Lecture 7: Channel coding theorem for discrete-time continuous memoryless channel

Lectured by Dr. Saif K. Mohammed

Scribed by Mirsad Čirkić

Information Theory for Wireless Communication (ITWC) Spring 2012

Let us first define, for the random sequences $\boldsymbol{X} = [X_1, \ldots, X_n]$ and $\boldsymbol{Y} = [Y_1 \ldots, Y_n]$ and their corresponding sequence realizations $\boldsymbol{x} = [x_1, \ldots, x_n]$ and $\boldsymbol{y} = [y_1 \ldots, y_n]$ where $x_k, y_k \in \mathbb{R}$, the following probability density functions (pdfs): $f_{\boldsymbol{X}}(\boldsymbol{x}) \stackrel{\Delta}{=} \prod_{k=1}^{n} f_X(x_k)$ as the input pdf, $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{n} f_{Y|X}(y_k|x_k)$ as the memoryless channel (transition) pdf, and $f_{\boldsymbol{Y}}(\boldsymbol{y})$ as the output pdf. Each element X_k (or just X without the index k) are independent and identically distributed i.i.d. Further, due to the i.i.d. property of X and the specified memoryless property imposed on the channel transition pdf, we have $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{k=1}^{n} f_Y(y_k)$.

Define the so called and the mutual information quantity

Let

$$I(X;Y) \triangleq \int_{x,y} f_{X,Y}(x,y) \log\left(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\right) dxdy \tag{1}$$

define the mutual information between the random variables X and Y. Further, let $g : \mathbb{R} \to \mathbb{R}$ be the capacity cost function (mapping). Then, we define (n, N, λ) code with a capacity cost $S \in \mathbb{R}$ as a system

- $\{(\boldsymbol{u}_1, A_1), \ldots, (\boldsymbol{u}_N, A_N)\}$
- $\boldsymbol{u}_i = [u_{i,1} \ldots u_{i,n}] \in \mathbb{R}^n$, $A_i \subset \mathbb{R}^n$, $A_i \cap A_j = \emptyset \forall i \neq j$
- $\sum_{k=1}^{n} \frac{g(u_{i,k})}{n} \leq S$, $\forall i$
- $\bar{\lambda} \triangleq \frac{1}{N} \sum_{i=1}^{N} \lambda_i$, $\lambda_M \triangleq \max_{i=1,\ldots,N} \lambda_i$, $\lambda_i \triangleq P(\boldsymbol{Y} \notin A_i | \boldsymbol{X} = \boldsymbol{u}_i)$, and $\lambda_M \leq \lambda$.

Theorem 1. *Random Coding Theorem*

We define the capacity with cost constraint S as

$$\begin{aligned} C(S) \triangleq & \max_{f_X(\cdot)} I(X;Y) \\ & \text{s.t.} \int_x f_X(x) g(x) \leq S. \end{aligned} \tag{2}$$

The theorem states that, For any "rate" R which satisfies $0 \leq R < C(S)$, there exists a sequence of $(n, 2^{nR}, \lambda(n))$ codes for $n = 1, 2, \ldots$ such that $\lambda(n) \to 0$ as $n \to \infty$.

In what follows, a proof is presented and for that reason, we need to define

$$\begin{aligned} A_\epsilon(n) \triangleq \{(\boldsymbol{x}, \boldsymbol{y}) : & \Big| -\frac{\log f_{\boldsymbol{X}}(\boldsymbol{x})}{n} - h(X) \Big| < \epsilon, \\ & \Big| -\frac{\log f_{\boldsymbol{Y}}(\boldsymbol{y})}{n} - h(Y) \Big| < \epsilon, \\ & \Big| -\frac{\log f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x})}{n} - h(Y,X) \Big| < \epsilon \Big\}, \end{aligned}$$

where $h(X) \triangleq -\int_x f_X(x) \log f_X(x) dx$. Further, we need to introduce the following two lemmas.

Lemma 1.

$$\underset{f_X(x)}{\boxed{MS}} \to \boldsymbol{X} \to \underset{f_{Y|X}(y|x)}{\boxed{MC}} \to \boldsymbol{Y}$$

- $P\left((\boldsymbol{X}, \boldsymbol{Y}) \in A_\epsilon(n)\right) > 1 - \epsilon$ *for sufficiently large* n
- *Vol*$(A_\epsilon(n)) \triangleq \int_{(\boldsymbol{x},\boldsymbol{y}) \in A_\epsilon(n)} d\boldsymbol{x} d\boldsymbol{y} < 2^{n(h(Y,X)+\epsilon)}$
- *Vol*$(A_\epsilon(n)) > (1-\epsilon) 2^{n(h(Y,X)-\epsilon)}$ *for sufficiently large* n

We omit the explicit proofs since they become trivial from the discrete case proofs in previous lectures.

Lemma 2. *Known as "the Packing Lemma"*
For independent $\tilde{\boldsymbol{X}}$ and $\boldsymbol{Y}$, i.e.,

$$\underset{f_X(x)}{\boxed{MS}} \to \tilde{\boldsymbol{X}}$$

$$\underset{f_X(x)}{\boxed{MS}} \to \boldsymbol{X} \to \underset{f_{Y|X}(y|x)}{\boxed{MC}} \to \boldsymbol{Y}$$

the following inequalities hold for sufficiently large n,

$$(1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)} < P\left((\tilde{\boldsymbol{X}}, \boldsymbol{Y}) \in A_\epsilon(n)\right) < 2^{-n(I(X;Y)-3\epsilon)}.$$

$$P\left((\tilde{\boldsymbol{X}}, \boldsymbol{Y}) \in A_\epsilon(n)\right) = \int_{(\tilde{\boldsymbol{x}},\boldsymbol{y}) \in A_\epsilon(n)} \underbrace{f_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}) f_{\boldsymbol{Y}}(\boldsymbol{y})}_{f_{\tilde{\boldsymbol{X}},\boldsymbol{Y}}(\tilde{\boldsymbol{x}},\boldsymbol{y})} d\tilde{\boldsymbol{x}} d\boldsymbol{y}$$

From the definition of $A_\epsilon(n)$, we have $2^{-n(h(X)+\epsilon)} < f_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}) < 2^{-n(h(X)-\epsilon)}$, $2^{-n(h(Y)+\epsilon)} < f_{\boldsymbol{Y}}(\boldsymbol{y}) < 2^{-n(h(Y)-\epsilon)}$. This gives us the upper bound

$$\begin{aligned} P\left((\tilde{\boldsymbol{X}}, \boldsymbol{Y}) \in A_\epsilon(n)\right) &< 2^{-n(h(X)+h(Y)-2\epsilon)} \int_{(\tilde{\boldsymbol{x}},\boldsymbol{y}) \in A_\epsilon(n)} d\tilde{\boldsymbol{x}} d\boldsymbol{y} \\ &< 2^{-n(h(X)+h(Y)-2\epsilon)} \underbrace{2^{n(h(Y,X)+\epsilon)}}_{\text{lemma 1}} \\ &= 2^{-n(I(X;Y)-3\epsilon)}, \end{aligned}$$

and analogously the lower bound $P\left((\tilde{\boldsymbol{X}},\boldsymbol{Y})\in A_\epsilon(n)\right) > 2^{-n(I(X:Y)+3\epsilon)}$, for sufficiently large n.

The proof of Theorem 1. relies on Shannon's random coding argument, which utilizes the fact that if the error probability averaged over all possible codes goes to zero as $n\to\infty$, then there must exist at least one code that provides an error probability smaller than the average.

1) Random Code Generation:

- Choose an arbitrary $\epsilon > 0$
- Choose a pdf $f_X(x)$ s.t. $\int_x f_X(x)g(x)dx \,=\, S-\epsilon < S$
- Generate $u_{i,k}$ for all $i=1,\ldots,N$ and $k=1,\ldots,n$ independently using $f_X(x)$.

2) Encoding: Let us label the possible codewords $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N$ as messages $w=1,\ldots,N$. When the transmitter wants to communicate message k, it will transmit $\boldsymbol{u}_k$.

3) Decoding: Let us use the following typical set decoder (for the code $\mathcal{C}$) that, for a given channel output $\boldsymbol{y}$, estimates the transmitted message as or be in error if

$$\hat{w}_{\mathcal{C}}(\boldsymbol{y}) = \begin{cases} i, & \text{if } \boldsymbol{y}\in A_i(\mathcal{C}) \text{ and } \boldsymbol{y}\notin A_j(\mathcal{C})\ \forall j\neq i \\ & \text{and } \sum_{k=1}^n \frac{g(u_{i,k})}{n} \le S \\ 0, & \text{otherwise (i.e. decoding error)} \end{cases},$$

where $A_i(\mathcal{C}) \triangleq \{\boldsymbol{y}\in\mathbb{R}^n \big| (\boldsymbol{u}_i,\boldsymbol{y})\in A_\epsilon(n)\}$.

The average error probability $\bar{\lambda}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N) \triangleq \frac{1}{N}\sum_{i=1}^N \lambda_i(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)$ and the error probability $\lambda_i(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N) \triangleq P\left(\hat{w}(\boldsymbol{Y})\neq i \big| \boldsymbol{X}=\boldsymbol{u}_i\right)$ are both deterministic for a particular realization of the code book $\mathcal{C}=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)$. Each realization $\boldsymbol{u}_i$ is drawn from the input pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$, and this is done independently over i. If we define $\boldsymbol{C}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i,\ldots,\boldsymbol{X}_N)$, where $\boldsymbol{X}_i$ are i.i.d. with the pdf $f_{\boldsymbol{X}}$, we can say that $f_{\boldsymbol{C}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)=\prod_{i=1}^N f_{\boldsymbol{X}}(\boldsymbol{x}_i)$ and that the code book $\mathcal{C}$ is drawn from $f_{\boldsymbol{C}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$. Now, the expectation of $\bar{\lambda}$ (taken over $\boldsymbol{C}$, i.e., the whole ensemble of codes) is

$$\mathbb{E}_{\boldsymbol{C}}\{\bar{\lambda}(\boldsymbol{C})\} = \mathbb{E}_{\boldsymbol{C}}\left\{\frac{1}{N}\sum_{i=1}^N \lambda_i(\boldsymbol{C})\right\} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{C}}\{\lambda_i(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N)\} = \mathbb{E}_{\boldsymbol{C}}\{\lambda_1(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N)\},$$

where the last equality utilizes the fact that the expectation value is invariant to any reordering of $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$ due to the i.i.d. property of $\boldsymbol{X}_i$ over i. Before we continue further, let us introduce a compact notation $A_0(\mathcal{C})$ for the set $\{\boldsymbol{y}\big|\sum_{k=1}^n \frac{g(u_{1,k})}{n}\le S\}$, which may seem strange since the condition does not depend on $\boldsymbol{y}$. Essentially, the set $A_0(\mathcal{C})$ will either contain the whole $\boldsymbol{y}$-space if the transmitted codeword $\boldsymbol{u}_1$ fulfills the constraint on $g(\cdot)$, or it will be the null set when the

constraint is not met. Hence,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{C}}\{\lambda_1(\boldsymbol{C})\} &= \int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \left(\int_{\boldsymbol{y};\hat{w}_{\mathcal{C}}(\boldsymbol{y})\neq 1} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} \right) d\mathcal{C} \\
&= \int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \left(\int_{A_1^{\complement}(\mathcal{C})\cup A_2(\mathcal{C})\cdots\cup A_N(\mathcal{C})\cup A_0^{\complement}(\mathcal{C})} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} \right) d\mathcal{C} \\
&\leq \underbrace{\int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \int_{A_1^{\complement}(\mathcal{C})} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C}}_{T_1} \\
&\quad + \sum_{i=2}^{N} \underbrace{\int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \int_{A_i(\mathcal{C})} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C}}_{T_i} \\
&\quad + \underbrace{\int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \int_{A_0^{\complement}(\mathcal{C})} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C}}_{T_0}.
\end{aligned}$$

We can continue from this point on by evaluating the integrals $T_0, T_1, T_2, \ldots, T_N$ separately. Starting with T_0, note that the condition defining the set A_0 does not depend on $\boldsymbol{y}$, which gives

$$\begin{aligned}
T_0 &= \int_{\mathcal{C}\in\mathbb{R}^{Nn};\sum_{k=1}^{n}\frac{g(u_{1,k})}{n}\geq S} f_{\boldsymbol{C}}(\mathcal{C}) \int_{\mathbb{R}^n} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C} \\
&= \int_{\boldsymbol{u}_1\in\mathbb{R}^n;\sum_{k=1}^{n}\frac{g(u_{1,k})}{n}\geq S} f_{\boldsymbol{X}}(\boldsymbol{u}_1) d\boldsymbol{u}_1 < \epsilon,
\end{aligned}$$

for sufficiently large n where the last inequality follows from the weak law of large numbers. That is, $\frac{1}{n}\sum_{k=1}^{n} g(u_{1,k}) \to \mathbb{E}_X\{g(X)\} = S - \epsilon$ in probability as $n \to \infty$. Continuing with T_1,

$$\begin{aligned}
T_1 &= \int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \int_{\boldsymbol{y};(\boldsymbol{u}_1,\boldsymbol{y})\notin A_\epsilon(n)} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C} \\
&= \int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{u}_1) \int_{\boldsymbol{y};(\boldsymbol{u}_1,\boldsymbol{y})\notin A_\epsilon(n)} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\boldsymbol{u}_1 \\
&= \int_{A_\epsilon^{\complement}(n)} f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{u}_1) d\boldsymbol{y} d\boldsymbol{u}_1 = P\left((\boldsymbol{X},\boldsymbol{Y}) \notin A_\epsilon(n) | \boldsymbol{X} = \boldsymbol{u}_1\right) < \epsilon,
\end{aligned}$$

for sufficiently large n where the last inequality follows from lemma 1. Lastly T_i for $i = 2, \ldots, N$,

$$\begin{aligned}
T_i &= \int_{\mathbb{R}^{Nn}} f_{\boldsymbol{C}}(\mathcal{C}) \int_{\boldsymbol{y};(\boldsymbol{u}_i,\boldsymbol{y})\in A_\epsilon(n)} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{u}_1) d\boldsymbol{y} d\mathcal{C} \\
&= \int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{u}_i) \int_{\boldsymbol{y};(\boldsymbol{u}_i,\boldsymbol{y})\in A_\epsilon(n)} f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{u}_i \\
&= \int_{A_\epsilon(n)} f_{\boldsymbol{X}}(\boldsymbol{u}_i) f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{u}_i \leq 2^{-n(I(Y;X)-3\epsilon)},
\end{aligned}$$

were the last inequality follows from lemma 2. These upper bounds result in

$$\mathbb{E}_{\boldsymbol{C}}\{\bar{\lambda}(\boldsymbol{C})\} < 2\epsilon + N2^{-n(I(Y;X)-3\epsilon)} < 2\epsilon + 2^{nR}2^{-n(I(Y;X)-3\epsilon)} = 2\epsilon + 2^{-n(I(Y;X)-R-3\epsilon)},$$

where $R \triangleq \frac{\log(N)}{n} \geq 0$ represents the "code rate". For $R < I(Y;X) - 3\epsilon$, we can choose sufficiently large n such that $\mathbb{E}_{\boldsymbol{C}}\{\bar{\lambda}(\boldsymbol{C})\} < 3\epsilon$.

Since $\mathbb{E}_{\boldsymbol{C}}\{\bar{\lambda}(\boldsymbol{C})\} < 3\epsilon$, following Shannon's random coding argument, there exists a code $(n, 2^{nR}, \cdot)$ that has an average error probability that is smaller than 3ϵ. In essence, we will only pick a code book that will yield a small average error probability, which also means that only those codewords that satisfy the capacity cost constraint will be included. Note that the decoding regions A_i can be made disjoint without increasing the upper bound on the average error probability. So far, we have not said anything about the maximum error probability, which is the quantity of interest in order to communicate with a certain level of reliability. Nevertheless, given an upper bound on the average error probability, say 3ϵ, the best $N/2$ codewords (smallest error probabilities) have a maximum error probability that is smaller than 6ϵ. If it were not true, then the average error probability would not be smaller than 3ϵ due to the contribution of the worst $N/2$ codewords each having an error probability greater than 6ϵ, i.e., $\frac{1}{N}\frac{N}{2}6\epsilon = 3\epsilon \not< 3\epsilon$. Using this fact, we can construct a new code consisting only of the $N/2$ best codewords, which would then reduce the new rate R' to

$$R' = \frac{\log(N/2)}{n} = R - \frac{1}{n}.$$

We can conclude now that since $S - \epsilon < S$, it follows that $R' < I(X;Y) < C(S)$. However by choosing $\epsilon > 0$ to be arbitrarily small and n sufficiently large, we can achieve any rate less than $C(S)$.

Capacity cost of the discrete-time continuous memoryless AWGN channel with an average input power constraint: Consider the discrete-time memoryless additive white Gaussian noise channel (AWGN), i.e., $Y = X + N$, with $X \in \mathbb{R}$ as the input, $Y \in \mathbb{R}$ as the output and $N \in \mathbb{R}$ as the AWGN with variance σ^2. Further, let us consider an average power constraint on the input, i.e. $\mathbb{E}[X^2] \leq S$. The capacity cost for this channel can be computed using (2), with the cost function $g(x) = x^2$.

$$\begin{aligned} C(S) &= \max_{f_X(\cdot):\mathbb{E}_X\{X^2\}\leq S} I(X;Y) = \max_{f_X(\cdot):\mathbb{E}_X\{X^2\}\leq S} h(Y) - \underbrace{h(Y|X)}_{h(N)} \\ &= \max_{f_X(\cdot):\mathbb{E}_X\{X^2\}\leq S} h(Y) - \frac{1}{2}\log(2\pi e\sigma^2) = \max_{\mathrm{Var}\{Y\}:\mathbb{E}_X\{X^2\}\leq S} \frac{1}{2}\log\left(\frac{\mathrm{Var}\{Y\}}{\sigma^2}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{S}{\sigma^2}\right). \end{aligned}$$

where the maximum is achieved when X is Gaussian distributed with mean 0 and variance S.

Later in Lecture 9, we will see that $\frac{1}{2}\log\left(1 + \frac{S}{\sigma^2}\right)$ is indeed the capacity of this channel.