EE 231A Information Theory
Lecture 8

Achievability in the Channel Coding Theorem

A. Preview: Every Channel is a Noisy Typewriter, Jointly Typical Sets
B. Random Codes and Typical Set Decoding
C. Probability of Error

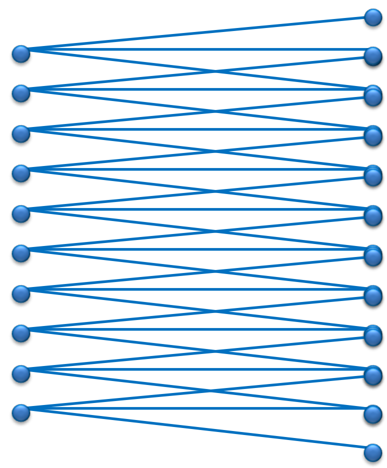
1

Part A: Preview:
Every Channel is a Noisy Typewriter
Jointly Typical Sets

2

Every Channel is a Noisy Typewriter.

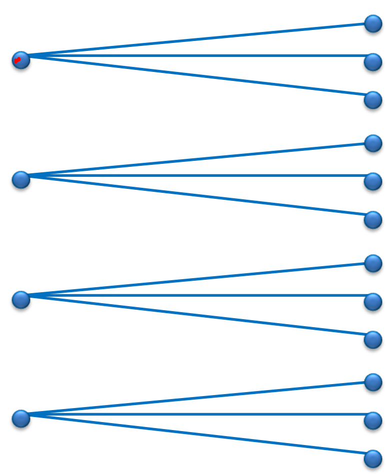
- For large block lengths, every channel looks like the noisy typewriter.



3

Code design = clever choice of inputs.

- Well-chosen inputs create essentially disjoint sets of outputs for each (well-chosen) input.



4

Jointly Typical Sequences

- The set $A_{\epsilon}^{(n)}$ of jointly typical sequences $\{x^n, y^n\}$ with respect to the distribution $p(x,y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:

$$A_{\epsilon}^{(n)} = \Bigg\{ (x^n, y^n): \quad \left| -\frac{1}{n}\log p(x^n) - H(X) \right| < \epsilon$$

$$\left| -\frac{1}{n}\log p(y^n) - H(Y) \right| < \epsilon$$

$$\left| -\frac{1}{n}\log p(x^n, y^n) - H(X,Y) \right| < \epsilon \Bigg\}$$

where $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$

5

$$\left| -\frac{1}{n}\log p(x^n) - H(X) \right| < \epsilon$$

$$-\frac{1}{n}\log p(x^n) - H(X) < \epsilon$$

$$\log p(x^n) + nH(X) > -n\epsilon$$

$$\log p(x^n) > -n\big(H(X) + \epsilon\big)$$

$$p(x^n) > 2^{-n\left(H(X)+\epsilon\right)}$$

6

$$\left| -\frac{1}{n}\log p(x^n) - H(X) \right| < \epsilon$$
$$\frac{1}{n}\log p(x^n) + H(X) < \epsilon$$
$$\log p(x^n) + nH(X) < n\epsilon$$
$$\log p(x^n) < -n\left(H(X) - \epsilon\right)$$
$$p(x^n) < 2^{-n\left(H(X)-\epsilon\right)}$$

7

Joint AEP (Thm 7.6.1)

- Let $\left(X^n, Y^n\right)$ be sequences of length n drawn according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. Then

1. $p\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1 \quad \text{as } n \to \infty$
2. $\left|A_\epsilon^{(n)}\right| \le 2^{n\left(H(X,Y)+\epsilon\right)}$
3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ (i.e. $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$) then $p\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \le 2^{-n(I(X;Y)-3\epsilon)}$.

8

Proof of 3

$$p\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n) p(y^n)$$

$$\leq \left| A_\epsilon^{(n)} \right| 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)}$$

$$\leq 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)}$$

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

9

How many typical sequences?

- $\sim 2^{nH(X)}$ typical X sequences
- $\sim 2^{nH(Y)}$ typical Y sequences
- $\sim 2^{nH(X,Y)}$ typical (X,Y) sequences
- Not all pairings of a typical X sequence with a typical Y sequence produce a typical (X,Y) sequence.
- In fact, when the X and Y sequences are chosen independently, the probability is $\sim 2^{-nI(X;Y)}$

10

Part B:

Random Codes and Typical Set Decoding

Proof of achievability in channel coding theorem

- Now we prove part 1 of channel coding theorem: All rates R<C are achievable.

Random Code Generation

- Fix $p(x)$. Generate a $(2^{nR}, n)$ code $\mathcal{C}$ at random according to $p(x)$.
- Specifically, we generate 2^{nR} codewords according to $p(x^n) = \prod_{i=1}^{n} p(x_i)$.

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix} \qquad p(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w))$$

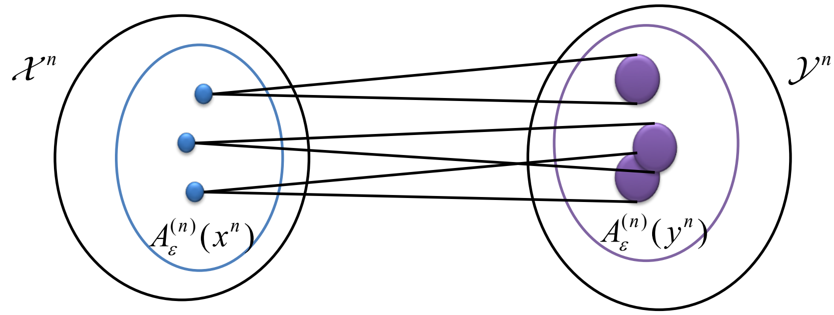
13

Typical Set Decoding

- Maximum likelihood decoding is optimal, but we will use decoding based on typical sets because it makes our analysis easier:
- Receiver declares $\hat{W}$ was sent if
  - 1) $\left(X^n(\hat{W}), Y^n\right) \in A_{\epsilon}^{(n)}$
  - 2) There is no other index k with $\left(X^n(k), Y^n\right) \in A_{\epsilon}^{(n)}$.
- If no such $\hat{W}$ exists declare an error.

14

Typical Set Decoding



- Typical set decoding:

$$\hat{W}(Y^n) = \begin{cases} W_i \text{ if } \left(X^n(W_i), Y^n\right) \in A_\varepsilon^{(n)} \text{ and } \left(X^n(W_j), Y^n\right) \notin A_\varepsilon^{(n)} \text{ for } j \neq i \\ \text{decoding failure otherwise} \end{cases}$$

15

The Game Plan

- We will compute the average probability of error, averaging over all random codes, under typical set decoding.
- This probability of error will converge to zero as $n \to \infty$ guaranteeing the existence of at least one good code.

16

Part C: Computing probability of error

Probabilities of error

- Probability of error for a specified input:

$$\lambda_i = P(\hat{W} \neq i \mid x^n = x^n(i))$$

- Maximal probability of error $\lambda^{(n)} = \max_{i \in \{1,2,\ldots,2^{nR}\}} \lambda_i$

- Average prob. of error for a specified $\mathcal{C}$

$$P_e^{(n)}(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_W(\mathcal{C})$$

- Average prob. of error over all possible $\mathcal{C}$'s.

$$P(\mathcal{E}) = \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C})$$

We only need to consider a single input.

$$
\begin{aligned}
P(\mathcal{E}) &= \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_W(\mathcal{C}) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_W(\mathcal{C}) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \underbrace{\sum_{\mathcal{C}} P(\mathcal{C}) \lambda_W(\mathcal{C})}_{\text{same for every } W} \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) \\
&= P(\mathcal{E} \mid W = 1)
\end{aligned}
$$

19

Symmetry of codes

- The symmetry of the code construction guarantee s that for every code $\mathcal{C}$, there are $2^{nR}!$ equally likely codes that are simply permutations (same codewords, different mapping of inputs to codewords).
- Considering a single index (input) over all $2^{nR}!$ permutations provides the same average as considering all indices over all permutations. (or all indices on a single permutation).

20

Example of why one input is sufficient.

- Let $\mathcal{X} = \{a,b,c,d\}$ $n = 1$ $2^{nR} = 3$
- There are 24 possible codes $\mathcal{C}$, each with probability 1/24.

$$\left.\begin{matrix}[a\ b\ c]^T\\ [a\ c\ b]^T\\ [b\ a\ c]^T\\ [b\ c\ a]^T\\ [c\ a\ b]^T\\ [c\ b\ a]^T\end{matrix}\right\}\bar{d} \quad \left.\begin{matrix}[b\ c\ d]^T\\ [b\ d\ c]^T\\ [c\ b\ d]^T\\ [c\ d\ b]^T\\ [d\ b\ c]^T\\ [d\ c\ b]^T\end{matrix}\right\}\bar{a} \quad \left.\begin{matrix}[a\ c\ d]^T\\ [a\ d\ c]^T\\ [c\ d\ a]^T\\ [c\ a\ d]^T\\ [d\ a\ c]^T\\ [d\ c\ a]^T\end{matrix}\right\}\bar{b} \quad \left.\begin{matrix}[a\ b\ d]^T\\ [a\ d\ b]^T\\ [b\ a\ d]^T\\ [b\ d\ a]^T\\ [d\ a\ b]^T\\ [d\ b\ a]^T\end{matrix}\right\}\bar{c}$$

Discussion of Example

- We have divided these codes into 4 sets that are mutually exclusive and collectively exhaustive.
- Each set contains all the permutations that use the same set of codewords.
- Within each permutation, $\lambda_i = \lambda_j$ for all i, j.
- We can do this for any set of random codes used in the channel coding theorem proof.
- The codeword error $\lambda_W(\mathcal{C})$ is a only a function of the codeword for W and the set of other codewords in $\mathcal{C}$.

Error Events and Correct Transmission

- Define the following events:

$$E_i = \{(X^n(i), Y^n) \in A_\epsilon^{(n)}\}, \quad i = 1, \ldots, 2^{nR}$$

- E_i is the event that the i^{th} codeword is typical with the received sequence.

- E_i is an error unless i=1.

Computing Probability of Error

$$\begin{aligned} P(\mathcal{E} \mid W = 1) &= P(E_1^c \cup E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}}) \\ &\le P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \end{aligned}$$

$$P(E_1^c) \to 0 \quad \text{as } n \to \infty$$

$$P(E_i) \le 2^{-n(I(X;Y)-3\epsilon)} \quad \text{for } i \ne 1$$

Error goes to zero if $R<I$.

$$
\begin{aligned}
P(\mathcal{E}) &= P(\mathcal{E} \mid W=1) \\
&\le P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\
&\le \text{Ú} + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\text{Ú})} \quad \text{for } n \text{ sufficiently large} \\
&= \text{Ú} + (2^{nR}-1)2^{-n(I(X;Y)-3\text{Ú})} \\
&\le \text{Ú} + 2^{-n(I(X;Y)-(R+3\text{Ú}))} \\
&\le 2\text{Ú} \quad \text{for } R < I(X;Y) - 3\text{Ú}
\end{aligned}
$$

- If $R<I(X;Y)$ we can always choose ϵ and n so that the average probability is less than 2ϵ, i.e. arbitrarily small.

25

Error goes to zero if $R<I$.

$$
\begin{aligned}
P(\mathcal{E}) &= P(\mathcal{E} \mid W=1) \\
&\le P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\
&\le \epsilon + \textstyle\sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \quad \text{for } n \text{ sufficiently large} \\
&= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\
&\le \epsilon + 2^{-n}\big(I(X;Y) - (R+3\epsilon)\big) \\
&\le 2\epsilon \quad \text{for } R < I(X;Y) - 3\epsilon
\end{aligned}
$$

- If $R<I(X;Y)$ we can always choose ϵ and n so that the average probability is less than 2ϵ, i.e. arbitrarily small.

26

Maximizing I gives C.

- Now choose $p(x)$ to maximize $I(X;Y)$ and everything we did works for C in place of $I(X;Y)$.

27

At least one good code…

- Since the average probability of error overall codes is small <2ϵ, there exists at least one code C^* that achieves $P_e^{(n)} < 2$Ú by itself.
- For C^*, discard the half of the codewords with the largest λ_i's.
- Because the average of the λ_i's is 2ϵ, the best half must have all their λ_i's less than 4ϵ.
- Hence the best half of C^* has maximal $\lambda^{(n)} < 4$Ú. We have reduced rate negligibly from R to $R - \frac{1}{n}$.

28