EL-GY 6063: Information Theory
Lecture 8

April 17, 2019

Extra Class on Tuesday, May 1, 6:00PM-8:30PM. The location will be announced on course website.

1 The Channel Coding Setup

In the following we describe the four components of the channel coding setup starting with information source, the encoder, channel and finally the decoder.

1. **Information Source (Message):** At the very beginning of the communication system there is a source of information. This is the source which produces the information to be sent to the receiver. It is generally assumed that the source produces a string of independent and identically distributed binary random variables $Z_1, Z_2, \cdots, Z_k$ which are distributed according to the Bernoulli distribution with parameter $\frac{1}{2}$, where k is a very large number. This is called the message. We often write $W = Z^k = (Z_1, Z_2, \cdots, Z_k)$ to keep in line with the textbook's notation.

   Definition 1. *In the channel coding problem, for a fixed natural number $k \in \mathbb{N}$, the random vector $W = (Z_1, Z_2, \cdots, Z_k)$ is called the message, where $Z_i, i \in [1, k]$ are a vector of independent and identically distributed Bernoulli variables with parameter $\frac{1}{2}$. Sometimes, instead of W, the message is represented by the variable M which is a random variable distributed uniformly over the alphabet $\{1, 2, \cdots, 2^k\}$.*

2. **Encoder:** The next element in the channel coding setup is called the channel encoder or transmitter. The channel encoder takes the message W as the input and outputs an n-length vector X^n of variables which is defined on the alphabet $\mathcal{X}^n$ where n is a natural number. It is generally assumed that X^n is a function of W. The channel encoder prepares the data to be sent over the channel or the communication medium. The elements of X^n are then sent sequentially starting from X_1 and after that X_2 all the way through X_n.

   Definition 2. *Given the natural numbers $n, k \in \mathbb{N}$, and the input alphabet $\mathcal{X}$, the encoder in the channel coding setup is characterized by the encoding function $e : \{0, 1\}^k \to \mathcal{X}^n$. The encoding function takes the message W as the input and produces $e(W) = e(Z^k)$ as its output. We sometimes write X^n or $e(M)$ instead of $e(W)$ to represent the output of the encoder. Furthermore, the natural number n is called the blocklength of e.*

   Example 1. *Let $k = 1$, $n = 2$, and $\mathcal{X} = \{0, 1\}$ be the binary alphabet. Then, the function $e : \{0, 1\} \to \{0, 1\}^2$ with the encoding rule $e(0) = 00$ and $e(1) = 11$ is an encoding function. This is called a repetition code (repetition encoder) of length two.*

3. **Channel:** The third component of the channel coding setup is the channel itself. The channel takes the input X^n from the encoder and produces the output Y^n defined on the alphabet $\mathcal{Y}^n$. It is assumed that the channel has a random or stochastic behavior. This is characterized by a conditional probability distribution $P_{Y^n|X^n}$. Sometimes, it is assumed that the channel takes the ith element of the input X_i

and produces the output Y_i randomly based on the transition probability matrix $P_{Y_i|X_i} = P_{Y|X}$. In other words it is assumed that

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).$$

That is the channel produces each output symbol independently of the previous ones and only dependent on the current channel input. This is called a stationary and memoryless channel. Stationarity means that $P_{Y^n|X^n} = P_{Y_{t+1}^{n+t}|X_{t+1}^{n+t}}$ for any $t \in \mathbb{N}$, where $X_{t+1}^{n+t} = (X_{t+1}, X_{t+2}, \cdots, X_{n+t})$, and memoryless property means that the output Y_i is independent of $X_1, X_2, \cdots, X_{i-1}$ given X_i for any $i \in \mathbb{N}$. In other words, the Markov chain $Y_i \leftrightarrow X_i \leftrightarrow X_1, X_2, \cdots, X_{i-1}$ holds for all $i \in \mathbb{N}$.

Definition 3. *Given the input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, a stationary and memoryless channel (without feedback) is characterized by the transition probability matrix $P_{Y|X}$.*

Exercise. Verify that the channel described above is truly stationary and memoryless.

In this course we only consider stationary and memoryless channels.

Example. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Assume that N is a Bernoulli random variable with parameter $p \in [0, 1]$ which is independent of X. Then, $Y = X \oplus_2 N$ characterizes a channel with transition probability matrix:

Table 1: $P_{Y|X}$

$Y \backslash X$	0	1
0	1-p	p
1	p	1-p

.

This is called the binary symmetric channel with parameter p and is denoted by BSC(p). You have seen in your digital communications course that the BSC(p) appears naturally in many modulation/demodulation setups.
(draw)

Example. Let $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, e\}$. The symbol e is called the erasure symbol. Let E be a Bernoulli random variable with parameter $\epsilon \in [0, 1]$. Define the relation between Y and X by:

$$Y = \begin{cases} X & \text{if } E = 0 \\ e & \text{if } E = 1. \end{cases}$$

Then, the channel is characterized by the transition probability matrix:

Table 2: $P_{Y|X}$

$Y \backslash X$	0	1
0	1-e	0
1	0	1-e
e	e	e

.

This is called the binary erasure channel with parameter e.

4. **Decoder:** The fourth and final component of the channel is the channel decoder or receiver. It takes the output of the channel Y^n and produces the reconstruction $\hat{Z}_1, \hat{Z}_2, \cdots, \hat{Z}_k$. We often write $\hat{W} = (\hat{Z}_1, \hat{Z}_2, \cdots, \hat{Z}_k)$. $\hat{W}$ is a function of Y^n.

Definition 4. *Given the natural numbers $n, k \in \mathbb{N}$, and the output alphabet $\mathcal{Y}$, the decoder in the channel coding setup is characterized by the decoding function $f : \mathcal{Y}^n \to \{0, 1\}^k$. The decoding function* *takes the channel output Y^n as the input and produces $f(Y^n) = \hat{Z}^k$ as its output. We sometimes write $\hat{W}$ or $\hat{M}$ instead of $f(Y^n)$ to represent the output of the decoder. Furthermore, the natural number n is called the blocklength of d.*

Example 2. *Let $k = 1$, $n = 2$, and $\mathcal{Y} = \{0, 1\}$ be the binary alphabet. Then, the function $f : \{0, 1\}^2 \rightarrow \{0, 1\}$ with the decoding rule $e(00) = e(01) = 0$ and $e(11) = e(10) = 1$ is a decoding function. This is called the majority rule decoder of length 2.*

Generally, the input alphabet $\mathcal{X}$, the channel transition probability $P_{Y|X}$ and the output alphabet $\mathcal{Y}$ are fixed and are not in the control of the communication system designer. These parameters characterize the channel coding problem.

Definition 5. *A (stationary and memoryless) channel coding problem is characterized by the triple $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, where $\mathcal{X}$ is called the input alphabet, $\mathcal{Y}$ is called the output alphabet and the conditional distribution $P_{Y|X}$ is called the channel transition probability matrix.*

On the other hand, the encoding and decoding functions used at the channel encoder and decoder are to be designed by the communication system designer.

Definition 6. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ and natural numbers $n, k \in \mathbb{N}$, ab (n, k)-coding strategy is characterized by a pair of functions $e : \{0, 1\}^k \rightarrow \mathcal{X}^n$ and $f : \mathcal{Y}^n \rightarrow \{0, 1\}^k$ are called the encoding and decoding functions, respectively. The natural number n is called the blocklength of the coding strategy.*

Definition 7. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ and natural numbers $n, k \in \mathbb{N}$, and the coding strategy (e,f), the probability of error is defined as*

$$P_e = P(\mathcal{E}) = P(\hat{Z}^k \neq Z^k) = P(f(Y^n) \neq Z^k) = P(f(Y^n) \neq \hat{W}).$$

Exercise. For the channel coding problem characterized by $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and BSC(p), and natural numbers $k = 1$ and $n \in \mathbb{N}$, consider the coding strategy (e, f) where the encoding function e is the repetition code of length n and the decoding function f is the majority rule decoder of length n. Prove that $\lim_{n\rightarrow\infty} P_e = 0$.

Definition 8. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ and natural numbers $n, k \in \mathbb{N}$, and coding strategy (e, f), where $e : \{0, 1\}^k \rightarrow \mathcal{X}^n$ and $f : \mathcal{Y}^n \rightarrow \{0, 1\}^k$, the rate of the coding strategy is defined as $r = \frac{k}{n}$. Alternatively, $r = \frac{1}{n} \log |\mathcal{W}| = \frac{1}{n} \log |\mathcal{M}|$.*

The rate can be interpreted as the average number of bits of information transmitted per channel use.

This apparent tradeoff is formalized below:

Definition 9. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$. The minimal (n, k)-achievable error $P_{n,k}$ is the minimum value for which there exists an (n, k)-coding strategy (e, f) with probability of error $P_{n,k}$. Alternatively,*

$$P_{n,k} = \min_{\substack{e:\{0,1\}^k\rightarrow\mathcal{X}^n \\ f:\mathcal{Y}^n\rightarrow\{0,1\}^k}} P_e.$$

Exercise. Argue that the minimum in the above definition always exists.

Exercise. Prove that $P_{2n,2k} \leq 2P_{n,k} - P_{n,k}^2$.

Definition 10. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ and natural number n, a rate R is called (n, P)-achievable, if there exists a natural number k such that $R \leq \frac{k}{n}$ and $P_e \leq P_{n,k}$.*

The above definition can be interpreted as follows. The rate R is (n, P)-achievable if there exists a coding strategy with blocklength n for which the average information per symbol is greater than R and the probability of error is less than P.

The n-length optimal rate function is defined next:

Definition 11. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ and natural number n, the n-length optimal rate function is the function $C_n(P) : [0,1] \to \mathbb{R}$, where*

$$C_n(P) = max_{(n,P)\text{-}achievable} R,$$

where the maximum is taken over the set of all (n, P)-achievable rates.

The n-length optimal rate function takes the desired probability of error as its input and outputs the maximum achievable rate using coding strategies with blocklength n and error probability P.

Example 3. *Show that for any natural number $n \in \mathbb{N}$ and real number $P \in [0,1]$, the inequality $C_n(P) \leq C_{2n}(2P - P^2)$ holds.*

Solution. Remember that $C_n(P)$ is the maximum rate of the coding strategies with blocklength n and error probability at most P. Assume that this maximum is achieved by the coding strategy (e_n, f_n), where $e_n : \{0,1\}^k \to \mathcal{X}^n$ and $f_n : \mathcal{Y}^n \to \{0,1\}^k$. In other words, the coding strategy (e_n, f_n) is such that $\frac{k}{n} = C_n(P)$ and $P(f_n(Y^n) \neq Z^k) \leq P$.

Define the coding strategy (e_{2n}, f_{2n}) as the concatenation of (e_n, f_n) with itself. That is, $e_{2n} : \{0,1\}^{2k} \to \mathcal{X}^{2n}$ and $f_{2n} : \mathcal{Y}^{2n} \to \{0,1\}^{2k}$, where $e_{2n}(Z^{2k}) = (e_n(Z_1, Z_2, \cdots, Z_k), e_n(Z_{k+1}, Z_{k+2}, \cdots, Z_{2k}))$ and similarly $f_{2n}(Y^{2n}) = (f_n(Y_1, Y_2, \cdots, Y_n), f_n(Y_{n+1}, Y_{n+2}, \cdots, Y_{2n}))$. Then, clearly, the rate of this coding strategy is $\frac{2k}{2n} = \frac{k}{n} = C_n(P)$. Also, we have

$$\begin{aligned} P(f_{2n}(Y^{2n}) \neq Z^{2k}) &= P((f_n(Y^b), f_n(Y_{n+1}^{2n})) \neq (Z^k, Z_{k+1}^{2}k)) \\ &= P(f_n(Y^n) \neq Z^k \text{ } or \text{ } f_n(Y_{n+1}^{2n}) \neq Z_{k+1}^{2k}) \\ &= P(f_n(Y^n) \neq Z^k) + P(f_n(Y_{n+1}^{2n}) \neq Z_{k+1}^{2k})) - P(f_n(Y^n) \neq Z^k \text{ } and \text{ } f_n(Y_{n+1}^{2n}) \neq Z_{k+1}^{2k}) \\ &= P(f_n(Y^n) \neq Z^k) + P(f_n(Y_{n+1}^{2n}) \neq Z_{k+1}^{2k})) - P(f_n(Y^n) \neq Z^k)P(f_n(Y_{n+1}^{2n}) \neq Z_{k+1}^{2k}) \\ &= 2P - P^2. \end{aligned}$$

Exercise. Show that for any given $m, n \in \mathbb{N}$, we have $nC_n(P) + mC_m(P) \leq (n+m)C_{n+m}(2P - P^2)$. Also, find a channel and suitable n such that $C_n(P) > C_{2n}(P)$. Find a channel and suitable n such that $C_n(P) > C_{n+1}(2P - P^2)$.

The property that for any given $m, n \in \mathbb{N}$, we have $nC_n(P) + mC_m(P) \leq (n+m)C_{n+m}(2P - P^2)$ is called sub-additivity. It means that although $C_n(P)$ may decrease in n at some points, it is increasing in the long term. Also, $C_n(P)$ is bounded from above since the maximum amount of information per channel use cannot exceed $min(\log|\mathcal{X}|, \log|\mathcal{Y}|)$ (why?). These two facts show that $\limsup_{n\to\infty} C_n(P)$ exists.

Exercise. Show that the function $C_n(P)$ is non-decreasing in P.

The above statement means that if there is a coding strategy with rate $R = C_n(P)$ and probability of error P, then for any $P' > P$, there is a coding strategy with the same rate R and probability of error P'.

As stated before, we are mostly interested in the case when $P \approx 0$. Let P_n be a sequence of error probabilities for which $P_n \to 0$ as $n \to \infty$. We are interested in finding $\limsup_{n\to\infty} C_n(P_n)$ that is the maximum rate achievable with probability of error P_n as $n \to \infty$.

Definition 12. *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, the rate $R \geq 0$ is said to be achievable if there exists a sequence of coding strategies (e_n, f_n) with blocklength n and rates R_n and error probabilities P_n such that:*
i) $P_n \to 0$ as $n \to \infty$.
ii) $\limsup_{n\to\infty} R_n \geq R$.
The channel capacity is defined as the supremum of all achievable rates.

2 The Probabilistic Approach

Sometimes, it is difficult to provide an algorithm or a constructive method to solve the problems such as finding the channel capacity described above. In these cases, the probabilistic approach turns out to be an extremely powerful tool.

Example 4. *10% of the surface of a sphere is colored blue, the rest is colored red. Show that irrespective of the manner in which the colors are distributed on the sphere, it is always possible to inscribe a cube in the sphere such that all of the eight vertices of the cube will lie in the red region.*

There are two ways to answer this problem, either one must come up with an algorithm to inscribe the cube in a suitable way or use the probabilistic approach as follows. To solve the problem in the probabilistic method we first define a random experiment. Assume that one point is chosen randomly and uniformly over the sphere and the first vertex of the cube is put on this point and the rest of the vertices are placed in the clockwise direction. Given this random experiment, one can ask what is the probability that all vertices are on the red paint. Let the event $\mathcal{E}_i, i \in \{1, 2, \cdots, 8\}$ be the event that the ith vertex is on red paint. Then,

$$P(\cap_{i=1}^{8} \mathcal{E}_i) = 1 - P(\cup_{i=1}^{8} \mathcal{E}_i^c) \geq 1 - \sum_{i=1}^{8} P(\mathcal{E}_i) = 1 - \sum_{i=1}^{8} \frac{1}{10} = \frac{1}{5}.$$

By a similar argument as in the previous example, it is always possible to fit the cube such that all vertices are on the red region.