EE376A/STATS376A Information Theory Lecture 11 - 02/13/2018

Lecture 11: Channel Coding Theorem: Converse Part

Lecturer: Tsachy Weissman *Scribe: Erdem Bıyık*

In this lecture[1], we will continue our discussion on channel coding theory. In the previous lecture, we proved the direct part of the theorem, which suggests if $R < C^{(I)}$, then R is achievable. Now, we are going to prove the converse statement: If $R > C^{(I)}$, then R is not achievable. We will also state some important notes about the direct and converse parts.

1 Recap: Communication Problem

Recall the communication problem:

$$J \sim \text{Unif}\{1, 2, \ldots, M\} \longrightarrow \boxed{\text{encoder}} \xrightarrow{X^n} \boxed{\begin{array}{c}\text{Memoryless Channel}\\ P(Y|X)\end{array}} \xrightarrow{Y^n} \boxed{\text{decoder}} \longrightarrow \hat{J}$$

- Rate $= R = \frac{\log M}{n} \; \frac{\text{bits}}{\text{channel use}}$
- Probability of error $= P_e = P(\hat{J} \neq J)$

The main result is $C = C^{(I)} = \max_{P_X} I(X;Y)$. Last week, we showed R is achievable if $R < C^{(I)}$. In this lecture, we are going to prove that if $R > C^{(I)}$, then R is not achievable.

2 Fano's Inequality

Theorem *(Fano's Inequality)*. Let X be a discrete random variable and $\hat{X} = \hat{X}(Y)$ be a guess of X based on Y. Let $P_e = P(X \neq \hat{X})$. Then,

$$H(X|Y) \leq h_2(P_e) + P_e \log(|\mathcal{X}| - 1)$$

where h_2 is the binary entropy function.

Proof. Let $V = 1_{\{X \neq \hat{X}\}}$, i.e. V is 1 if $X \neq \hat{X}$ and 0 otherwise.

$$\begin{aligned}
H(X|Y) &\leq H(X, V|Y) && (1)\\
&= H(V|Y) + H(X|V, Y) && (2)\\
&\leq H(V) + H(X|V, Y) && (3)\\
&= H(V) + \sum_{v,y} H(X|V=v, Y=y)P(V=v, Y=y) && (4)\\
&= H(V) + \sum_{y} H(X|V=0, Y=y)P(V=0, Y=y) + \sum_{y} H(X|V=1, Y=y)P(V=1, Y=y) && (5)\\
&= H(V) + \sum_{y} H(X|V=1, Y=y)P(V=1, Y=y) && (6)\\
&\leq H(V) + \log(|\mathcal{X}| - 1)\sum_{y} P(V=1, Y=y) && (7)\\
&= H(V) + \log(|\mathcal{X}| - 1)P(V=1) && (8)\\
&= h_2(P_e) + P_e \log(|\mathcal{X}| - 1) && (9)
\end{aligned}$$

[1] *Reading:* Chapter 7.9 and 7.12 of Cover, Thomas M., and Joy A. Thomas. Elements of information theory. Wiley, 2006.

where (1) is from data processing inequality, (2) is due to chain rule, (3) is because conditioning can only reduce (or not change) entropy. (4) directly follows from the definition of conditional entropy. (6) is because when $V = 0$, $X = \hat{X}$ and X is a function of Y, so $H(X|V=0, Y=y) = 0$. Note that $H(X|V = 1, Y = y)$ is maximized when $P(X|V = 1, Y = y)$ is uniformly distributed, which yields to $\log(|\mathcal{X}| - 1)$. Hence, (7) follows. The next step is just law of total probability, and completes the proof. □

Note a weaker version of Fano's inequality is

$$H(X|Y) \leq 1 + P_e \log|\mathcal{X}| \tag{10}$$

which will be useful later in proving the converse theorem. This is also stated as

$$P_e \geq \frac{H(X|Y) - 1}{\log \mathcal{X}} \tag{11}$$

Fano's inequality basically says that if $H(X|Y)$ is large, i.e., if given Y, X has a lot of uncertainty, then any estimator of X based on Y must have a large probability of error.

3 Proof of Converse Part

For any scheme,

$$\begin{aligned}
\log M - H(J|Y^n) &= H(J) - H(J|Y^n) && (12)\\
&= I(J; Y^n) && (13)\\
&= H(Y^n) - H(Y^n|J) && (14)\\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, J) && (15)\\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1}, J) && (16)\\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1}, X_i, J) && (17)\\
&= \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) && (18)\\
&= \sum_{i=1}^{n} I(X_i; Y_i) && (19)\\
&\leq nC^{(I)} && (20)
\end{aligned}$$

where (12) is because J is uniformly distributed, (13), (14) and (19) are directly from the definition of mutual information, (15) is from the properties of joint/conditional entropy, (16) and (17) are due to the fact that conditioning can only decrease (or not change) entropy. Since the channel is memoryless (i.e. Y_i—X_i—(Y^{i-1}, J)), (18) follows. Next, (20) is because the capacity is the maximum of the mutual information between input and output.

Now, for schemes with $\frac{\log M}{n} \geq R$,

$$\begin{aligned} P_e &\geq \frac{H(J|Y^n) - 1}{\log M} && (21) \\ &\geq \frac{\log M - nC^{(I)} - 1}{\log M} && (22) \\ &= 1 - \frac{nC^{(I)}}{\log M} - \frac{1}{\log M} && (23) \\ &\geq 1 - \frac{C^{(I)}}{R} - \frac{1}{nR} && (24) \\ &\overset{n\to\infty}{\longrightarrow} 1 - \frac{C^{(I)}}{R} && (25) \end{aligned}$$

where (21) is due to weaker version of Fano's inequality, (22) is from the result obtained with (20), and (24) is because $\frac{\log M}{n} \geq R$. The result shows that when $R > C^{(I)}$, there exists a positive lower bound on the probability of error, so R is not achievable. □

4 Geometric Interpretation of Converse Part

To interpret the converse theorem geometrically, consider Fig. 1, where a communication channel is visualized with the corresponding typical sets. In this scheme, 2^{nR} codewords are selected from $\mathcal{X}^n$. After transmission through the channel, each codeword has a typical set of size $2^{nH(Y|X)}$ (we'll study the notion of conditional typicality in more detail later). Also, note that the typical set of Y^n has size $2^{nH(Y)}$, naturally independent from X. Now, note that the typical channel outputs given $X^n(i)$'s should not intersect in order to have zero probability of error. Hence, there must be at least $2^{nR}2^{nH(Y|X)}$ elements in the typical channel outputs. By this volume argument,

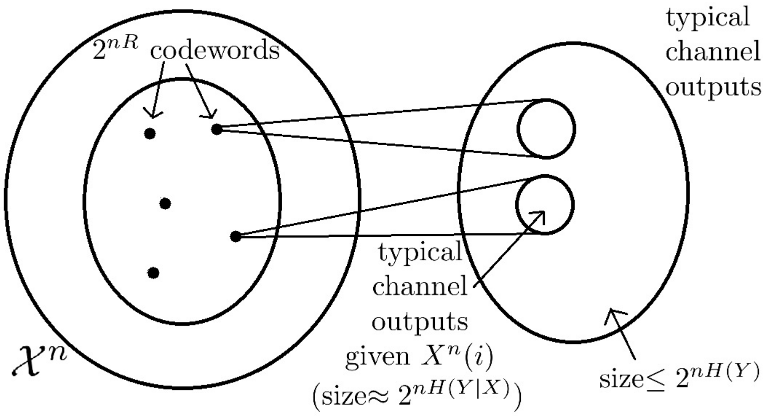

Figure 1: Geometric interpretation of communication problem

$$\begin{aligned} 2^{nH(Y)} &\geq 2^{nR}2^{nH(Y|X)} && (26) \\ 2^{nH(Y)-nH(Y|X)} &\geq 2^{nR} && (27) \\ 2^{nI(X;Y)} &\geq 2^{nR} && (28) \\ I(X;Y) &\geq R && (29) \end{aligned}$$

Since $I(X;Y) \leq C$, we must have $R \leq C$.

5 Some Notes on the Direct and Converse Parts

5.1 Communication with Feedback

Now assume X_i is a function of both J and Y^{i-1} (previously, it was a function of only J), so the encoder knows what decoder receives. This is obviously a stronger encoder, as it has more information. However, it can be verified that the proof of the converse theorem is valid for memoryless channels with feedback, as well. This can be directly seen from that the proof uses the properties of the channel only at Eq. (18), which also holds when feedback is allowed (because the Markov property still holds: Y_i—X_i—Y^{i-1}, J). Moreover, achievability result is obvious as the feedback can be ignored. Therefore, the maximum achievable rate remains the same with feedback.

On the other hand, this setting increases the reliability of the system, i.e. the probability of error vanishes faster; and the schemes become simpler.

Example. Recall the binary erasure channel (BEC) shown in Fig. 2. Also recall that the capacity of BEC is $C = 1 - p$ bits/channel use where p is the erasure probability. Consider a binary erasure channel with feedback. A very simple scheme that achieves capacity would be to repeat each information bit until it is correctly received by the decoder. With this scheme the probability that a bit is correctly sent through the channel at one attempt is $1 - p$, at two attempts is $p(1 - p)$, and so on. Hence, it follows a geometric distribution, whose mean is $\frac{1}{1-p}$. Therefore, we have

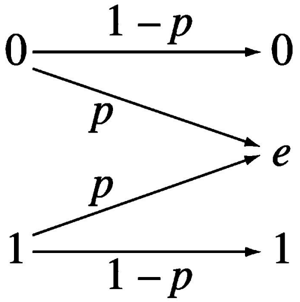

Figure 2: Binary erasure channel (image from Wikipedia)

$$\frac{1}{1-p} \text{ channel uses per information bit}$$

Equivalently, $R = C$. This approach can be extended to all memoryless channels[2].

5.2 Practical Schemes

In the proof of the direct part, we showed mere existence of schemes C_n of size at least 2^{nR} and arbitrarily small probability of error. We did not explicitly give particular code constructions that achieve this. For practical schemes that enable communication with rates arbitrarily close to the maximum mutual information with reasonable complexity, we refer to:

[2] *Reading:* Horstein, Michael. "Sequential transmission using noiseless feedback." IEEE Transactions on Information Theory 9.3 (1963): 136-143.
Additional References: Schalkwijk, J., and Thomas Kailath. "A coding scheme for additive noise channels with feedback–I: No bandwidth constraint." IEEE Transactions on Information Theory 12.2 (1966): 172-182.
Shayevitz, Ofer, and Meir Feder. "Optimal feedback communication via posterior matching." IEEE Transactions on Information Theory 57.3 (2011): 1186-1222.
Li, Cheuk Ting, and Abbas El Gamal. "An efficient feedback coding scheme with low error probability for discrete memoryless channels." IEEE Transactions on Information Theory 61.6 (2015): 2953-2963.

1. Low-density parity-check (LDPC) codes
2. Polar codes

The course EE388 - Modern Coding Theory is encouraged for much more detail. In this course, we are going to dedicate some time to either of these codes in lectures or homeworks.

5.3 Generalization to Infinite Alphabets

Proof of the direct part assumed finite alphabets; however it carries over to general case by approximation / quantization.

5.4 $\mathbf{P_e}$ vs $\mathbf{P_{max}}$

We previously talked about P_{max} which may be more reasonable for some practical systems, and puts a more stringent condition than P_e:

$$\begin{aligned} P_e = P(\hat{J} \neq J) &= \frac{1}{m} \sum_{j=1}^{m} P(\hat{J} \neq j | J = j) \\ P_{max} &= \max_j P(\hat{J} \neq j | J = j) \end{aligned}$$

Claim. Given a coding scheme C_n such that $P_e \to 0$ as $n \to \infty$, we can find another scheme C'_n, that achieves $P_{max} \to 0$.

Proof. By Markov's inequality, we have

$$\left| \{1 \leq j \leq m : P(\hat{J} \neq j | J = j) \leq 2P_e\} \right| \geq \frac{m}{2}$$

Hence, given C_n with $|C_n| = m$, R_C and P_e, there exists a C'_n such that $|C'_n| \geq m/2$ and $P_{max} \leq 2P_e$. The rate of this new code is then $R_{c'_n} \geq \frac{\log(M/2)}{n} = \frac{\log M}{n} - \frac{1}{n} \geq R - \epsilon$ for arbitrarily small ϵ when n is large. Note P_{max} of C'_n goes to zero as $P_e \to 0$. □

Thus, capacity is the same regardless of whether reliable communication is with respect to P_e or P_{max}.