right)\left(\frac{2}{3}\right)$ at time 2, and so on. How large, to first order in the exponent, is the piece of cake after n cuts?

Solution. Let $C_1, C_2, \ldots, C_n$ denote the choice of each cut. Then after n cuts, the size of the cake $W_n = \prod_{i=1}^{n} C_i$. By taking the logarithm at the equation, we have $\log W_n = \sum_{i=1}^{n} \log C_i$. Since C_is are i.i.d., we can apply the law of large numbers as follows.

$$\lim_{n\to\infty} \frac{1}{n} \log W_n = E(\log C) = \frac{3}{4} \log \frac{2}{3} + \frac{1}{4} \log \frac{3}{5} \approx -0.494 \tag{46}$$

Note that the equation above indicates that

$$\frac{1}{n} \log W_n = -0.494 + o(1) W_n \quad = 2^{-0.494t + o(n)} \tag{47}$$

So the first order in the exponent is $\frac{3}{4} \log \frac{2}{3} + \frac{1}{4} \log \frac{3}{5} \approx -0.494$. □

Exercise 30 (AEP) Let X_i be iid $\sim p(x), x \in \{1, 2, \ldots, m\}$. Let $\mu = EX$ and $H = -\sum p(x) \log p(x)$. Let $A^n = \left\{x^n \in \mathcal{X}^n : |-\frac{1}{n} \log p\left(x^n\right) - H| \leq \epsilon\right\}$. Let $B^n = \left\{x^n \in \mathcal{X}^n : \left|\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\right| \leq \epsilon\right\}$

1. Does $\Pr\left\{X^n \in A^n\right\} \longrightarrow 1$?
2. Does $\Pr\left\{X^n \in A^n \cap B^n\right\} \longrightarrow 1$?
3. Show that $|A^n \cap B^n| \leq 2^{n(H+\epsilon)}$ for all n
4. Show that $|A^n \cap B^n| \geq \left(\frac{1}{2}\right) 2^{n(H-\epsilon)}$ for n sufficiently large.

Solution.

1. Yes. By the Large Number Law,

$$-\frac{1}{n} \log p(x^n) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \to H(X) \quad \text{in probability} \tag{48}$$

, which implies that

$$\Pr\left\{X^n \in A^n\right\} = \Pr\left\{|-\frac{1}{n} \log p\left(x^n\right) - H| \leq \epsilon\right\} \to 1 \tag{49}$$

2. Yes. Part (1) implies that $\lim_{n\to\infty} \Pr\left\{X^n \in A^n\right\} = 1$.

   By strong law of large number we have $\lim_{n\to\infty} \Pr\left\{X^n \in B^n\right\} = 1$

   For arbitrary $\delta > 0$, there exists N_1, such that $\Pr\left\{X^n \in A^n\right\} > 1 - \frac{\delta}{2}$ for all $n > N_1$, and there exists N_2, such that $\Pr\left\{X^n \in B^n\right\} > 1 - \frac{\delta}{2}$ for all $n > N_2$. We take $N = \max\{N_1, N_2\}$, for any $n > N$, we have

$$\begin{aligned} \Pr\left\{X^n \in A^n \cap B^n\right\} &= \Pr\left\{X^n \in A^n\right\} + \Pr\left\{X^n \in B^n\right\} - \Pr\left\{X^n \in A^n \cup B^n\right\} \\ &> 1 - \frac{\delta}{2} + 1 - \frac{\delta}{2} - 1 = 1 - \delta \end{aligned} \tag{50}$$

   , which indicates that $\Pr\left\{X^n \in A^n \cap B^n\right\} \to 1$

3. From $x^n \in A^n$ we know that

$$2^{-n(H+\epsilon)} \leq p(x^n) \leq 2^{-n(H-\epsilon)} \tag{51}$$

$$\begin{aligned} 1 &= \sum_{x^n \in \S^n} p(x) \\ &\geq \sum_{x^n \in A^n \cap B^n} p(x) \\ &\geq 2^{-n(H(X)+\epsilon)} |A^n \cap B^n| \end{aligned} \tag{52}$$

   It follows that $|A^n \cap B^n| \leq 2^{-n(H+\epsilon)}$.

4. From $\Pr\{X^n \in A^n \cap B^n\} \to 1$, we take $\delta = \frac{1}{2}$, then there exists sufficiently large n, such that

$$\begin{aligned} \frac{1}{2} &\leq \Pr(X^n \in A^n \cap B^n) \\ &\leq \sum_{x^n \in A^n \cap B^n} p(x^n) \\ &\leq |A^n \cap B^n| 2^{-n(H(X)-\epsilon)} \end{aligned} \tag{53}$$

It follows that $|A^n \cap B^n| \geq \left(\frac{1}{2}\right) 2^{-n(H-\epsilon)}$.

□

Exercise 31 (Doubly stochastic matrices) An $n \times n$ matrix $P = [P_{ij}]$ is said to be doubly stochastic if $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$ for all i and $\sum_i P_{ij} = 1$ for all j. An $n \times n$ matrix P is said to be a permutation matrix if it is doubly stochastic and there is precisely one $P_{ij} = 1$ in each row and each column. It can be shown that every doubly stochastic matrix can be written as the convex combination of permutation matrices.

1. Let $\mathbf{a}^t = (a_1, a_2, \ldots, a_n), a_i \geq 0, \sum a_i = 1$, be a probability vector. Let $\mathbf{b} = \mathbf{a}P$, where P is doubly stochastic. Show that $\mathbf{b}$ is a probability vector and that $H(b_1, b_2, \ldots, b_n) \geq H(a_1, a_2, \ldots, a_n)$. Thus, stochastic mixing increases entropy.
2. Show that a stationary distribution μ for a doubly stochastic matrix P is the uniform distribution.
3. Conversely, prove that if the uniform distribution is a stationary distribution for a Markov transition matrix P, then P is doubly stochastic.

Proof.

1. From $\mathbf{b} = \mathbf{a}P$ we know that

$$b_j = \sum_{i=1}^{m} a_i p_{ij} \tag{54}$$

Then we have.

$$\begin{aligned} H(\mathbf{b}) - H(\mathbf{a}) &= -\sum_{j=1}^{m} \left(\sum_{i=1}^{m} a_i p_{ij} \right) \log b_j + \sum_{i=1}^{m} a_i \log a_i \\ &= -\sum_{i=1}^{m} a_i \left(\sum_{j=1}^{m} p_{ij} \log b_j \right) + \sum_{i=1}^{m} a_i \left(\sum_{j=1}^{m} p_{ij} \log a_i \right) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} a_i p_{ij} \log \frac{a_i}{b_j} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} b_j p_{ij} \left(\frac{a_i}{b_j} \log \frac{a_i}{b_j} \right) \\ &\geq \left(\sum_{i=1}^{m} \sum_{j=1}^{m} b_j p_{ij} \frac{a_i}{b_j} \right) \log \left(\sum_{i=1}^{m} \sum_{j=1}^{m} b_j p_{ij} \frac{a_i}{b_j} \right) \\ &= \left(\sum_{i=1}^{m} \sum_{j=1}^{m} p_{ij} a_i \right) \log \left(\sum_{i=1}^{m} \sum_{j=1}^{m} p_{ij} a_i \right) = 1 \cdot \log 1 = 0 \end{aligned} \tag{55}$$

2. By condition we have $\mu_i = \frac{1}{m}$ for any i. Since for any j,

$$\sum_{i=1}^{m} \mu_i p_{ij} = \frac{1}{m} \sum_{i=1}^{m} p_{ij} = \frac{1}{m} = \mu_i \tag{56}$$

We have that $\mu P = \mu$. The uniform distribution is a stationary distribution for a doubly stochastic matrix.

3. From $\mu P = \mu$ and $\mu = \frac{1}{m}$, we know that

$$\sum_{i=1}^{m} \frac{1}{m} p_{ij} = \frac{1}{m} \tag{57}$$

holds for any j, which implies

$$\sum_{i=1}^{m} p_{ij} = 1 \quad \text{for any} j \tag{58}$$

Then P is doubly stochastic.

□

Exercise 32 (Shuffes increase entropy) Argue that for any distribution on shuffles T and any distribution on card positions X that

$$\begin{aligned} H(TX) &\geq H(TX|T) \\ &= H\left(T^{-1}TX|T\right) \\ &= H(X|T) \\ &= H(X) \end{aligned}$$

if X and T are independent.

Proof. The first line holds because condition reduces entropy. The second line holds since T^{-1} can be given by the condition T. The last line holds if X and T are independent, which finishes the proof. □

Exercise 33 (Entropy rates of Markov chains)

1. Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{10} & 1 - p_{10} \end{bmatrix}$$

2. What values of p_{01}, p_{10} maximize the entropy rate?

3. Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 - p & p \\ 1 & 0 \end{bmatrix}$$

4. Find the maximum value of the entropy rate of the Markov chain of part (c). We expect that the maximizing value of p should be less than $\frac{1}{2}$, since the 0 state permits more information to be generated than the 1 state.

5. Let $N(t)$ be the number of allowable state sequences of length t for the Markov chain of part (c). Find $N(t)$ and calculate

$$H_0 = \lim_{t \to \infty} \frac{1}{t} \log N(t)$$

[Hint: Find a linear recurrence that expresses $N(t)$ in terms of $N(t-1)$ and $N(t-2)$. Why is H_0 an upper bound on the entropy rate of the Markov chain? Compare H_0 with the maximum entropy found in part (d).]

Solution.

1. We first calculate the stationary distribution μ.

$$\begin{cases} \mu P = \mu \\ \mu \mathbf{1}^T = 1 \end{cases} \Rightarrow \mu = \left[\frac{p_{10}}{p_{01}+p_{10}}, \frac{p_{01}}{p_{01}+p_{10}} \right] \tag{59}$$

By the entropy rate of Markov Chain,

$$\begin{aligned} H(\mathcal{X}) &= \sum_{i=1,2} \mu_i \left(\sum_{j=1,2} -p_{ij} \log p_{ij} \right) \\ &= \frac{p_{10} H(p_{01}) + p_{01} H(p_{10})}{p_{01}+p_{10}} \end{aligned} \tag{60}$$

2. Note that the entropy is upper bounded by its alphabet, since

$$\frac{1}{n} H\left(X_1, X_2, \ldots, X_n\right) \leq \frac{1}{n} \log |\mathcal{X}|^n = \log |\mathcal{X}| \tag{61}$$

Hence the maximal entropy rate for this problem is $\log 2 = 1$. This can be obtained when $p_{01} = p_{10} = \frac{1}{2}$, where

$$\begin{aligned} H(\mathcal{X}) &= \frac{p_{10} H(p_{01}) + p_{01} H(p_{10})}{p_{01}+p_{10}} \\ &= \frac{1}{2} + \frac{1}{2} = 1 \end{aligned} \tag{62}$$

Therefore, $p_{01} = p_{10} = \frac{1}{2}$ maximize the entropy rate.

3. We first calculate the stationary distribution μ.

$$\begin{cases} \mu P = \mu \\ \mu \mathbf{1}^T = 1 \end{cases} \Rightarrow \mu = \left[\frac{1}{p+1}, \frac{p}{p+1} \right] \tag{63}$$

By the entropy rate of Markov Chain,

$$\begin{aligned} H(\mathcal{X}) &= \sum_{i=1,2} \mu_i \left(\sum_{j=1,2} -p_{ij} \log p_{ij} \right) \\ &= \frac{-p \log p - (1-p) \log(1-p)}{p+1} \end{aligned} \tag{64}$$

4. We take the derivative of $H(\mathcal{X})$.

$$\frac{dH(\mathcal{X})}{dp} = \frac{\log(1-p) - \log p + \log(1-p)}{(p+1)^2} := 0 \quad \Rightarrow \quad p = \frac{3-\sqrt{5}}{2} \tag{65}$$

The maximal entropy rate is $\log \frac{1+\sqrt{5}}{2} \approx 0.6942$.

5. The transition matrix implies that there is no possibility that the last state is 1 with the last but one state to be 1. We can calculate the $N(t)$ recursively. If the last state is 0, the previous state possibilities add up to $N(t-1)$. If the last state is 1, however, the last but one state can only be 0. Then all the previous state possibilities will add up to $N(t-2)$. Further more, we can manually check the initial length that $N(1) = 2$, $N(2) = 3$. Now we have

$$N(t) = N(t-1) + N(t-2) \tag{66}$$

By solving the characteristic equation we know that $N(t)$ must be in the form like

$$N(t) = C_1 (\frac{1+\sqrt{5}}{2})^n + C_2 (\frac{1-\sqrt{5}}{2})^t \tag{67}$$

Then

$$\begin{aligned} H_0 &= \lim_{t\to\infty} \frac{1}{t} \log N(t) \\ &= \lim_{t\to\infty} \frac{1}{t} \log \left(C_1 (\frac{1+\sqrt{5}}{2})^t + C_2 (\frac{1-\sqrt{5}}{2})^t \right) \\ &= \lim_{t\to\infty} \frac{1}{t} \log C_1 (\frac{1+\sqrt{5}}{2})^t \\ &= \lim_{t\to\infty} \frac{1}{t} t \log \frac{1+\sqrt{5}}{2} = \log \frac{1+\sqrt{5}}{2} \end{aligned} \tag{68}$$

We find that H_0 is the upper bound of the entropy rate, and can be obtained with the maximum entropy found in pard (d). This is because by the property of entropy,

$$H\left(X_1, X_2, \ldots, X_n\right) \leq \log |X_1, X_2, \ldots, X_n| = N(t) \tag{69}$$

□

Exercise 34 (Maximal entropy graphs) Consider a random walk on a connected graph with four edges.

1. Which graph has the highest entropy rate?
2. Which graph has the lowest?

Solution. In a random walk, the next vertex will be arbitrarily chosen from the adjancent vertices of the current vertex. That is to say, all the edges are given the same weight.

We can formulate this problem as follows. By the inclusion-exclusion principle, there are at most 5 vertices in the graph given four edges. Their adjancent relation can be represented in a 5×5 0-1 matrix$W$, where $W_{ij} = 0$ or 1 and $W_{ij} = W_{ji}$.

Then the transition probability matrix P and the stationary distributionμ can be calculated.

$$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}} \tag{70}$$

$$\mu_i = \frac{W_i}{2W} \tag{71}$$

where $W_i = \sum_j w_{ij}$ and $W = \sum_i \frac{W_i}{2}$

Then the entropy rate can be calculated as

$$\begin{aligned} H(\mathcal{X}) &= H\left(X_2 | X_1\right) \\ &= H(X_2, X_1) - H(X_1) \\ &= H(\frac{1}{8}, \frac{1}{8}, \ldots, \frac{1}{8}) - H(\mu) \\ &= 3 - H(\mu) \end{aligned} \tag{72}$$

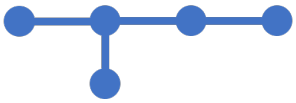
For 4 edges, there are five possible graphs, as has been shown in Figure 5.

Their corresponding entropy rates are 0.75, 0.84436, 1, 1.09436 and 1.

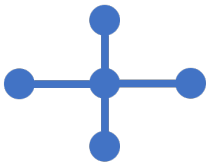
1. The fourth graph has the largest entropy rate.
2. The first graph has the lowest entropy rate.

□

$$W_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \mu_1^T = \begin{bmatrix} 1/8 \\ 1/4 \\ 1/4 \\ 1/4 \\ 1/8 \end{bmatrix}$$

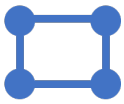


$$W_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \mu_2^T = \begin{bmatrix} 1/8 \\ 3/8 \\ 1/4 \\ 1/8 \\ 1/8 \end{bmatrix}$$



$$W_3 = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad P_3 = \begin{bmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mu_3^T = \begin{bmatrix} 1/2 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \end{bmatrix}$$

$$W_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad P_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \quad \mu_4^T = \begin{bmatrix} 1/8 \\ 3/8 \\ 1/4 \\ 1/4 \end{bmatrix}$$



$$W_5 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad P_5 = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \quad \mu_5^T = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

Figure 5: Five Possible Graphs for Four Edges

6 Data Compression

Exercise 35 (Slackness in the Kraft inequality) An instantaneous code has word lengths $l_1, l_2, \ldots, l_m$, which satisfy the strict inequality $\sum_{i=1}^{m} D^{-l_i} < 1$ The code alphabet is $\mathcal{D} = \{0, 1, 2, \ldots, D-1\}$. Show that there exist arbitrarily long sequences of code symbols in D^* which cannot be decoded into sequences of codewords.

Proof. W.l.o.g., we assume $l_1 \leq l_2 \leq \ldots \leq l_m$. Since the coding method is instantaneous, for any coding with the length of l_i, there exists no other coding that begins with the l_i corresponding coding. That is to say, if there exists a coding with l_i length, then $D^{l_m - l_i}$ codewords of the D^{l_m} codewords will be decodable. Adding all these decodable words up we have

$$\sum_{i=1}^{m} D^{l_m - l_i} = D^{l_m} \sum_{i=1}^{m} D^{-l_i} < D^{l_m} \tag{73}$$

,which implies that there exists some prefix words in D^{l_m} that are undecodable. Any arbitrary word that begins with such prefix will be unable to decode into sequence of codewords. □

Exercise 36 (Fix-free codes) A code is a fix-free code if it is both a prefix code and a suffix code. Let $l_1, l_2, \ldots, l_m$ be m positive integers. Prove that if $\sum_{k=1}^{m} 2^{-l_k} \leq \frac{1}{2}$ then there exists a binary fix-free code with codeword length $l_1, l_2, \ldots, l_m$

Proof. W.l.o.g., we assume $l_1 \leq l_2 \leq \ldots \leq l_m$. We prove by induction on m.

When $m = 1$, the conclusion is trivial.

We assume that when $m = k-1$, for any increasingly ordered positive integers $l_1, l_2, \ldots, l_{k-1}$, if $\sum_{i=1}^{k-1} 2^{-l_i} \leq \frac{1}{2}$, then there exists a binary fix-free code with codeword length $l_1, l_2, \ldots, l_{k-1}$.

Now let $m = k$. For every particular codeword, in terms of prefix codes, it will occupy the space of $2^{l_m - l_i}$ in the tree, while in terms of suffix codes, it wll also occpy the space of $2^{l_m - l_i}$ in the 2^{l_m} nodes. These two sets may overlap, but they will surely be less than $2 \times 2^{l_m - l_i}$. Hence for all codewords, we have

$$2 \sum_{i=1}^{k} 2^{l_m - l_i} \leq 2^{l_m} \tag{74}$$

It follows that $\sum_{i=1}^{k} 2^{-l_i} \leq \frac{1}{2}$. By removing l_k we have $\sum_{i=1}^{k-1} 2^{-l_i} < \frac{1}{2}$. By the induction hypothesis, we know that for $l_1, \ldots, l_{k-1}$ lengths of codewords, they can form a set of fix-free codes. Furthermore, the strict less relation tells us that there still remains space for the codeword l_k. Hence, the conclusion follows. □

Exercise 37 ($\frac{3}{4}$ fix-free codes) Prove that when

$$\sum_{k=1}^{m} 2^{-l_k} \leq \frac{3}{4}$$

the conclusion above holds.

Proof. The proof here is not complete. We try to solve the problem from two perspectives. The proof idea is based on this paper[1].

There exists no upper bounds greater than $\frac{3}{4}$. For any $\frac{3}{4} + \epsilon > \frac{3}{4}$, we choose k such that $2^{-k} < \epsilon$. We construct a list of codelengths with 1 and $2^{k-2} + 1$ ks. Then we have

$$\sum_{i=1}^{N} 2^{-l_i} = \frac{1}{2} + 2^{-k}(2^{k-2} + 1) = \frac{3}{4} + \epsilon \tag{75}$$

Our choice of codeword lengths satisfies our assumption, however, with the first codeword with length of 1 as a prefix and suffix, there are at most 2^{k-2} words of length k, which implies that our choice is invalid, contradiction.

[1] R. Ahlswede and B. Balkenhol and L. Khachatrian. Some properties of Fix-Free Codes

The conclusion with $\frac{3}{4}$ holds under some restrictions. We suppose that for all code lengths, either $l_i = l_{i+1}$ or $2l_i \leq l_{i+1}$. We prove that under this restriction the conclusion will hold.

W.l.o.g., we assume that $l_1 \leq l_2 \leq \ldots \leq l_m$. We prove by induction on m. The base case of $m = 1$ is trivial.

Assume that for any $n \leq m-1$, with $\sum_{i=1}^{n} 2^{-l_i} \leq \frac{3}{4}$ we can construct n different codeword lengths. We prove that this holds for the case of m.

Let m' be the largest index i with $l_i < l_{m'}$. By induction hypothesis we can construct a fix-free code C' with the lengths $l_1, \ldots, l_{m'}$. Note for every particular word with length l_i, it will occupy at most $2 \times 2^{l_m - l_i}$ nodes due to the prefix and suffix rule. However, with our restriction, $2^{l_m - 2l_i}$ nodes will be returned, since they will not actually be used as codewords. Therefore, at the l_m level, at most $2\sum_{i=1}^{m'} 2^{l_m - l_i} - \sum_{i=1}^{m'} 2^{l_m - 2l_i}$ nodes will be occupied.[2]

To ensure that the remaining $l_{m'+1}, \ldots, l_m$ can be added to the orginal code system we should have

$$2\sum_{i=1}^{m'} 2^{l_m - l_i} - \sum_{i=1}^{m'} 2^{l_m - 2l_i} \leq 2^{l_m} - (m - m') \tag{76}$$

Writing $K = m - m'$ and $\alpha = \sum_{i=1}^{m'} 2^{-l_i}$. (76) can be written as

$$2\alpha - \alpha^2 \leq 1 - \frac{K}{2^{l_m}} \tag{77}$$

With abbreviation $\beta = \sum_{i=1}^{m} 2^{-l_i} = \alpha + \frac{K}{2^{l_m}}$ and $\delta = \frac{K}{2^{l_m}}$ we get the equivalent inequality

$$\beta \leq 1 + \delta - \sqrt{\delta} \tag{78}$$

Note $\delta \in (0, 1)$, the right side has the minimal value of $\frac{3}{4}$ at $\delta = \frac{1}{4}$. Thus for any $\sum_{i=1}^{m} 2^{-l_i} \leq \frac{3}{4}$, the conclusion holds.

□

Exercise 38 (More Huffman codes) Find the binary Huffman code for the source with probabilities $\left(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}\right)$. Argue that this code is also optimal for the source with probabilities $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$

Figure 6: Huffman Code for $\left(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}\right)$.

Length	*CodeWord*	*X*	*Probability*				
2	11	1	1/3	1/3	1/3	3/5	1
2	01	2	1/5	1/5	2/5	2/5	
2	00	3	1/5	1/5	4/15		
3	101	4	2/15	4/15			
3	100	5	2/15				

Solution. The huffman code of $\left(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}\right)$. is found as Figure 6 shows.

Figure 7: Huffman Code for $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$

Length	*CodeWord*	*X*	*Probability*				
2	11	1	1/5	1/5	1/5	3/5	1
2	01	2	1/5	1/5	2/5	2/5	
2	00	3	1/5	1/5	2/5		
3	101	4	1/5	2/5			
3	100	5	1/5				

[2]The paper gives a more detailed formula, but I can't fully understand how the last component formula is derived. Fortunately leaving the that component out will not affect the proof here.

We see that for $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$, we can follow an identical process to construct the huffman code, as Figure 7 shows. Since huffman code is always optimal, the conclusion holds. □

Exercise 39 (Bad codes) Which of these codes cannot be Huffman codes for any probability assignment?

1. {0,10,11}
2. {00,01,10,110}
3. {01,10}

Solution.

1. {0,10,11} can be the huffman code for distribution $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right\}$.
2. {00,01,10,110} can be shortened by {00,01,10,11}. It cannot be a huffman code.
3. {01,10} can be shortened by {0,1}. It cannot be a huffman code.

□

Exercise 40 (Huffman 20 questions) Consider a set of n objects. Let $X_i = 1$ or 0 accordingly as the i th object is good or defective. Let $X_1, X_2, \ldots, X_n$ be independent with $\Pr\{X_i = 1\} = p_i$; and $p_1 > p_2 > \cdots > p_n > \frac{1}{2}$. We are asked to determine the set of all defective objects. Any yes-no question you can think of is admissible.

1. Give a good lower bound on the minimum average number of questions required.
2. If the longest sequence of questions is required by nature's answers to our questions, what (in words) is the last question we should ask? What two sets are we distinguishing with this question? Assume a compact (minimum average length) sequence of questions.
3. Give an upper bound (within one question) on the minimum average number of questions required.

Solution.

1. The asking process can be modeled into the problem of constructing a compressed code for the sequence $X_1, X_2, \ldots, X_n$. The set of all defective objects can be determined if we determine the codeword we've constructed. The question we are asking is a yes-no question, so the codeword is based on a 2-ray alphabet. We can use entropy of $X_1, X_2, \ldots, X_n$ to give a lower bound on the average codeword length.

$$\begin{aligned} L^* &\geq H_2(X_1, X_2, \ldots, X_n) \\ &= \sum_{i=1}^{n} H(X_i) = \sum_{i=1}^{n} H(p_i) \end{aligned} \tag{79}$$

2. The longest sequence implies that we are distinguishing between the two least cases in this problem, i.e. the case where all objects are good ($\prod_{i=1}^{n}(1-p_i)$) and the case where all objects are good except the one that has the least probability to be defective ($p_n \prod_{i=1}^{n-1}(1-p_i)$). The question will be like "Is X_n defective?".

3. Using the same notion in part 1, the upper bound of the minimum average number of questions (codeword length) will be

$$\begin{aligned} L^* &\leq H_2(X_1, X_2, \ldots, X_n) + 1 \\ &= \sum_{i=1}^{n} H(X_i) + 1 = \sum_{i=1}^{n} H(p_i) + 1 \end{aligned} \tag{80}$$

□

Exercise 41 (Simple optimum compression of a Markov source) Consider the three-state Markov process $U_1, U_2, \ldots$ having transition matrix

U_n \ U_{n-1}	S_1	S_2	S_3
S_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
S_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
S_3	0	$\frac{1}{2}$	$\frac{1}{2}$

Thus, the probability that S_1 follows S_3 is equal to zero. Design three codes C_1, C_2, C_3 (one for each state 1,2 and 3, each code mapping elements of the set of S_i 's into sequences of 0 's and 1 's, such that this Markov process can be sent with maximal compression by the following scheme:

1. Note the present symbol $X_n = i$
2. Select code C_i
3. Note the next symbol $X_{n+1} = j$ and send the codeword in C_i corresponding to j
4. Repeat for the next symbol.

What is the average message length of the next symbol conditioned on the previous state $X_n = i$ using this coding scheme? What is the unconditional average number of bits per source symbol? Relate this to the entropy rate $H(\mathcal{U})$ of the Markov chain.

Solution. We can design the codes using the Huffman method.

	S_1	S_2	S_3
C_1	1	01	00
C_2	01	1	00
C_3	N/A	0	1

We first calculate the stationery distribution of the Markov Chain.

$$\begin{cases} \mu P = \mu \\ \mu \mathbf{1}^T = 1 \end{cases} \Rightarrow \mu = \left[\frac{2}{9}, \frac{4}{9}, \frac{1}{3}\right] \tag{81}$$

The average length is

$$\sum_{i=1}^{3} \mu_i \sum_{j=1}^{3} L(C_{ij}) = \frac{2}{9} \cdot \frac{3}{2} + \frac{4}{9} \cdot \frac{3}{2} + \frac{1}{3} \cdot 1 = \frac{4}{3} \tag{82}$$

The entropy rate can be calculated as

$$H(\mathcal{U}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij} = \frac{4}{3} \tag{83}$$

They are the same because the optimal code length is its entropy, which can be expressed with the notion of "average" entropy $\frac{1}{n} H(U_1, U_2, \ldots, U_n, \ldots)$, which approximates to $H(\mathcal{U})$. □

Exercise 42 (Shannon codes and Huffman codes) Consider a random variable X that takes on four values with probabilities $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right)$

1. Construct a Huffman code for this random variable.

2. Show that there exist two different sets of optimal lengths for the codewords; namely, show that codeword length assignments (1,2,3,3) and (2,2,2,2) are both optimal.

3. Conclude that there are optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\left\lceil \log \frac{1}{p(x)} \right\rceil$

Solution.

1. A Huffman code is constructed as Figure 8 shows.

Figure 8: Huffman Code for $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right)$

Length	***CodeWord***	***X***	***Probability***			
1	1	1	1/3 —	1/3 —	1/3 —	1
2	01	2	1/3 —	1/3 —	2/3	
3	001	3	1/4 —	1/3		
3	000	4	1/12			

2. Another Huffman code is constructed as Figure 9 shows. They are both optimal for the distribution but with distinct length assignments.

Figure 9: Huffman Code for $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right)$

Length	***CodeWord***	***X***	***Probability***			
2	11	1	1/3 —	1/3 —	2/3 —	1
2	10	2	1/3 —	1/3	1/3	
2	01	3	1/4 —	1/3		
2	00	4	1/12			

3. The case of $X = 3$ in Figure 8 serves as an example. It has the length of 3, exceeding the Shannon length $\left\lceil \log \frac{1}{p(x)} \right\rceil = 2$. It can be concluded that in some cases, we can construct optimal codes with codeword lengths for some symbols that exceed the Shannon code length.

□

Exercise 43 (Data compression) Find an optimal set of binary codeword lengths $l_1, l_2, \ldots$ (minimizing $\sum p_i l_i$) for an instantaneous code for each of the following probability mass functions:

1. $\mathbf{p} = \left(\frac{10}{41}, \frac{9}{41}, \frac{8}{41}, \frac{7}{41}, \frac{7}{41}\right)$
2. $\mathbf{p} = \left(\frac{9}{10}, \left(\frac{9}{10}\right)\left(\frac{1}{10}\right), \left(\frac{9}{10}\right)\left(\frac{1}{10}\right)^2, \left(\frac{9}{10}\right)\left(\frac{1}{10}\right)^3, \ldots\right)$

Solution. The optimal code is given in Figure 10 by the Huffman rule. Note that in problem (2) we have that any probability is greater than the sum of the probabilities less than itself. Hence, we can construct the Huffman code in a monotonous order □

Figure 10: Huffman Code for Exercise 43

Length	*CodeWord*	*X*	*Probability*				
2	11	1	10/41	10/41	10/41	24/41	1
2	01	2	9/41	9/41	17/41	17/41	
2	00	3	8/41	8/41	14/41		
3	101	4	7/41	14/41			
3	100	5	7/41				

Length	*CodeWord*	*X*	*Probability*				
1	1	1	0.9	…	0.9	…	0.9	1
2	01	2	0.09	…	0.09	…	0.099…	
…	…	…	…	…	…	…		
i	00…01	i	0.00..09	…	0.00…099			
i+1	00…001	i+1	0.00…009	…				
…	…	…	…	…				

7 Information Channel

Exercise 44 (BSC) Calculate the channel capacity of BSC.

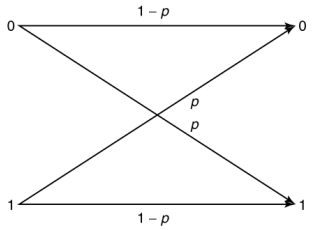


Figure 11: Binary Symmetric Channel

Solution.

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} (H(Y) - \sum p(x) H(Y|X=x)) \\ &= \max_{p(x)} (H(Y) - \sum p(x) H(p)) \\ &= \max_{p(x)} (H(Y) - H(p)) = \log 2 - H(p) \end{aligned} \tag{84}$$

The maximal value can be obtained when X and Y are uniformly distributed. □

Exercise 45 (BSC) Calculate the channel capacity of BEC.

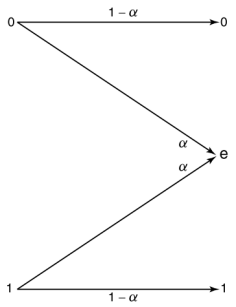


Figure 12: Binary Erasure Channel

Solution.

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} (H(Y) - H(\alpha)) \end{aligned} \tag{85}$$

Let $\Pr(X=1) = \pi$, then

$$\begin{aligned} H(Y) &= H((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha)) \\ &= H(\alpha) + (1-\alpha)H(\pi) \\ C &= \max_{p(x)} (H(Y) - H(\alpha)) \\ &= \max_{\pi} ((1-\alpha)H(\pi) + H(\alpha) - H(\alpha)) \\ &= \max_{\pi} (1-\alpha)H(\pi) = 1-\alpha \end{aligned} \tag{86}$$

The maximal value can be obtained when X is uniformly distributed. □

Exercise 46 (Using two channels at once) Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2 respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1|x_1) \times p(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ are sent simultaneously, resulting in $\mathcal{Y}_1, \mathcal{Y}_2$. Find the capacity of this channel.

Solution. By condition we know that $p(y_1, y_2|x_1, x_2) = p(y_1|x_1) \times p(y_2|x_2)$. It follows by definition that

$$H(Y_1, Y_2|X_1, X_2) = H(Y_1|X_1) + H(Y_2|X_2). \tag{87}$$

Therefore,

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\ &= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\ &\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\ &= I(Y_1; X_1) + I(Y_2; X_2) \\ &\leq C_1 + C_2 \end{aligned} \tag{88}$$

Hence $C = C_1 + C_2$. The equality holds when $p(x_1, x_2) = p^*(x_1)p^*(x_2)$, where $p^*(x_1)$ and $p^*(x_2)$ are the optimal distribution corresponding to the original channel capacities. □

Exercise 47 (Z-channel) The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, x, y \in \{0, 1\}$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

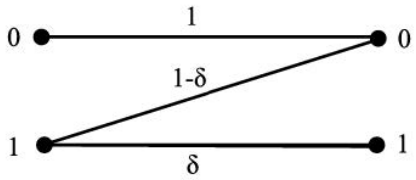


Figure 13: Z-Channel

Solution. Assume that $\Pr(X=1)=\pi$, given the transition matrix Q, we have that $\Pr(Y=1)=\frac{\pi}{2}$, $\Pr(Y=1)=1-\frac{\pi}{2}$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(\frac{\pi}{2}) - \pi H(\frac{1}{2}) - (1-\pi)H(1) \\ &= -\frac{\pi}{2}\log(\frac{\pi}{2}) - \left(1-\frac{\pi}{2}\right)\log\left(1-\frac{\pi}{2}\right) - \pi \end{aligned} \tag{89}$$

Since the function is a concave function, we find its maximal value by taking the derivative.

$$\begin{aligned} \frac{dI(X;Y)}{d\pi} &= -\frac{1}{2}\log(\frac{\pi}{2}) - \frac{1}{2\ln 2} + \frac{1}{2\ln 2} + \left(1-\frac{1}{2}\right)\log\left(1-\frac{\pi}{2}\right) - 1 \\ &= \frac{1}{2}\log\frac{2-\pi}{\pi} - 1 := 0 \end{aligned} \tag{90}$$

It follows that the mutual information is at its maximum when $\pi=\frac{2}{5}$. The channel capacity is $C \approx 0.32193$. □

Exercise 48 (Erasures and errors in a binary channel) Consider a channel with binary inputs that has both erasures and errors. Let the probability of error be ϵ and the probability of erasure be α, so the channel is follows:

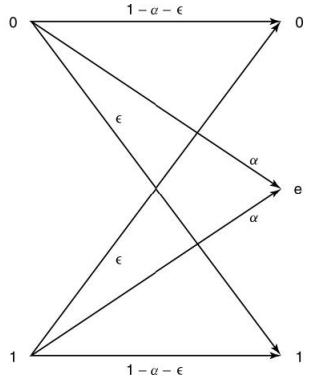


Figure 14: Erasures and Errors

Find the capacity of this channel.

Solution. The transition matrix is as follows.

X \ Y	0	e	2
0	$1-\alpha-\epsilon$	α	ϵ
1	ϵ	α	$1-\alpha-\epsilon$

The matrix is row symmetric, we have that

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\alpha, \epsilon, 1-\alpha-\epsilon) \tag{91}$$

Furthermore, note that the distribution of Y in terms of the distribution of X is symmetric, i.e.

$$H(Y)|_{\Pr(X=0)=\pi} = H(Y)|_{\Pr(X=0)=1-\pi} \tag{92}$$

and that the entropy function is a concave function. It follows that the maximal value must be obtained at $p(x) = \frac{1}{2}$ for $x = 0, 1$.

$$C = H(\frac{1-\alpha}{2}, \alpha, \frac{1-\alpha}{2}) - H(\alpha, \epsilon, 1-\alpha-\epsilon) \tag{93}$$

□

Exercise 49 (Additive noise channel) Find the channel capacity of the following discrete memoryless channel:

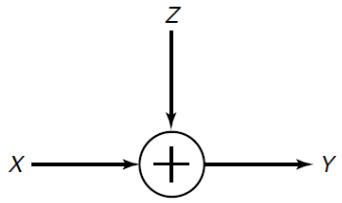


Figure 15: Discrete Memoryless Channel

where $\Pr\{Z = 0\} = \Pr\{Z = a\} = \frac{1}{2}$. The alphabet for x is $\mathbf{X} = \{0,1\}$. Assume that Z is independent of X. Observe that the channel capacity depends on the value of a.

Solution. The condition implies that $a \neq 0$, but after the plus operation the distribution of Y may vary. Hence we discuss the value of a in several cases.

1. $a = \pm 1$, the result of $X + a$ and $X + 0$ may overlap. We take the case of $a = 1$ as example. The transition matrix is as follows.

X \ Y	0	1	2
0	$\frac{1}{2}$	$\frac{1}{2}$	0
1	0	$\frac{1}{2}$	$\frac{1}{2}$

The matrix is row symmetric, assume that $\Pr(X = 0) = \pi$we have that

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(\frac{1}{2}\pi, \frac{1}{2}, \frac{1}{2}\pi) - H(\frac{1}{2}) \\ &\leq H(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\pi) - H(\frac{1}{2}) = \frac{1}{2} \end{aligned} \tag{94}$$

$C = \frac{1}{2}$. The maximal value is obtained at $\pi = \frac{1}{2}$.

2. $a \neq \pm 1$. Then the output Y will not overlap for different X. Hence Y is a function of X. We have

$$C = \max I(X;Y) = \max H(X) = 1 \tag{95}$$

The maximal value is obtained when $p(x) = \frac{1}{2}$ for $x = 0, 1$.

□

Exercise 50 (Channel capacity) Consider the discrete memoryless channel $Y = X + Z (\text{mod} 11)$, where

$$Z = \begin{pmatrix} 1, & 2, & 3 \\ \frac{1}{3}, & \frac{1}{3}, & \frac{1}{3} \end{pmatrix}$$

and $X \in \{0, 1, \ldots, 10\}$. Assume that Z is independent of X.

1. Find the capacity.
2. What is the maximizing $p^*(x)$?

Solution. Note that the transition matrix of this channel is

X \ Y	0	1	2	...	9	10
0	0	$\frac{1}{3}$	$\frac{1}{3}$	...	0	0
1	0	0	$\frac{1}{3}$	...	0	0
2	0	0	0	...	0	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮
9	$\frac{1}{3}$	$\frac{1}{3}$	0	...	0	$\frac{1}{3}$
10	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	...	0	0

Note that the matrix is both column and row symmetric. Therefore we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\mathbf{r}) \\ &\leq \log |\mathcal{Y}| - H(\mathbf{r}) = \log 11 - \log 3 = \log \frac{11}{3} \end{aligned} \tag{96}$$

$C = \log \frac{11}{3}$. The maximal value is obtained when $p^*(x) = \frac{1}{11}$ for every $x \in \{0, 1, \ldots, 10\}$ □

Exercise 51 (Zero-error capacity) A channel with alphabet {0,1,2,3,4} has transition probabilities of the form

$$p(y|x) = \begin{cases} 1/2 & \text{if } y = x \pm 1 \bmod 5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the capacity of this channel in bits.

(b) The zero-error capacity of a channel is the number of bits per channel use that can be transmitted with zero probability of error. Clearly, the zero-error capacity of this pentagonal channel is at least 1 bit (transmit 0 or 1 with probability 1/2). Find a block code that shows that the zero-error capacity is greater than 1 bit. Can you estimate the exact value of the zero-error capacity? (Hint: Consider codes of length 2 for this channel.)

Solution. 1. Note that the transition matrix is row and column symmetric. It follows that

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\frac{1}{2}) \\ &\leq \log |\mathcal{Y}| - H(\frac{1}{2}) = \log 5 - 1 \approx 1.322 \end{aligned} \tag{97}$$

$C = 1.322$ bits. The maximal value is obtained when X and Y are uniformly distributed.

2. According to the hint, we try to build a 2-tuple code with the alphabet {0,1,2,3,4}. Note that for every input codeword (X_1, X_2), there are four possibilities of output, namely $((X_1 \pm 1) \bmod 5, (X_1 \pm 1) \bmod 5)$, but the good news is that if we can ensure that there are no other input codewords (and their possible outputs) occupying the 4 entries, we can infer the input determintly.

   Figure 16 gives an example of zero-error codes, where cells identified by different colors represent a codeword. The input codewords are (0,0),(1,2),(2,4),(3,1) and (4,3). Given an output codeword, its corresponding input codeword can be determined directly from the table.

(X_1, X_2)	0	1	2	3	4
0	(0,0)				
1			(1,2)		
2					(2,4)
3		(3,1)			
4				(4,3)	

Figure 16: A Construction of Zero-Error Codes

   With the codes above, we can find that the number of bits per channel use is $\frac{1}{2}\log 5 > 1$.

□

8 Channel Capacity

Exercise 52 (Capacity of the carrier pigeon channel) Consider a commander of an army besieged in a fort for whom the only means of communication to his allies is a set of carrier pigeons. Assume that each carrier pigeon can carry one letter (8 bits), that pigeons are released once every 5 minutes, and that each pigeon takes exactly 3 minutes to reach its destination.

1. Assuming that all the pigeons reach safely, what is the capacity of this link in bits/hour?
2. Now assume that the enemies try to shoot down the pigeons and that they manage to hit a fraction α of them. since the pigeons are sent at a constant rate, the receiver knows when the pigeons are missing. What is the capacity of this link?
3. Now assume that the enemy is more cunning and that every time they shoot down a pigeon, they send out a dummy pigeon carrying a random letter (chosen uniformly from all 8-bit letters). What is the capacity of this link in bits/hour? Set up an appropriate model for the channel in each of the above cases, and indicate how to go about finding the capacity.

Solution.

1. The capacity is 8 bits/5mins = 96 bits/hour.
2. The process can be modeled as an erasure model. Consider the transmission of a pigeon of 8 bits (256 alphabets), it has the probability of α to be erased. We assume the message sent is $X \in \{0, \ldots, 256\}$, the

message received is $Y \in \{e, 0, \ldots, 256\}$. The capacity of a single transmission is

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} (H(Y) - H(\alpha)) \\ &= \max_{\pi} ((1-\alpha) H(Y|X=Y) + H(\alpha) - H(\alpha)) \\ &= \log(256)(1-\alpha) \end{aligned} \tag{98}$$

Hence the capacity is $8(1-\alpha)$ bits per pigeon, or $96(1-\alpha)$ bits per hour.

3. The process can be modeled as a binary symmetric channel. Consider the transmission of a pigeon of 8 bits (256 alphabets), it has the probability of $\frac{\alpha}{256}$ to be changed to a different value. We assume the message sent is $X \in \{0, \ldots, 256\}$, the message received is $Y \in \{0, \ldots, 256\}$. The capacity of a single transmission is

$$\begin{aligned} C &= \max I(X;Y) \\ &= \max H(Y) - H(Y|X) \\ &= \max H(Y) - \sum p(x) H(Y|X=x) \\ &= \max H(Y) - H\left(1 - \frac{255\alpha}{256}, \frac{\alpha}{256}, \ldots, \frac{\alpha}{256}\right) \\ &= 16 + \frac{255}{256}\alpha \log \frac{256 - 255\alpha}{\alpha} - \log(256 - 255\alpha) \end{aligned} \tag{99}$$

Hence the capacity is $16 + \frac{255}{256}\alpha \log \frac{256-255\alpha}{\alpha} - \log(256 - 255\alpha)$ bits per pigeon, or $12(16 + \frac{255}{256}\alpha \log \frac{256-255\alpha}{\alpha} - \log(256 - 255\alpha))$ bits per hour.

□

Exercise 53 (Channel with two independent looks at Y) Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X

1. Show that $I(X; Y_1, Y_2) = 2I(X; Y_1) - I(Y_1, Y_2)$
2. Conclude that the capacity of the channel $(X \mapsto Y_1, Y_2)$ is less than twice the capacity of channel $(X \mapsto Y_1)$

Proof.

1.

$$\begin{aligned} I(X; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X) \\ &= H(Y_1, Y_2) - H(Y_1|X) - H(Y_2|X) \\ &= H(Y_1) + H(Y_2) - I(Y_1; Y_2) - H(Y_1|X) - H(Y_2|X) \\ &= I(X; Y_1) + I(X; Y_2) - I(Y_1, Y_2) \\ &= 2I(X; Y_1) - I(Y_1, Y_2) \end{aligned} \tag{100}$$

2.

$$\begin{aligned} C_1 &= \max_{p(x)} I(X; Y_1, Y_2) \\ &= \max_{p(x)} (2I(X; Y_1) - I(Y_1, Y_2)) \\ &\leq \max_{p(x)} 2I(X; Y_1) \\ &= 2C_2 \end{aligned} \tag{101}$$

□

Exercise 54 (Binary multiplier channel)

1. Consider the channel $Y = XZ$, where X and Z are independent binary random variables that take on values 0 and 1.Z is Bernoulli(α) [i.e., $P(Z = 1) = \alpha$]. Find the capacity of this channel and the maximizing distribution on X.

2. Now suppose that the receiver can observe Z as well as Y. What is the capacity?

Solution.

1. Let $p(X = 1) = \pi$

$$\begin{aligned} I(Y;X) &= H(Y) - H(Y|X) \\ &= H(\alpha\pi) - \pi H(\alpha) \\ &= -\alpha\pi \log \alpha\pi - (1-\alpha\pi)\log(1-\alpha\pi) - H(\alpha)\pi \end{aligned} \tag{102}$$

We find the maximal value by taking the derivative.

$$\begin{aligned} \frac{dI(Y;X)}{d\pi} &= -\alpha \log \alpha\pi - \frac{\alpha}{\ln 2} + \alpha \log(1-\alpha\pi) + \frac{\alpha}{\ln 2} - H(\alpha) \\ &= \alpha \log \frac{1-\alpha\pi}{\alpha\pi} - H(\alpha) := 0 \\ \frac{1-\alpha\pi}{\alpha\pi} &= 2^{\frac{H(\alpha)}{\alpha}} \\ \pi^* &= \frac{1}{\alpha\left(2^{\frac{H(\alpha)}{\alpha}} + 1\right)} \\ C &= \log\left(2^{\frac{H(\alpha)}{\alpha}} + 1\right) - \frac{H(\alpha)}{\alpha} \end{aligned} \tag{103}$$

2. Let $p(X = 1) = \pi$

$$\begin{aligned} I(Y,Z;X) &= H(Y,Z) - H(Y,Z|X) \\ &= H(Y|Z) + H(Z) - H(Y|Z,X) - H(Z|X) \\ &= H(Y|Z) - H(XZ|Z,X) = H(XZ|Z) \\ &= \alpha H(\pi) \leq \alpha \end{aligned} \tag{104}$$

The maximal value is obtained when X is normally distributed.

□

Exercise 55 (Noise Channel) Consider the channel $\mathcal{X} = \{0,1,2,3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$

1. What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

2. What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

Solution.

1.

$$\begin{aligned} I(Y;X) &= H(X) - H(X|X+Z) \\ &\leq H(x) \leq \log 4 = 2 \end{aligned} \tag{105}$$

The maximal capacity will be achieved when X is a function of Y, i.e. given Y, X can be determined, and that X is normally distributed. A choice of $\mathcal{Z}$ can be $\mathcal{Z} = \{0, 5, 10\}$

2.

$$\begin{aligned} I(Y;X) &= H(Y) - H(X+Z|X) = H(X+Z) - H(Z) \\ &= H(X+Z) - \log 3 \end{aligned} \tag{106}$$

Note that for any single possible X, there will be at most 3 distinct corresponding Y. In order to minimize the maximal $H(X+Z)$, we need to set Z to be 3 continuous numbers, so that the uncertainty of $X+Z$ will be minimized. We take $\mathcal{Z} = \{0,1,2\}$ for example. Let the distribution of $\mathcal{X}$ to be $\{p_1, p_2, p_3, p_4\}$ Then

$$H(X+Z) = H(\frac{p_1}{3}, \frac{p_1+p_2}{3}, \frac{p_1+p_2+p_3}{3}, \frac{p_2+p_3+p_4}{3}, \frac{p_3+p_4}{3}, \frac{p_4}{3}) \leq \log 6 \tag{107}$$

The maximal value is obtained when $p_1 = p_4 = \frac{1}{2}$. Hence the channel capacity is 1.

□

Exercise 56 (Erasure channel.) Let $\{\mathcal{X}, p(y|x), \mathcal{Y}\}$ be a discrete memoryless channel with capacity C. Suppose that this channel is cascaded immediately with an erasure channel $\{\mathcal{Y}, p(s|y), \mathcal{S}\}$ that erases α of its symbols. Specifically, $\mathcal{S} = \{y_1, y_2, \ldots, y_m, e\}$, and

$$\begin{aligned} &\Pr\{S = y|X = x\} = \bar{\alpha}p(y|x), \quad y \in \mathcal{Y} \\ &\Pr\{S = e|X = x\} = \alpha \end{aligned}$$

Determine the capacity of this channel.

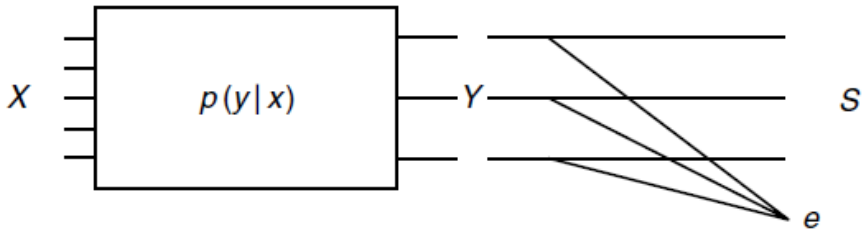


Figure 17: Erasure Channel

Solution.

$$
\begin{aligned}
I(X;S) &= H(S) - H(S|X) \\
&= H(S) - \sum_{x\in\mathcal{X}} p(x) \left(\left(\sum_{y\in\mathcal{Y}} \bar{\alpha} p(y|x) \log \bar{\alpha} p(y|x) \right) - \alpha \log \alpha \right) \\
&= H(S) - \sum_{x\in\mathcal{X}} p(x) \left(\sum_{y\in\mathcal{Y}} \bar{\alpha} p(y|x) \left(\log \bar{\alpha} + \log p(y|x) \right) \right) - \alpha \log \alpha \\
&= H(S) - \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} \bar{\alpha} p(y|x) \log p(y|x) - \alpha \log \alpha - \bar{\alpha} \log \bar{\alpha} \\
&= H(S) - H(\alpha) - (1-\alpha) H(Y|X) \\
&= -\sum_{y\in\mathcal{Y}} \bar{\alpha} p(y) \log \bar{\alpha} p(y) - \alpha \log \alpha - H(\alpha) - (1-\alpha) H(Y|X) \\
&= -\sum_{y\in\mathcal{Y}} \bar{\alpha} p(y) \left(\log \bar{\alpha} + \log p(y) \right) - \alpha \log \alpha - H(\alpha) - (1-\alpha) H(Y|X) \\
&= -\sum_{y\in\mathcal{Y}} \bar{\alpha} p(y) \log p(y) - \bar{\alpha} \log \bar{\alpha} - \alpha \log \alpha - H(\alpha) - (1-\alpha) H(Y|X) \\
&= (1-\alpha) H(Y) + H(\alpha) - H(\alpha) - (1-\alpha) H(Y|X) \\
&= (1-\alpha) I(X;Y)
\end{aligned}
\tag{108}
$$

Therefore the capacity of this channel is $(1-\alpha)C$. □

Exercise 57 (Choice of channels) Find the capacity C of the union of two channels $(\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2)$, where at each time, one can send a symbol over channel 1 or channel 2 but not both. Assume that the output alphabets are distinct and do not intersect.

1. Show that $2^C = 2^{C_1} + 2^{C_2}$. Thus, 2^C is the effective alphabet size of a channel with capacity C
2. Compare with Problem 2.10 where $2^H = 2^{H_1} + 2^{H_2}$, and interpret part (a) in terms of the effective number of noise-free symbols.
3. Use the above result to calculate the capacity of the following channel.

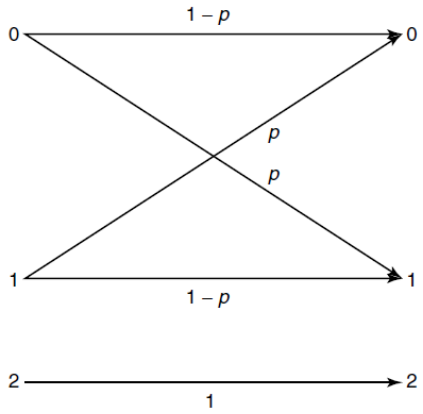


Figure 18: Choice of Channels

Solution.

1. We assume

$$p(X = x) = \begin{cases} \pi, x = X_1 \\ 1 - \pi, x = X_2 \end{cases} \tag{109}$$

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= -\sum_{x \in \mathcal{X}_\infty \cup \mathcal{X}_\in} p_X(x) \log p_X(x) - \pi H(X_1|Y_1) - \bar{\pi} H(X_2|Y_2) \\ &= -\sum_{x \in \mathcal{X}_\infty} \pi p_{X_1}(x) \log \pi p_{X_1}(x) - \sum_{x \in \mathcal{X}_\in} \bar{\pi} p_{X_2}(x) \log \bar{\pi} p_{X_2}(x) - \pi H(X_1|Y_1) - \bar{\pi} H(X_2|Y_2) \\ &= H(\pi) + \pi H(X_1) + \bar{\pi} H(X_2) - -\pi H(X_1|Y_1) - \bar{\pi} H(X_2|Y_2) \\ &= H(\pi) + \pi I(X_1;Y_1) + \bar{\pi} I(X_2;Y_2) \\ &\leq H(\pi) + \pi C_1 + \bar{\pi} C_2 \end{aligned} \tag{110}$$

The equality holds when $p(x_1)$ and $p(x_2)$ are at their optimal distribution. We take the derivative on π to calculate the capacity.

$$\begin{aligned} \frac{dI(X;Y)}{d\pi} &= -\log \pi + \log(1-\pi) + C_1 - C_2 := 0 \\ \frac{\pi^*}{1-\pi^*} &= 2^{C_1 - C_2} \\ \pi^* &= \frac{2^{C_1}}{2^{C_1} + 2^{C_2}} \\ \Rightarrow C &= \log(2^{C_1} + 2^{C_2}) \end{aligned} \tag{111}$$

It follows that $2^C = 2^{C_1} + 2^{C_2}$

2. Using the same notion as Assignment3.7, 2^C is the effective number of noise-free symbols. Since the two channels are disjoint in their alphabets, the noise-free symbols will also not overlap. Hence the combined capacity is equal to the sum of two sub-channels' noise-free symbols, i.e. $2^C = 2^{C_1} + 2^{C_2}$

3. From the Binary Symmetric Channel we know that $C_1 = 1 - H(p)$. $C_2 = 0$. Hence we have $C = \log(2^{1-H(p)} + 1)$

□

Exercise 58 (Capacity) Suppose that channel $\mathcal{P}$ has capacity C, where $\mathcal{P}$ is $m \times n$ channel matrix.

1. What is the capacity of

$$\tilde{\mathcal{P}} = \begin{bmatrix} \mathcal{P} & 0 \\ 0 & 1 \end{bmatrix}?$$

2. What about the capacity of

$$\hat{\mathcal{P}} = \begin{bmatrix} \mathcal{P} & 0 \\ 0 & I_k \end{bmatrix}?$$

where I_k if the $k \times k$ identity matrix.

Solution.

1. Note that the new channel is composed of two disjoint sub-channels, $\mathcal{P}$ and a one-to-one channel whose capacity is 1. By the conclusion of Exercise 57, we know that $C_{\mathcal{P}} = \log(2^C + 2^0) = \log(2^C + 1)$.

2. Note that the new channel is composed of two disjoint sub-channels, $\mathcal{P}$ and a one-to-one channel whose capacity is $\log k$. By the conclusion of Exercise 57, we know that $C_{\mathcal{P}} = \log(2^C + 2^{\log k}) = \log(2^C + k)$.

□

Exercise 59 (Channel with memory) Consider the discrete memoryless channel $Y_i = Z_i X_i$ with input alphabet $X_i \in \{-1, 1\}$

1. What is the capacity of this channel when $\{Z_i\}$ is i.i.d. with

$$Z_i = \begin{cases} 1, p = 0.5 \\ -1, p = 0.5? \end{cases}$$

Now consider the channel with memory. Before transmission begins, Z is randomly chosen and fixed for all time. Thus, $Y_i = ZX_i$

2. What is the capacity if

$$Z = \begin{cases} 1, p = 0.5 \\ -1, p = 0.5? \end{cases}$$

Solution.

1. The channel can be modeled as a binary symmetric channel with false rate 0.5, hence the capacity is $1 - H(0.5) = 0$.
2. Since Z is fixed for all time, we can set X_0 to be 1, and use Y_i to determine what Z is. Then for the rest of the transmissions, Y becomes a function of X. Therefore the capacity will approach $\log |\mathcal{X}| = 1$ as $n \to \infty$.

□

Exercise 60 (Tall, fat people) Suppose that the average height of people in a room is 5 feet. Suppose that the average weight is 100 lb.

1. Argue that no more than one-third of the population is 15 feet tall.
2. Find an upper bound on the fraction of 300 -Ib 10 -footers in the room.

Solution.

1. The height of people can be regarded as a random variable X. By Markov's inequality,

$$\Pr\{X \geq 15\} \leq \frac{5}{15} = \frac{1}{3}$$

2. The weight of people can be regarded as a random variable Y. By Markov's inequality,

$$\Pr\{Y \geq 300\} \leq \frac{100}{300} = \frac{1}{3}$$

and that

$$\Pr\{X \geq 10\} \leq \frac{5}{10} = \frac{1}{2}$$

Hence there are at most $\frac{1}{3}$ 300-lb, 10-footers among all the people.

□

9 Differential Entropy

Exercise 61 (Differential entropy) Evaluate the differential entropy $h(X) = -\int f \ln f$ for the following:

1. The exponential density, $f(x) = \lambda e^{-\lambda x}, x \geq 0$
2. The Laplace density, $f(x) = \frac{1}{2}\lambda e^{-\lambda |x|}$
3. The sum of X_1 and X_2, where X_1 and X_2 are independent normal random variables with means μ_i and variances $\sigma_i^2, i = 1, 2$

Solution.

1.
$$\begin{aligned} h(X) &= -\int_0^\infty \lambda e^{-\lambda x} \ln \lambda e^{-\lambda x} dx \\ &= -\int_0^\infty \lambda e^{-\lambda x} (\ln \lambda - \lambda x) dx \\ &= \ln \lambda \int_0^\infty e^{-\lambda x} d(-\lambda x) - \lambda \int_0^\infty x de^{-\lambda x} \\ &= \ln \lambda e^{-\lambda x}\Big|_0^\infty - \lambda x e^{-\lambda x}\Big|_0^\infty - \int_0^\infty e^{-\lambda x} d(-\lambda x) \\ &= -\ln \lambda + 1 = \ln \frac{e}{\lambda} \end{aligned} \tag{112}$$

2.
$$\begin{aligned} h(X) &= -\int_{-\infty}^{+\infty} \frac{1}{2}\lambda e^{-\lambda|x|} \ln \frac{1}{2}\lambda e^{-\lambda|x|} dx \\ &= -2\int_0^{+\infty} \frac{1}{2}\lambda e^{-\lambda x} \ln \frac{1}{2}\lambda e^{-\lambda x} dx \\ &= -\int_0^{+\infty} \lambda e^{-\lambda x} \ln \lambda e^{-\lambda x} dx + \ln 2 \int_0^\infty \lambda e^{-\lambda x} dx \\ &= \ln \frac{e}{\lambda} - \ln 2 \cdot e^{-\lambda x}\Big|_0^\infty \\ &= \ln \frac{2e}{\lambda} \end{aligned} \tag{113}$$

3. By condition we know that
$$X_1, X_2 \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \tag{114}$$
Since $X_1 + X_2 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, it follows that $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{(x-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}} \\ h(X) &= -\int f(x) \log f(x) dx \\ &= -\int f(x) \log \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} + f(x) \left(-\frac{(x-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)} \right) dx \end{aligned} \tag{115}$$
By property of normal distribution, we have
$$\int f(x) dx = 1 \text{ and } \int (x - (\mu_1 + \mu_2))^2 f(x) dx = \sigma_1^2 + \sigma_2^2 \tag{116}$$
Hence,
$$h(X) = \frac{1}{2} \log 2\pi(\sigma_1^2 + \sigma_2^2) + \frac{1}{2} \tag{117}$$

□

Exercise 62 (Concavity of determinants) Let K_1 and K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove:
$$\left|\lambda K_1 + \bar{\lambda} K_2\right| \geq |K_1|^{\lambda} |K_2|^{\bar{\lambda}} \quad \text{for } 0 \leq \lambda \leq 1, \quad \bar{\lambda} = 1 - \lambda$$
where $|K|$ denotes the determinant of K. [Hint: Let $\mathbf{Z} = \mathbf{X}_\theta$ where $\mathbf{X}_1 \sim N(0, K_1), \mathbf{X}_2 \sim N(0, K_2)$ and $\theta =$ Bernoulli (λ). Then use $h(\mathbf{Z}|\theta) \leq h(\mathbf{Z})$.]

Proof. Let $Z = \theta X_1 + (1-\theta)X_2$, where θ, X_1 and X_2 are independent, $\mathbf{X}_1 \sim N(0, K_1), \mathbf{X}_2 \sim N(0, K_2)$ and $\theta =$ Bernoulli (λ). Note that for every entry in the covariance matrix of Z,

$$\begin{aligned} cov(Z_i, Z_j) &= E(Z_i Z_j) E((\theta X_{1i} + \bar{\theta} X_{2i})(\theta X_{1j} + \bar{\theta} X_{2j})) \\ &= E(\theta^2) E X_{1i} X_{1j} + E(\bar{\theta}^2) E X_{2i} X_{2j} \\ &= \lambda cov(X_{1i}, X_{1j}) + \bar{\lambda} cov(X_{2i}, X_{2j}) \end{aligned} \tag{118}$$

Hence we have $K_Z = \lambda K_{X_1} + \bar{\lambda} K_{X_2}$. Also note that $EZ = \mathbf{0}$. By Theorem 8.6.5 [Cover] we have that

$$h(Z) \leq \frac{1}{2} \log(2\pi e)^n |K_Z| = \frac{1}{2} \log(2\pi e)^n \left|\lambda K_1 + \bar{\lambda} K_2\right| \tag{119}$$

By the formula for entropy of a multivariate normal distribution, we have that

$$\begin{aligned} h(X_1) &\leq \frac{1}{2} \log(2\pi e)^n |K_1| \\ h(X_2) &\leq \frac{1}{2} \log(2\pi e)^n |K_2| \\ \Rightarrow h(Z|\theta) &= \lambda h(X_1) + \bar{\lambda} h(X_2) \\ &= \frac{\lambda}{2} \log(2\pi e)^n |K_1| + \frac{\bar{\lambda}}{2} \log(2\pi e)^n |K_2| \\ &= \frac{1}{2} \log(2\pi e)^n |K_1|^{\lambda} |K_2|^{\bar{\lambda}} \end{aligned} \tag{120}$$

Note that $h(Z|\theta) \leq h(Z)$. The result follows from Equation 119 and 120. □

Exercise 63 (Uniformly distributed noise) Let the input random variable X to a channel be uniformly distributed over the interval $-\frac{1}{2} \leq x \leq +\frac{1}{2}$ Let the output of the channel be $Y = X + Z$, where the noise random variable is uniformly distributed over the interval $-a/2 \leq z \leq +a/2$

1. Find $I(X;Y)$ as a function of a
2. For $a = 1$ find the capacity of the channel when the input X is peak-limited; that is, the range of X is limited to $-\frac{1}{2} \leq x \leq +\frac{1}{2}$. What probability distribution on X maximizes the mutual information $I(X;Y)$?
3. (Optional) Find the capacity of the channel for all values of a, again assuming that the range of X is limited to $-\frac{1}{2} \leq x \leq +\frac{1}{2}$

Solution.

1. We first calculate the distribution of Y.

$$p_Y(y) = \int_{-\frac{a}{2}}^{\frac{a}{2}} \frac{1}{a} \mathbf{1}_{\{y - \frac{1}{2} \leq z \leq y + \frac{1}{2}\}} dz \tag{121}$$

If $a \leq 1$, we have that

$$p_Y(y) = \begin{cases} \frac{1}{a}\left(y + \frac{a+1}{2}\right) & -\frac{a+1}{2} \leq y \leq \frac{a-1}{2} \\ 1 & \frac{a-1}{2} < y \leq \frac{1-a}{2} \\ \frac{1}{a}\left(-y + \frac{a+1}{2}\right) & \frac{1-a}{2} < y \leq \frac{a+1}{2} \end{cases} \tag{122}$$

If $a > 1$, we have that

$$p_Y(y) = \begin{cases} \frac{1}{a}\left(y + \frac{a+1}{2}\right) & -\frac{a+1}{2} \leq y \leq \frac{1-a}{2} \\ \frac{1}{a} & \frac{1-a}{2} < y \leq \frac{a-1}{2} \\ \frac{1}{a}\left(-y + \frac{a+1}{2}\right) & \frac{a-1}{2} < y \leq \frac{a+1}{2} \end{cases} \tag{123}$$

For $a \le 1$,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(Z) = h(Y) - \ln a \\ &= \int_{-\frac{a+1}{2}}^{\frac{a+1}{2}} p(y) \ln p(y) dy - \ln a \\ &= -2 \int_0^{\frac{1-a}{2}} 1 \ln 1 dy - 2 \int_{\frac{1-a}{2}}^{\frac{1+a}{2}} \frac{1}{a}(-y + \frac{a+1}{2}) \log \frac{1}{a}(-y + \frac{a+1}{2}) dy - \ln a \\ &= 0 - 2a \int_0^1 t \ln t dt - \ln a \quad (t \triangleq \frac{1}{a}\left(-y + \frac{a+1}{2}\right) \\ &= -2a \left(\frac{1}{2} t^2 \ln t - \frac{1}{4} t^2 \right) \Big|_0^1 - \ln a \\ &= \frac{a}{2} - \ln a \end{aligned} \tag{124}$$

For $a > 1$,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(Z) = h(Y) - \ln a \\ &= \int_{-\frac{a+1}{2}}^{\frac{a+1}{2}} p(y) \ln p(y) dy - \ln a \\ &= -2 \int_0^{\frac{a-1}{2}} \frac{1}{a} \ln \frac{1}{a} dy - 2 \int_{\frac{a-1}{2}}^{\frac{1+a}{2}} \frac{1}{a}(-y + \frac{a+1}{2}) \log \frac{1}{a}(-y + \frac{a+1}{2}) dy - \ln a \\ &= \frac{a-1}{a} \ln a - 2a \int_0^{\frac{1}{a}} t \ln t dt - \ln a \quad (t \triangleq \frac{1}{a}\left(-y + \frac{a+1}{2}\right) \\ &= -\frac{1}{a} \ln a - 2a \left(\frac{1}{2} t^2 \ln t - \frac{1}{4} t^2 \right) \Big|_0^{\frac{1}{a}} \\ &= \frac{1}{2a} \end{aligned} \tag{125}$$

2. Since X and Z are both limited to $\left[-\frac{1}{2}, \frac{1}{2}\right]$, Y is limited to $[-1, 1]$. By the Example 12.2.4 in [Cover], the maximal differential entropy $h(Y)$ is $\ln 2$, which can be obtained when Y is uniformly distributed, and thus $p(X = -\frac{1}{2}) = p(X = \frac{1}{2}) = \frac{1}{2}$.

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z) = h(Y) - \ln 1 \le \ln 2 \text{ nats} \tag{126}$$

   The capacity is $\ln 2$ nats $= 1$ bit.

3. If $a = \frac{1}{k}, k \in \mathbf{N}$, then we can construct X to be uniformly distributed on $\left\{-\frac{1}{2}, -\frac{1}{2} + \frac{1}{k}, \ldots, \frac{1}{2}\right\}$, with Y uniformly distributed on $\left[-\frac{1}{2} - \frac{1}{2k}, -\frac{1}{2} + \frac{1}{2k}\right]$. In this case, the maximal mutual information can be obtained. $C = \log\left(1 + \frac{1}{k}\right)$.

□

Exercise 64 (Channel with uniformly distributed noise) Consider a additive channel whose input alphabet $\mathcal{X} = \{0, \pm 1, \pm 2\}$ and whose output $Y = X + Z$, where Z is distributed uniformly over the interval $[-1, 1]$. Thus, the input of the channel is a discrete random variable, whereas the output is continuous. Calculate the capacity $C = \max_{p(x)} I(X;Y)$ of this channel.

Solution. Note that

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z) = h(Y) - \log 2 \tag{127}$$

Note that the support set of Y is $[-3, 3]$.By the Example 12.2.4 in [Cover], the maximal differential entropy $h(Y)$ is $\log 6$, which can be obtained when $p(X = -2) = p(X = 0) = p(X = 2) = \frac{1}{3}$, and Y is uniformly distributed. The maximal mutual information is $\log 3$. □

Lemma 1 (Conditional Expectation of Two Normal Random Variables)**.** *Let X and Y be jointly Gaussian with variances σ_1^2 ,σ_2^2 and correlation coefficient ρ. We have that $E(X|Y) = \frac{\sigma_1 \rho}{\sigma_2} Y$.*

Proof. We calculate the conditional probability density of $p(x|y)$

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p(y)} \\ &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right]\right\}}{\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\frac{y^2}{\sigma_2^2}\right\}} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{\rho^2 y^2}{\sigma_2^2}\right] - \frac{1}{2}\frac{y^2}{\sigma_2^2} + \frac{1}{2}\frac{y^2}{\sigma_2^2}\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left(x - \frac{\sigma_1\rho y}{\sigma_2}\right)\right\} \end{aligned} \tag{128}$$

With Y given, the mean value is $\frac{\sigma_1\rho}{\sigma_2}Y$ □

Exercise 65 (Gaussian mutual information) Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$

Solution. By the formula of mutual information between correlated Gaussian random variables, we have

$$I(X;Z) = -\frac{1}{2}\log(1-\rho_{xz}^2) \tag{129}$$

With the lemma above, we can derive ρ_{xz} as follows.

$$\begin{aligned} \rho_{xz} &= \frac{\mathrm{E}\{XZ\}}{\sigma_x\sigma_z} \\ &= \frac{\mathrm{E}\{\mathrm{E}\{XZ|Y\}\}}{\sigma_x\sigma_z} && \text{Nested Expectation} \\ &= \frac{\mathrm{E}\{\mathrm{E}\{X|Y\}\mathrm{E}\{Z|Y\}\}}{\sigma_x\sigma_z} && \text{Markov Chains} \\ &= \frac{\mathrm{E}\left\{\left(\frac{\sigma_x\rho_{xy}}{\sigma_y}Y\right)\left(\frac{\sigma_z\rho_{zx}}{\sigma_y}Y\right)\right\}}{\sigma_x\sigma_z} && \text{Apply the Lemma} \\ &= \rho_{xy}\rho_{zy} \end{aligned} \tag{130}$$

Hence $I(X;Z) = -\frac{1}{2}\log(1-\rho_{xy}^2\rho_{zy}^2)$ □

10 Gaussian Channel

Exercise 66 (Channel with two independent looks at Y) Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X

1. Show that $I(X; Y_1, Y_2) = 2I(X; Y_1) - I(Y_1; Y_2)$
2. Conclude that the capacity of the channel $X \mapsto Y_1, Y_2$ is less than twice the capacity of the channel $X \mapsto Y_1$.

Proof.

1.

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) \\ &= H(Y_1,Y_2) - H(Y_1|X) - H(Y_2|X) \\ &= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) \\ &= I(X;Y_1) + I(X;Y_2) - I(Y_1,Y_2) \\ &= 2I(X;Y_1) - I(Y_1,Y_2) \end{aligned} \tag{131}$$

2.

$$\begin{aligned} C_1 &= \max_{p(x)} I(X;Y_1,Y_2) \\ &= \max_{p(x)} (2I(X;Y_1) - I(Y_1,Y_2)) \\ &\leq \max_{p(x)} 2I(X;Y_1) \\ &= 2C_2 \end{aligned} \tag{132}$$

□

Exercise 67 (Two-look Gaussian channel) Given $X \mapsto Y_1, Y_2$. Consider the ordinary Gaussian channel with two correlated looks at X, that is, $Y = (Y_1, Y_2)$, where

$$\begin{aligned} Y_1 &= X + Z_1 \\ Y_2 &= X + Z_2 \end{aligned}$$

with a power constraint P on X, and $(Z_1, Z_2) \sim \mathcal{N}_2(0, K)$, where

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}$$

Find the capacity C for

1. $\rho = 1$
2. $\rho = 0$
3. $\rho = -1$

Solution. From Theorem 8.6.5 [Cover] we know that the Gaussian distribution maximizes the entropy over all distributions with the same variance. Hence it is clear that normally distributed $X \sim \mathcal{N}(0, P)$ will maximize the mutual information. In this case $(Y_1, Y_2) \sim \left(0, \begin{bmatrix} P+N & P+\rho N \\ P+\rho N & P+N \end{bmatrix}\right)$

$$\begin{aligned} \max I(X;Y_1,Y_2) &= h(Y_1,Y_2) - h(Y_1,Y_2|X) \\ &= h(Y_1,Y_2) - h(Z_1,Z_2) \\ &= \frac{1}{2}\log(2\pi e)^2 \begin{vmatrix} P+N & P+\rho N \\ P+\rho N & P+N \end{vmatrix} - \frac{1}{2}\log(2\pi e)^2 \begin{vmatrix} N & N\rho \\ N\rho & N \end{vmatrix} \\ &= \frac{1}{2}\log\left(1 + \frac{2P}{(1+\rho)N}\right) \end{aligned} \tag{133}$$

1. $\rho = 1$, $C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$
2. $\rho = 0$, $C = \frac{1}{2}\log\left(1 + \frac{2P}{N}\right)$
3. $\rho = -1$, $C = +\infty$.

□

Exercise 68 (Output power constraint) Consider an additive white Gaussian noise channel with an expected output power constraint P. Thus, $Y = X + Z, Z \sim N\left(0, \sigma^2\right), Z$ is independent of X, and $EY^2 \leq P$ Find the channel capacity.

Solution.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z) \\ &= h(Y) - \frac{1}{2}\log\left(2\pi e\sigma^2\right) \\ &\leq \frac{1}{2}\log\left(2\pi eP\right) - \frac{1}{2}\log\left(2\pi e\sigma^2\right) \\ &= \frac{1}{2}\log\frac{P}{\sigma^2} \end{aligned} \tag{134}$$

The equailty holds when Y is normally distributed. In this case $X \sim \mathcal{N}(0, P - \sigma^2)$. □

Exercise 69 (Exponential noise channels) $Y_i = X_i + Z_i$, where Z_i is i.i.d. exponentially distributed noise with mean μ. Assume that we have a mean constraint on the signal (i.e., $EX_i \leq \lambda$). Show that the capacity of such a channel is $C = \log\left(1 + \frac{\lambda}{\mu}\right)$

Proof.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z) \\ &= h(Y) - \sum_i h(Z_i) \\ &\leq \sum_i \left(h(Y_i) - h(Z_i)\right) \end{aligned} \tag{135}$$

The equailty holds when Y_is are independent, which can be obtained if X_is are independent. Hence we can only consider the channel for the input and output to be single-valued. Still, $I(X;Y) = h(Y) - h(Z)$ holds.

Note for exponentially distributed Z,

$$\begin{aligned} h(Z) &= -\int_0^{+\infty} g(z)\ln\frac{1}{\mu}e^{-\frac{z}{\mu}}dz \\ &= -\int_0^{+\infty} g(z)\ln\frac{1}{\mu}dz - \int_0^{+\infty} g(z)\frac{z}{\mu}dz \\ &= 1 + \ln\mu \end{aligned} \tag{136}$$

Note that $EY = EX + EZ \leq \lambda + \mu$. For mean-value bounded Y, by Theorem 12.1.1 and Example 12.2.5 [Cover], the maximizing differential entropy is $h^*(Y) = 1 + \ln(\lambda + \mu)$, with distribution $p^*(y) = \frac{1}{\lambda+\mu}e^{-\frac{y}{\lambda+\mu}}$. Therefore

$$I(X;Y) \leq \sum_{i=1}^{n}\left((1 + \ln(\lambda + \mu)) - (1 + \ln(\mu))\right) = n\ln\frac{\lambda + \mu}{\mu} \tag{137}$$

The equailty holds when X_i are independent with mean value λ and $Y_i \sim \exp\left(\frac{1}{\lambda+\mu}\right)$. We need to find such distribution for X_i. Since X_i and Z_i are independent and $Y_i = X_i + Z_i$, it follows that the characteristc functions hold the following relation.

$$\phi_Y(t) = \phi_X(t) \cdot \phi_Z(t) \tag{138}$$

Therefore

$$\begin{aligned}\phi_X(t) &= \frac{\phi_Y(t)}{\phi_Z(t)} \\ &= \frac{(1-i(\lambda+\mu)t)^{-1}}{(1-i\mu t)^{-1}} \\ &= \frac{1}{\lambda+\mu}\frac{[\mu - i\mu(\lambda+\mu)t]+\lambda}{1-i(\lambda+\mu)t} \\ &= \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}(1-i(\lambda+\mu)t)^{-1}\end{aligned} \tag{139}$$

The characteristc function above is a linear combination of two kinds of distribution. We can set every X_i to be 0 with the probability of $\frac{\mu}{\lambda+\mu}$, and to be exponentially distributed with mean value $\lambda+\mu$ by the probability of $\frac{\lambda}{\mu+\lambda}$. Then the channel capacity $n \ln \frac{\lambda+\mu}{\mu}$ can be obtained.

□