LECTURE 9

Last time:

- Channel Capacity
- BSC and BEC

Lecture outline

- Continue on Joint AEP
- Coding Theorem

Reading: Reading: Scts. 8.4- 8.7, 8.9

Definition The set of the joint typical sequences $\{(\underline{x}^n, \underline{y}^n)\}$ is

$$
\begin{aligned}
A_\epsilon^{(n)} = & \Big\{(\underline{x}^n, \underline{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \left| -\frac{1}{n} \log P_{\underline{X}^n}(\underline{x}^n) - H(X) \right| < \epsilon \\
& \left| -\frac{1}{n} \log P_{\underline{Y}^n}(\underline{y}^n) - H(Y) \right| < \epsilon \\
& \left| -\frac{1}{n} \log P_{\underline{X}^n, \underline{Y}^n}(\underline{x}^n, \underline{y}^n) - H(X,Y) \right| < \epsilon \Big\}
\end{aligned}
$$

Theorem

Consider sequences $(\underline{X}^n, \underline{Y}^n)$ drawn i.i.d. from $P_{X,Y}$, then

- $P((\underline{X}^n, \underline{Y}^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
- $|A_\epsilon^{(n)}| \leq 2^{nH(X,Y)+\epsilon}$

Simple extension of the AEP for single random variable.

Joint AEP

- Pass $\underline{x}^n$ through the channel, to obtain $\underline{y}^n$, with high probability $(\underline{x}^n, \underline{y}^n)$ are jointly typical.
- We only observe $\underline{y}^n$, try to find the $\underline{x}^n$ that is jointly typical.

Question:
If I randomly pick a typical sequence $\underline{\tilde{X}}^n \in A_\epsilon^{(n)}(X)$, and a $\underline{\tilde{Y}}^n \in A_\epsilon^{(n)}(Y)$, are they joint typical?

$$A_\epsilon^{(n)}(X) \quad \times A_\epsilon^{(n)}(Y) \stackrel{?}{=} \quad A_\epsilon^{(n)}(X, Y)$$

$$2^{nH(X)} \quad \times 2^{nH(Y)} \geq \quad 2^{nH(X,Y)}$$

Joint AEP

Theorem If $(\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n)$ are independent with the same marginal distributions, i.e., $(\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \sim P_{\underline{X}^n}(\underline{x}^n) P_{\underline{Y}^n}(\underline{y}^n)$,

$$P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

and for n large enough,

$$P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \geq (1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)}$$

Proof

$$\begin{aligned} & P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \\ = & \sum_{A_\epsilon^{(n)}(X,Y)} P_{\underline{X}^n}(\underline{x}^n) P_{\underline{Y}^n}(\underline{y}^n) \\ \leq & \; 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\ = & \; 2^{-n(I(X;Y)-3\epsilon)} \end{aligned}$$

Example:Binary Source and BSC

Consider a binary source $\underline{X}^n$ passes through a binary symmetric channel with flipping probability p.

Fix an arbitrary input sequence $\underline{y}^n$, what $\underline{x}^n$'s are jointly typical?

Write $\underline{w}^n = \underline{y}^n - \underline{x}^n$, there should be approximately np 1's in $\underline{w}^n$.

How many of such $\underline{x}^n$ are there? $\doteq 2^{nH(p)}$.

Check: This set is much smaller than $\mathcal{X}^n = 2^n$.

Discussions

- Fix a typical $\underline{y}^n$, out of the $2^{nH(X)}$ typical X sequences, there are approximately $2^{nH(X|Y)}$ sequences that are jointly typical with $\underline{y}^n$.
- Pass $\underline{x}^n$ through the channel to obtain a joint typical pair $(\underline{x}^n, \underline{y}^n)$. Assume now that only $\underline{y}^n$ is observed. If we pick randomly another $\underline{\tilde{x}}^n$ and ask if this is the originally transmitted sequence, checking joint typically gives a probability of confusion $2^{-nI(X;Y)}$.
- Checking the joint typicality is a good way of decoding.

Overview

- Consider a DMC with transition probabilities $P_{Y|X}(y|x)$
- We say a data rate R is achievable if there exists a sequence of code books, $\mathcal{C}^{(n)}$, each has 2^{nR} codewords, for which the probability of error goes to 0 as $n \to \infty$.
- Notice whenever a code book is chosen, it is revealed to the receiver.
- The relation between the "probability of error" and the "capacity"

Construction

- Random code books: for each n, choose 2^{nR} codewords, each with length n, i.i.d. from an input distribution $P_X(x)$.

$$P(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} P_X(x_i(w))$$

- All messages are equiprobable. Define message W

$$P(W = w) = 2^{-nR}, \quad \text{for } w = 1, \ldots, 2^{nR}$$

Encoding and Decoding

Transmitter

Depend on the message $W = w$, transmit the codeword $\underline{x}^n(w)$ through n usages of the channel.

- Distinguish i.i.d. random code and transmitting independent symbols.

Receiver

Typical set decoding:

- For a given $\underline{y}^n$, if there exists unique $\underline{x}^n(w)$ in the code book that is jointly typical with $\underline{y}^n$, decode $\widehat{W} = w$.
- otherwise, declare an error

Random coding

Calculating the Probability of Error

$$
\begin{aligned}
P(\text{error}) &= \sum_{\mathcal{C}} P(\mathcal{C}) P(\text{error}|\mathcal{C}) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P(\text{error}|\mathcal{C}, W = w) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) P(\text{error}|\mathcal{C}, W = w) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) P(\text{error}|\mathcal{C}, W = 1) \\
&= P(\text{error}|W = 1)
\end{aligned}
$$

where the error probability is computed by averaging over the ensemble of the codes.

- Random coding creates symmetry.
- Instead of the error probability of one particular code, we compute the average error probability of the random codes.
- If the average probability of error is small, then there exists one code with small probability of error.

Using Joint AEP

Define

$$E_i = \{(\underline{x}^n(i), \underline{y}^n) \in A_\epsilon^{(n)}\}$$

for $i = 1, \ldots, 2^{nR}$.

$$\begin{aligned} P(\text{error}|W=1) &= P(E_1^c \cup E_2 \cup \ldots \cup E_{2^{nR}}) \\ &\leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \end{aligned}$$

Event E_i is in the space $\mathcal{X}^n \times \mathcal{Y}^n$. In computing the probability of E_i, what the distribution of $\underline{X}^n, \underline{Y}^n$ we should use?

- For E_1, $\underline{X}^n, \underline{Y}^n$ are drawn from the joint distribution $P_{\underline{X}^n, \underline{Y}^n}$.
- For E_i with $i \geq 2$, $\underline{X}^n, \underline{Y}^n$ are drawn from independent marginal distributions, i.e. $\underline{X}^n, \underline{Y}^n \sim P_{\underline{X}^n}(\underline{x}^n) P_{\underline{Y}^n}(\underline{y}^n)$.

Think about joint AEP

Using Joint AEP

Fixed $\epsilon > 0$, for large enough n,

$$P(E_1^c) \leq \epsilon$$

for $i \geq 2$,

$$P(E_i) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

The probability of error

$$\begin{aligned} P(\text{error}) &\leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\ &\leq \epsilon + 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)} \end{aligned}$$

If $R < I(X;Y) - 3\epsilon$, we can choose n large enough such that the second term is arbitrarily small.

Summary

- The average performance of the random code is "good", so there exists a "good" code.
- The above can be derived for any input distribution, so pick P_X^* that maximizes the mutual information, as long as $R \leq C = \max_{P_X} I(X;Y)$, the rate is achievable.

Remaining Questions

- Can we reliably transmit any $R > C$?
- Long codewords are good, but how long. What is the performance for finite codeword length.
- Average error probability is small, but may have particularly bad codewords.
- Random codes seems good, but how do we decode? Is there any better structured code that can do as well? The randomness in coding is a computational trick or a fundamental requirement to achieve the capacity?

Converse of Coding Theorem

Recall Fano's inequality: For any r.v.'s X, Y, we try to guess X by $\hat{X} = g(Y)$. The error probability $P_e = P(X \neq \hat{X})$ satisfies

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

Proof by applying chain rule on $H(X, E|Y)$.

$$\begin{aligned}
H(E, X|Y) &= H(E|X, Y) + H(X|Y) = H(X|Y) \\
H(E, X|Y) &= H(X|E, Y) + H(E|Y) \\
&\leq H(X|E, Y) + H(E) \\
&\leq P_e H(X|Y, E = 1) \\
&\quad + (1 - P_e) H(X|Y, E = 0) + H(E) \\
&= P_e H(X|Y, E = 1) + H(E) \\
&\leq P_e \log(|\mathcal{X}| - 1)
\end{aligned}$$

Converse of Coding Theorem

Now consider trying to guess the message W based on the observation on $\underline{Y}^n$.

$$H(W|\underline{Y}^n) \leq 1 + P_e^{(n)} nR$$

$$\begin{aligned} nR &= H(W) = H(W|\underline{Y}^n) + I(W;\underline{Y}^n) \\ &\leq H(W|\underline{Y}^n) + I(\underline{X}^n(W);\underline{Y}^n) \\ &\leq 1 + P_e^n nR + I(\underline{X}^n;\underline{Y}^n) \\ &\leq 1 + P_e^n nR + nC \end{aligned}$$

Rearrange

$$P_e^n \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

- For n large enough, P_e is bounded away from 0.
- Suppose we can achieve $P_e^n = 0$ for some finite n, than we can concatenate such codes to have 0 error probability for large n, which gives contradiction.