EL-GY 6063: Information Theory
Lecture 9

April 13, 2018

1 Summary of previous lecture

Definition 1. *Given an (n,k)-coding strategy (e,f), the image of the encoding function $\mathcal{C} = \{e(z^k)|z^k \in \{0,1\}^k\}$ is called the codebook corresponding to e.*

Example. Let e be the repetition encoder of order two, then $\mathcal{C} = \{00, 11\}$.

Generally the codebook is written as a matrix. The matrix has 2^k rows and n columns. Each row is called a codeword. The ith row is equal to $e(i)$, the image of the ith binary sequence. We define $\underline{c}_i = e(i)$ as the ith codeword. Then, $c_{i,j}, j \in \{1, 2, \cdots, n\}$ is the jth element of this n-length codeword.

Example. Let e be the repetition encoder of order two, then we write $\mathcal{C} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$. Also, $\underline{c}_1 = e(0) = 00$ and $\underline{c}_2 = e(1) = 11$. We write $c_{1,1} = 0$, $c_{1,2} = 0$, $c_{2,1} = 1$ and $c_{1,2} = 1$.

Example 1. *A typicality encoder $e : \{0,1\}^k \to \mathcal{X}^n$ is characterized by parameters (n, P_X, ϵ), where n is the blocklength, P_X is a distribution on the channel input alphabet and ϵ is a (small) positive real number. The encoding function has the property that its set of possible outputs is a subset of the typical set $\mathcal{T}_\epsilon^n(X)$ for a given ϵ and n. As a reminder,*

$$\mathcal{T}_\epsilon^n(X) = \{x^n | \big| \frac{1}{n} N(a|x^n) - P_X(a) \big| \leq \epsilon, \forall a \in \mathcal{X}\}.$$

As an example, assume that the source X is binary and let P_X be the binary symmetric distribution and $n = 100$ and $\epsilon = 0.03$. Then,

$$\mathcal{T}_{0.03}^1 00(X) = \{x^{100} | \big| \frac{1}{100} N(1|x^{100}) - \frac{1}{2} \big| \leq 0.03\} = \{x^{100} | 47 \leq N(1|x^n) \leq 53\}.$$

So, the typicality encoder has outputs which are binary sequences of length 100 and have between 47 to 53 ones. Note that the encoder does not output all of the typical set elements, rather its outputs are a subset of the typical set.

For the typicality encoder, the rows of the codebook matrix are all typical vectors.

Lemma 1. *Assume that a coding strategy uses the typicality encoder with parameters (n, P_X, ϵ). Then, the output of the channel Y^n is jointly typical with the input $X^n = e(Z^k)$ with probability approaching one as $n \to \infty$. More precisely, $\lim_{n\to\infty} P((X^n, Y^n) \in \mathcal{A}_{2\epsilon}^n(X,Y)) = 1$.*

proof. The result follows from lemma 1 in lecture 5 which show that the probability of $P(Y^n \in \mathcal{T}_{2\epsilon}^n(Y|x^n))$ approaches 1 as $n \to \infty$.

Example 2. *The typicality decoder $f : \mathcal{Y}^n \to \{0,1\}^k$ operates based on the above lemma. It searches over all messages Z^k and finds a message for which $(e(Z^k), Y^n) \in \mathcal{T}_{2\epsilon}^n(X,Y)$. If there is exactly one such message it declares that message as the reconstruction, otherwise it declares an error in decoding.*

Example 3. *Assume that a typicality encoder with blocklength 100 and parameter ϵ is used to transmit data through a binary symmetric channel with parameter* 0.1*. Assume that the encoder transmits the channel input $X^{100} = 00\cdots011\cdots1$ the vector which has 50 zeros at the beginning and 50 ones at the end. Note that for the binary symmetric channel with paramater 0.1, we have $P_{Y|X}(0|0) = P_{Y|X}(1|0) = 0.9$ and $P_{Y|X}(1|0) = P_{X|Y}(0|1) = 0.1$. So, the output Y^{100} has roughly 5 bit flips in the first 50 bits and 5 more flips in the next 50 bits. That is, it has roughly* 45 *zeros and 5 ones in the first fifty bits and* 45 *ones and* 5 *zeros in the next 50 bits. In this case, $(X^{100}, Y^{100}) \in \mathcal{T}_{2\epsilon}^{100}(X,Y)$.*

Theorem 1 (Shannon's Channel Coding Theorem)**.** *For the channel coding problem characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, the channel capacity is given as:*

$$C = max_{P_X} I(X;Y).$$

1.1 Proof of Achievability

We first prove the achievability statement. That is we prove that for any fixed $R < C$, the rate R is achievable.

Fix the distribution P_X as the distribution which maximizes the capacity expression. Also, fix the blocklength n and error probability 2ϵ. Let $k = \lceil n \cdot R \rceil$ so that $\frac{k-1}{n} \leq R < \frac{k}{n}$. Alternatively, $R \approx \frac{k}{n}$ for large enough n.

Codebook Generation: Construct a codebook containing 2^k codewords each of length n where each codeword is chosen independently based on the distribution $\prod_{i=1}^{n} P_X(x_i)$.

We construct a codebook $\mathcal{C}$ containing 2^k codewords of length n by choosing each entry $C_{i,j}, i \in \{1,2,\cdots,2^k\}, j \in \{1,2,\cdots,n\}$ randomly and independently. This results in a random matrix $\mathcal{C}$. The matrix $\mathcal{C}$ is generated by choosing each of its entries independently of other entries and based on the distribution P_X. In other words, each $C_{i,j} \in \mathcal{X}$ is given a value which is chosen based on the distribution P_X and is independent of all other entries of the matrix $\mathcal{C}$.

Example. Assume that $\mathcal{X} = \{0,1\}$ and P_X is the binary symmetric distribution. Then, each entry in the codebook $\mathcal{C}$ is a one with probability $\frac{1}{2}$ and a zero with probability $\frac{1}{2}$ and the value is independent of all other entries. As a result, for large n, each codeword almost has $\frac{n}{2}$ zeros and $\frac{n}{2}$ ones.

Encoding: In order to transmit the message M, the encoder sends $e(M) = \underline{C}_M$, which is the Mth row of the codebook over the channel.

So, the vector $X^n = \underline{C}_M$ is input to the channel. The output Y^n is produced based on the channel transition probability Y^n. From previous lectures, we know that the pair of vectors (X^n, Y^n) are jointly typical with respect to the distribution $P_{X,Y}$ with high probability.

Decoding: The decoder uses typicality decoding to decode the message. That is, having received the channel output Y^n, the decoder searches for a codeword $\underline{C}_{\hat{M}} \in \mathcal{C}$ such that $(\underline{C}_{\hat{M}}, y^n) \in \mathcal{T}_{\epsilon}^{n}(X,Y)$. If a unique codeword exists which is jointly typical with Y^n, then $\hat{M}$ is declared as the reconstruction of the message. Otherwise, if such a codeword does not exist or if there are more than one such codewords, an error is declared.

First, note that the rate of the coding strategy is $\frac{k}{n}$ which is greater than or equal to R. So, the rate constrain is satisfied. We need to investigate the probability of error.

Note that the coding strategy was generated randomly by choosing the random codebook. We calculate the probability of the event that the probability of error of the coding strategy is greater than ϵ.

Without loss of generality, we assume that $M = 1$.

Exercise. Show that $P(\mathcal{E}) = P(\mathcal{E}|M=1) = P(\mathcal{E}|M=i), \forall i \in \{1,2,\cdots,2^k\}$.

Define the error event $\mathcal{E}_0$ as the event that the channel input and output are not jointly typical, that is $\mathcal{E}_0 = \{(X^n, Y^n) \notin \mathcal{T}_{2\epsilon}^{n}(X,Y)\}$. Furthermore, define the error event $\mathcal{E}_i, i \in \{2,3,\cdots,2^k\}$ as the event that the codeword $\underline{C}_i, i \neq 1$ is jointly typical with Y^n. That is $\mathcal{E}_i = \{\exists i \neq 1 : (\underline{C}_i, Y^n) \in \mathcal{T}_{2\epsilon}^{n}(X,Y)\}, i \in \{2,3,\cdots,2^k\}$. Then, it is clear that $\mathcal{E} = \mathcal{E}_0 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cdots \cup \mathcal{E}_{2^k}$. Then, we have:

$$P(\mathcal{E}) = P(\mathcal{E}|M=1) = P(\mathcal{E}_0 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cdots \cup \mathcal{E}_{2^k}|M=1) \overset{(a)}{\leq} P(\mathcal{E}_0|M=1) + \sum_{i=2}^{2^k} P(\mathcal{E}_i|M=1) \qquad (1)$$

where in (a) we have used the union bound. We investigate each of the error events separately. First, note that from the AEP we have:

$$P(\mathcal{E}_0|M=1) = P((X^n, Y^n) \notin \mathcal{T}_{2\epsilon}^n(X)) \leq 2\epsilon.$$

Additionally, we have:

$$\begin{aligned}
P(\mathcal{E}_i|M=1) = P((\underline{C}_i, Y^n) \in \mathcal{T}_{2\epsilon}^n(X,Y)\}|M=1) &= \sum_{(x^n,y^n)\in\mathcal{T}_{2\epsilon}^n(X,Y)} P((\underline{C}_i, Y^n) = (x^n, y^n)) \\
&\overset{(a)}{=} \sum_{(x^n,y^n)\in\mathcal{T}_{2\epsilon}^n(X,Y)} P(\underline{C}_i = x^n)P(Y^n = y^n) \\
&\overset{(b)}{\leq} \sum_{(x^n,y^n)\in\mathcal{T}_{2\epsilon}^n(X,Y)} 2^{-n(H(X)-2\epsilon|\mathcal{X}||\mathcal{Y}|)}2^{-n(H(Y)-2\epsilon|\mathcal{X}||\mathcal{Y}|)} = 2^{-n(H(X)+H(Y)-4\epsilon|\mathcal{X}||\mathcal{Y}|)} \sum_{(x^n,y^n)\in\mathcal{T}_{\epsilon}^n(X,Y)} 1 \\
&= 2^{-n(H(X)+H(Y)-4\epsilon|\mathcal{X}||\mathcal{Y}|)}|(x^n, y^n) \in \mathcal{T}_{\epsilon}^n(X,Y)| \\
&\overset{(b)}{\leq} 2^{-n(H(X)+H(Y)-4\epsilon|\mathcal{X}||\mathcal{Y}|)}2^{n(H(X,Y)2\epsilon)} \\
&= 2^{-n(H(X)+H(Y)-H(X,Y)-6\epsilon|\mathcal{X}||\mathcal{Y}|)} = 2^{-n(I(X;Y)-6\epsilon|\mathcal{X}||\mathcal{Y}|)},
\end{aligned}$$

where in (a) we have used the fact that the codewords are produced by choosing their entries independently. Note that Y^n is produced by the channel after receiving $X^n = \underline{C}_1$. Additionally, $\underline{C}_1$ is independent of $\underline{C}_i, i \neq 1$. So, Y^n is also independent of $\underline{C}_i, i \neq 1$. In (b), we have used the following theorem from previous lectures:

Theorem 2. *The set $\mathcal{T}_{\epsilon}^n(X,Y)$ satisfies the following properties:*
i) $P(\mathcal{T}_{\epsilon}^n(X,Y)) \to 1$ as $n \to \infty$. Alternatively, for any $\epsilon > 0$ and for sufficiently large n, we have $P(\mathcal{T}_{\epsilon}^n(X,Y)) > 1-\epsilon$,
ii) $|\mathcal{T}_{\epsilon}^n(X,Y)| \leq 2^{n(H(X,Y)+\epsilon)}$,
iii)$2^{-n(H(X)+\epsilon|\mathcal{X}|)} \leq P(x^n) \leq 2^{-n(H(X)-\epsilon|\mathcal{X}|)}$ for any $x^n \in \mathcal{T}_{\epsilon}^n(X)$.

Next, we replace the error probabilities in equation (1) with the above bounds:

$$\begin{aligned}
P(\mathcal{E}) \leq P(\mathcal{E}_0) + \sum_{i=2}^{2^k} P(\mathcal{E}_i) &\leq \epsilon + \sum_{i=2}^{2^k} 2^{-n(I(X;Y)-6\epsilon|\mathcal{X}||\mathcal{Y}|)} \\
&= \epsilon + 2^k \cdot 2^{-n(I(X;Y)-6\epsilon)} = \epsilon + 2^{-n(I(X;Y)-\frac{k}{n}-6\epsilon|\mathcal{X}||\mathcal{Y}|)}.
\end{aligned}$$

Note that $2^{-n(I(X;Y)-\frac{k}{n}-6\epsilon|\mathcal{X}||\mathcal{Y}|)}$ goes to 0 as $n \to \infty$ as long as $I(X;Y) - \frac{k}{n} - 6\epsilon|\mathcal{X}||\mathcal{Y}| > 0$. By our assumption, $\frac{k}{n} \approx R < C = I(X;Y)$. So, if $6\epsilon|\mathcal{X}||\mathcal{Y}| < C - R$, then, for large enough n, $2^{-n(I(X;Y)-\frac{k}{n}-6\epsilon|\mathcal{X}||\mathcal{Y}|)} < \epsilon$ and we get:

$$P(\mathcal{E}) < 3\epsilon.$$

Note that we are using the probabilistic method to solve the problem. So far, we have show that:

$$P(\mathcal{E}) \leq 3\epsilon \Rightarrow \sum_{\mathcal{C}} P(\mathcal{C})P(\mathcal{E}|\mathcal{C}) \leq 3\epsilon.$$

That is the weighted average of the probability of error of the coding strategies resulting from $\mathcal{C}$'s is less than or equal to 3ϵ. So, there exists at least one coding strategy for which the probability of error is less than 3ϵ. This concludes the proof.

2 Examples

Example 4 (Binary Symmetric Channel)**.** *Consider the BSC(p) where $p \in [0,1]$. Calculate the channel capacity.*

Solution. Note that for the BSC(p), we have $Y = X \oplus_2 N_p$, where N_p is Bernoulli(p) and independent of X. From the channel coding theorem we know that $C = max_{P_X} I(X;Y)$. So,

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(X \oplus_2 N_p|X) = H(Y) - H(N_p|X) = H(Y) - H(N_p).$$

On the other hand we know that $H(Y)$ is maximized if Y is binary symmetric and we also know that Y is binary symmetric if and only if X is binary symmetric. So,

$$C = max_{P_X} I(X;Y) = max_{P_X} H(Y) - H(p, 1-p) = 1 - H(p, 1-p).$$

Note that capacity is a convex function of $P_{Y|X}$ for a fixed P_X. When $p = 0$, the channel is noiseless and maximum information is transmitted. When $p = 1$ the information can be sent by flipping the output. When $p = \frac{1}{2}$, we have $C = 0$ since the output is completely noisy and independent of the input.

Example 5 (Binary Erasure Channel)**.** *Consider the BEC(ϵ), where $\epsilon \in [0,1]$. Calculate the channel capacity.*

Solution. As a reminder in the BEC we have: Let E be a Bernoulli random variable with parameter $\epsilon \in [0,1]$. The channel rule is given by:

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

So,

$$I(X;Y) = H(X) - H(X|Y) = H(X) - P(Y=1)H(Y|X=1) - P(Y=e)H(X|Y=e) - P(Y=0)H(X|Y=1).$$

Note that if $Y = 1$ then we know that $X = 1$ (i.e. $P(X=1|Y=1) = 1$) so $H(X|Y=1) = 0$. Similarly, $H(X|Y=0) = 0$. So,

$$I(X;Y) = H(X) - P(Y=e)H(X|Y=e) = H(X) - P(Y=e)H(P_{X|Y}(0|e), P_{X|Y}(1|e)).$$

Note that $P_{X|Y}(0|e) = \frac{P_{X,Y}(0,e)}{P(e)} = \frac{P(X_0)\epsilon}{\epsilon} = P(X_0)$. Similarly, $P_{X|Y}(1|e) = P_X(1)$. So,

$$I(X;Y) = H(X) - P_Y(e)H(X) = (1-\epsilon)H(X).$$

Which is maximized when X is binary symmetric. So, $I(X;Y) = 1 - \epsilon$.