Lecture 26*

Agenda for the lecture

- Random coding proofs of capacity theorem: information density based proof

26.1 Han and Verdú’s random coding proof of achievability

In the previous lecture, we gave a random coding based proof of achievability which showed that a “typical decoder” can distinguish roughly $2^{nI(P;W)}$ messages reliably by using the channel $n$ times, when codewords are randomly generated as i.i.d. with common distribution $P$. However, the proof was mysterious and did not clarify what is it about a good channel that enables the existence of such codes of large sizes. In this lecture, we present an alternative proof due to Han and Verdú which explicitly clarifies this point. Furthermore, the proof is single-shot, i.e., it works even for $n = 1$. We have already encountered such single-shot results in the form of little-big lemmas in source coding and hypothesis testing. Now, we extend that approach to channels, where achievable rate will be obtained by using the Chebyshev’s inequality with the single-shot achievability result.

To state the result, we introduce the notion of information density, a quantity which plays a similar role for mutual information as entropy density for entropy.

Copyright (2018) Himanshu Tyagi. Feel free to use with acknowledgement.
Definition 26.1. For a joint distribution $P_{XY}$, the information density $i_{P_{XY}}(X,Y)$ is a random variable given by

$$i_{P_{XY}}(x,y) = \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}.$$  

Note that

$$\mathbb{E}_{P_{XY}}[i_{P_{XY}}(X,Y)] = I(X \wedge Y).$$

Also, $i(x,y) \geq \lambda > 0$ indicates that $(x,y)$ is more likely to occur under $P_{XY}$ than under $P_X \times P_Y$. Thus, if with a large probability $i(X,Y) \geq \lambda$, then, heuristically, the random variables $X, Y$ are not independent. Our single shot result will show that a channel whose input and output is so dependent for any input distribution $P$ can be used to transmit $2^{\lambda}$ messages reliably in just $n = 1$ channel use.

Theorem 26.2 (Han-Verdú random coding theorem). For a discrete channel $(X,W,Y)$, a pmf $P$ on $\mathcal{X}$, and $\lambda > 0$, suppose that for $0 < \epsilon < 1$ and with $P_{XY}(x,y) = P(x)W(y|x)$ the set

$$\mathcal{T} = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : i_{P_{XY}}(x,y) \geq \lambda\},$$  

satisfies

$$\sum_{(x,y) \in \mathcal{T}} P(x)W(y|x) \geq 1 - \epsilon.$$  

Then, there exists a $(1, 2^{\lambda-\eta})$ code for the channel $W$ with average probability of error less than $\epsilon + 2^{-\eta}$.

The result above allows us to prove bounds for achievable rates by exhibiting $\lambda$ and $\epsilon$ which satisfy (1) and (2) for any distribution $P$ of our choice. In particular, for a DMC $(\mathcal{X}, W, \mathcal{Y})$, we can apply this result to the channel $(\mathcal{X}^n, W^n, \mathcal{Y}^n)$ using a $P$ corresponding to
an i.i.d. $X^n$ and get a $\lambda$ using Chebyshev's inequality. We obtain the following corollary.

**Corollary 26.3.** For a DMC $(\mathcal{X}, W, \mathcal{Y})$ and any pmf $P$ on $\mathcal{X}$, we have

$$C^{\text{ave}}(W) \geq I(P; W).$$

**Proof.** Let $X^n$ be i.i.d. with common distribution $P$ and $Y^n$ be the output of $n$ uses of the DMC $W$ when the input is $X^n$. For this joint distribution, the information density $i(X^n, Y^n)$ is given by

$$i(X^n, Y^n) = \log \frac{W^n(Y^n|X^n)}{(PW)^n(Y^n)} = \sum_{i=1}^{n} \log \frac{W(Y_i|X_i)}{(PW)(Y_i)} =: \sum_{i=1}^{n} Z_i,$$

where $Z_i$ denotes the log $W(Y_i|X_i)/PW(Y_i)$. Note that $i(X^n, Y^n) = \sum_{i=1}^{n} Z_i$ and that the rvs $Z_1, ..., Z_n$ are i.i.d. with $\mathbb{E}[Z_i] = I(X \wedge Y)$; let $V$ denote the variance of each $Z_i$. Then, since

$$\mathbb{E}\left[\sum_{i=1}^{n} Z_i\right] = nI(X \wedge Y) \text{ and } \text{Var}\left[\sum_{i=1}^{n} Z_i\right] = nV,$$

by Chebyshev's inequality

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq nI(X \wedge Y) - \sqrt{nV/\epsilon}\right) \leq \epsilon.$$

Thus, the conditions of Theorem 26.2 hold for the channel $(\mathcal{X}^n, W, \mathcal{Y}^n)$ with $\lambda = nI(X \wedge Y) - \sqrt{nV/\epsilon}$. Therefore, there exists a $(1, M)$ code for $W^n$, which is an $(n, M)$ code for $W$, with probability of error less than $\epsilon + 2^{-n\eta}$ and

$$\log M \leq nI(X \wedge Y) - \sqrt{\frac{nV}{\epsilon}} - n\eta,$$

i.e., a code of rate $I(X \wedge Y) - \sqrt{V/n\epsilon} - \eta$ for every $\eta > 0$. Therefore, $I(X \wedge Y)$ is an achievable rate. \qed
26.2 Proof of Theorem 26.2

The proof is very similar to the random coding proofs presented in the previous lecture. We begin by describing our encoder and decoder.

26.2.1 The random codebook and the decoder

Our random codebook $C$ consists of codewords $X_1, \ldots, X_M \in \mathcal{X}$ generated i.i.d. with common distribution $P$. The decoder once again uses the template of Figure 2 in lecture 7, with individual decision $B_m$ described as follows:

$$B_m(Y) = 1((X_m, Y) \in \mathcal{T}).$$

Note that $B_m$ is randomized mapping since it depends on $X_m$, which is a part of the random codebook.

26.2.2 Error analysis

By the considerations of the previous lecture, we have

$$\mathbb{E}[\epsilon(C)] \leq \mathbb{P}(B_1 = 0 | \text{1 sent}) + M \mathbb{P}(B_2 = 1 | \text{1 sent}).$$

We bound each term above. For the first term,

$$\mathbb{P}(B_1 = 0 | \text{1 sent}) = \mathbb{E}[\mathbb{E}[\epsilon(C) | X_1]]$$

$$= \mathbb{E}[\mathbb{E}[1(B_1 = 0) | \text{1 sent}, X_1]]$$

$$= \sum_x P_X(x) \mathbb{P}(B_1 = 0 | \text{1 sent}, X_1 = x)$$

$$= \sum_x P_X(x) \mathbb{P}(B_1(Y) = 0 | \text{1 sent}, X_1 = x)$$

$$= \sum_x P_X(x) \sum_{y : (x, y) \in \mathcal{T}^c} W(y | x)$$
where $P_{XY}(x, y) = P(x)W(y|x)$.

For the second term,

\[
\begin{align*}
P(B_2 = 1|1 \text{ sent}) &= \mathbb{E}[\mathbb{E}[\epsilon(C)|X_1, X_2]] \\
&= \sum_{x_1, x_2} P(x_1)P(x_2)\mathbb{P}(B_2(Y) = 0|1 \text{ sent}, X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1, x_2} P(x_1)P(x_2)\sum_{y:(x_2,y)\in\mathcal{T}} W(y|x_1) \\
&= \sum_{x_2} P(x_2)\sum_{y:(x_2,y)\in\mathcal{T}} \sum_{x_1} P(x_1)W(y|x_1) \\
&= \sum_{x_2} P(x_2)\sum_{y:(x_2,y)\in\mathcal{T}} (PW)(y) \\
&\leq \sum_{x_2} P(x_2)\sum_{y:(x_2,y)\in\mathcal{T}} 2^{-\lambda}W(y|x_2) \\
&\leq 2^{-\lambda},
\end{align*}
\]

where the last equality uses $\sum_x P(x)W(y|x) = (PW)(y)$ and the first inequality uses the definition of $\mathcal{T}$.

Combining the two bounds above, if $M \leq 2^{\lambda-\eta}$, we get

\[
\mathbb{E}[\epsilon(C)] \leq \epsilon + 2^{-\eta},
\]

which completes the proof.