Tufts University
Electrical and Computer Engineering

EE194 – Network Information Theory
Prof. Mai Vu

Lecture 2: Channel Capacity

1 Typicality and Joint Typicality

The Asymptotic Equipartition Property (AEP) in information theory is the analog of the Law of Large Numbers in probability theory. Assume that the binary random variable X has a probability mass function defined by $p(1) = 2/3$, and $p(0) = 1/3$, where X_1,X_2,...,X_n are *i.i.d.* random variables according to $p(x)$, and we want to observe a realization of the sequence X_1,X_2,...,X_n, where $n \longrightarrow \infty$. We will see that as n increases, the number of zeroes and ones in the sequence would be very close to $n/3$, and $2n/3$, respectively.

1.1 Typical Sets

We want to answer the following question: if X_1,X_2...,X_n are *i.i.d.* random variables according to $p(x)$, what is the probability of a sequence $(x_1, x_2..., x_n)$ to occur as n goes to infinity? This will lead us to divide the set of the sequences χ^n into two sets, the typical set, which contains the sequences which are very likely to occur, and the non-typical set which contains all the other sequences.

1.1.1 Theorem (AEP)

Theorem Let X_1,X_2,..., be *i.i.d.* random variables according to $p(x)$, then in probability

$$-\frac{1}{n}\log p(X_1, X_2, ..., X_n) \longrightarrow H(X)$$

Proof Based on the weak law of large number (WLLN), for i.i.d Z_i, $\frac{1}{n}\sum_{i=1}^{n} Z_i \longrightarrow E[Z]$ in probability. Specifically

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\right| \leq \epsilon\right] \longrightarrow 1 \; \forall\epsilon > 0, as\ n \longrightarrow \infty.$$

In other words, the average of realizations of *i.i.d.* random variables X_i converges in probability towards the Expected Value.

Then

$$-\frac{1}{n}\log p(X_1, X_2, ..., X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log(p(X_i)) \longrightarrow E[\log(p(X))] = H(X).$$

Definition Typical set $A_\epsilon^{(n)}$

The *typical set* $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, ..., x_n) \in \chi^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, ..., x_n) \leq 2^{-n(H(X)-\epsilon)}$$

We denote the set $A_\epsilon^{(n)}$ as an ϵ*-typical set* with respect to $p(x)$, and we have,

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

1.1.2 Properties of $A_\epsilon^{(n)}$

1. Probability of the typical set

$$\Pr\{A_\epsilon^{(n)}\} \longrightarrow 1, \text{as } n \longrightarrow \infty$$

2. Number of sequences in the typical set

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$$

1.2 Jointly Typical Sets

In communication we have a channel input sequence X^n, and a channel output sequence Y^n. We decode Y^n as the ith index if the codeword $X^n(i)$ is *jointly typical* with the received signal Y^n. Here we define the idea of joint typicality.

Definition Jointly typical set

The set $A_\epsilon^{(n)}$ of *jointly typical* sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is defined as

$$\begin{aligned} A_\epsilon^{(n)} = \{(x^n, y^n) : 2^{-n(H(X)+\epsilon)} &\leq p(x_1, x_2, ..., x_n) \leq 2^{-n(H(X)-\epsilon)} \\ 2^{-n(H(Y)+\epsilon)} &\leq p(y_1, y_2, ..., y_n) \leq 2^{-n(H(Y)-\epsilon)} \\ 2^{-n(H(X,Y)+\epsilon)} &\leq p(x^n, y^n) \leq 2^{-n(H(X,Y)-\epsilon)}.\} \end{aligned}$$

where,

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

1.2.1 Joint AEP theorem

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

$$\Pr\{A_\epsilon^{(n)}\} \longrightarrow 1, as\ n \longrightarrow \infty.$$

The size of the jointly typical set

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

If $(\tilde{X}^n, \tilde{Y}^n) \sim p(X^n)p(Y^n)$, then

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

Also for sufficiently large n,

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

Proof. The first two parts could be proved as for the AEP theorem. For the third part we have:

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

For sufficiently large n, $Pr(A_\epsilon^{(n)}) \geq 1-\epsilon$, and therefore

$$\begin{aligned} 1-\epsilon &\leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} P(x^n)P(y^n) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \end{aligned}$$

and

$$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}.$$

By similar arguments to the upper bound above, we can also show that for n sufficiently large,

$$\begin{aligned} pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n) &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} P(x^n)P(y^n) \\ &\geq (1-\epsilon) 2^{n(H(X,Y)-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(Y)+\epsilon)} \\ &= (1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)}. \end{aligned}$$

□

1.2.2 Intuition for Joint Typicality

We know from the *joint AEP theorem*, that there are only about $2^{nH(X,Y)}$ jointly typical sequences, whereas, there are about $2^{nH(X)}$ typical X sequences and about $2^{nH(Y)}$ typical Y sequences. This shows that not all pairs of typical X^n and typical Y^n are also jointly typical. This is because of the fact that $H(X,Y) \leq H(X) + H(Y)$.

Now assume that a specific X^n is given. For this X^n we can search through all the $2^{nH(Y)}$ typical Y sequences to find those which are jointly typical with X^n. For this sequence X^n, there are about $2^{nH(Y|X)}$ conditionally typical Y sequences.The probability that some randomly chosen signal Y^n is jointly typical with X^n is about $\frac{2^{nH(Y|X)}}{2^{nH(Y)}} = 2^{-nI(X;Y)}$. This suggests that we can search about $2^{nI(X;Y)}$ typical Y sequences in order to find one which is jointly typical with X^n.

2 Definitions

Channel capacity is a measure of maximum information per channel usage one can get through a channel. This one of the fundamental concepts in information theory.

Definition 1 A *Discrete channel*, denoted by ($\mathcal{X}$, $p(y|x)$, $\mathcal{Y}$), consists of two finite alphabet sets, $\mathcal{X}$ and $\mathcal{Y}$, and a conditional probability mass function $p(y|x)$, where $\mathcal{X}$ is the input set, $\mathcal{Y}$ is the output set, $p(y|x)$ is the channel transition matrix.

For each input $x_i \in \mathcal{X}$ to the channel, the output can be one of a number of posibilities $\{y_j\}$, each with probability $p(y_j|x_i)$.

The channel is said to be *memoryless* if the current output is conditionally independent of previous channel inputs and outputs, given the current input, i.e., it depends only on the current input. That is

$$p(Y_i|X^i, Y^{i-1}) = p(Y_i|X_i).$$

A discrete memoryless channel (DMC) is both discrete and memoryless.

Definition 2 An (M, n) code for a channel ($\mathcal{X}$, $p(y|x)$, $\mathcal{Y}$) consists of

1. A message set $\{1, 2, \ldots, M\}$,
2. An encoding function $X^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}$ which generates M codewords, $x^n(1), x^n(2), \ldots, x^n(M)$, and
3. A decoding function $g : \mathcal{Y} \rightarrow \{1, 2, \ldots, M\}$, $g(Y^n) = i$.

The set of codewords is called the codebook $\mathcal{C}$ =$\{x^n(1), x^n(2), \ldots, x^n(M)\}$.

Definition 3 *Probability of error* There are three several definitions for the probability of error

Conditional probability of error: The conditional probability of error given that index i was sent is defined as

$$\lambda_i = Pr\{g(Y^n) \neq i | X^n = X^n(i)\}. \tag{1}$$

Maximum probability of error: The maximum probability of error for an (M, n) is defined as

$$\lambda^{(n)} = \max \quad \lambda_i. \tag{2}$$

Average probability of error: The average probability of error e for an (M, n) code is defined as

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i. \tag{3}$$

Also, if W is chosen according to a uniform distribution over the set $\{1, 2, \ldots, M\}$, we can define $P_e^{(n)} = Pr\{g(Y^n) \neq W\}$.

Definition 4 The *rate* R of an (M, n) code is $R = \frac{\log_2 M}{n}$ *bits per transmission.*

Definition 5 A rate R is *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that the maximal probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$.

Definition 6 The *capacity* of a channel is the supremum of all achievable rates.

3 Channel coding theorem

Theorem. *The capacity of a DMC ($\mathcal{X}$, $p(y|x)$, $\mathcal{Y}$) channel is given by*

$$C = \max_{p(x)} I(X;Y), \tag{4}$$

where the maximum is taken over all possible input distributions.

By definition, codes with rate less than capacity $R < C$ *can* yield arbitrarily small probability of error for sufficiently large block lengths while for $R > C$ the probability of error is bounded away from 0. We will prove this theorem in the next section, but first lets look at some examples.

3.1 Examples

Example 1 *Four letter noisy typewriter*

We consider a four letter Noisy typewriter. In this case we have an input alphabet of 4 letters (A, B, C, and D) and each letter is either printed out correctly or changed to the next letter with probability $\frac{1}{2}$.

The capacity of this channel is

$$\begin{aligned} C = & \quad \max \quad I(X;Y) \\ = & \quad \max \big(H(Y) - H(Y|X)\big) \\ = & \quad \max H(Y) - H\left(\frac{1}{2}\right) \\ = & \quad \log|\mathcal{Y}| - 1 \\ = & \quad 1 \quad bits. \end{aligned}$$

A simple code that achieves capacity for this channel is to use either $\{A, C\}$ or $\{B, D\}$ as input letters so that no two letters can be confused. In each case, there are 2 codewords of block length 1. If we choose the codewords *i.i.d.* according to a uniform distribution on $\{A, C\}$ or $\{B, D\}$, then the output of the channel is also *i.i.d.* and uniformly distributed on $\{A, C, B, D\}$.

Example 2 *Binary Symmetric Channel (BSC)*

In A *BSC*, given any input sequence, every possible output sequence has some positive probability, so it will not be possible to distinguish even two codewords with zero probability of error. Hence the *zero-error* capacity of the BSC is zero.

The capacity of this channel is

$$\begin{aligned} C = & \quad \max \quad I(X;Y) \\ = & \quad \max \big(H(Y) - H(Y|X)\big) \\ = & \quad \max H(Y) - H(p) \\ = & \quad \log|\mathcal{Y}| - H(p) \\ = & \quad 1 - H(p) \quad bits. \end{aligned}$$

we can also model this channel as $Y = X \oplus Z$ where $Z \sim Bernoulli(p)$.

Example 3 *Binary Erasure Channel (BEC)*

The binary erasure channel (BEC) is the analog of BSC in which some bits are lost rather than being corrupted. The binary erasure channel is

$$\begin{aligned} C = & \quad \max \quad I(X;Y) \\ = & \quad \max \big(H(X) - H(X|Y\big) \\ = & \quad \max \left(H(X) - \alpha H(X)\right) \\ = & \quad \log|\mathcal{X}|(1-\alpha) \\ = & \quad 1 - \alpha \quad bits. \end{aligned}$$

3.2 Properties of the Capacity

1. $C \geq 0$ since $I(X;Y) \geq 0$.
2. $I(X;Y) \leq \log|\mathcal{X}|$ since $C = \max I(X;Y) \leq \max H(X) = \log|\mathcal{X}|$.
3. Similarly $I(X;Y) \leq \log|\mathcal{Y}|$.
4. $I(X;Y)$ is a continuous function of $p(x)$.
5. $I(X;Y)$ is a concave function of $p(x)$.

   Since $I(X;Y)$ is a concave function, a local maximum is a global maximum. This maximum can then be found by standard convex optimization techniques.

4 Proof of the channel coding theorem

To prove the capacity, we have to prove its achievability and converse. Achievability means that for a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exits a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$. Converse means that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

The proof uses the idea of random coding and joint typicality.

4.1 Achievability

Code Construction

Fix input distribution $p(x)$, generate a random code $X^n(\omega)$ such that each code word $X_i(\omega)$ with $p(x_i)$ i.i.d. $X^n(\omega) = (X_1(\omega), X_2(\omega), \cdots, X_n(\omega))$. We exhibit the $(2^{nR}, n)$ codewords as the matrix:

$$\text{code book}\mathcal{C} = \begin{pmatrix} X^n(1) \\ X^n(2) \\ \dots \\ X^n(2^{nR}) \end{pmatrix} = \begin{pmatrix} X_1(1) & X_2(1) & \dots & X_n(1) \\ X_1(2) & X_2(2) & \dots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \dots & X_n(2^{nR}) \end{pmatrix}$$

Thus the probability of generating this codebook:

$$\Pr(\mathcal{C}) = \prod_{\omega=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(\omega))$$

This codebook is revealed to boh sender and receiver:

Encoding

To send message ω, the transmitter sends codeword $X^n(\omega)$ over n channel uses.

Decoding

Using jointly typically decoding, the decoder searches the codebook and find

$$(X^n(i), Y^n) \in A_\epsilon^n.$$

If exactly one $(X^n(i), Y^n) \in A_\epsilon^n \rightarrow \widehat{\omega} = i$. Otherwise if either

(1) no codeword is jointly typical with Y^n or

(2) more than 1 codeword is jointly typical with Y^n

then set $\widehat{\omega} = 0$. Decoder error if $\widehat{\omega} \neq \omega$.

Error analysis

Define event that the ith codeword and Y^n are jointly typical

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

Suppose that message W is uniform distributed. Then by the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index that was sent. Thus we can assume, the message $W = 1$, the error occurs if either E_1^c occurs or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs. Probability of error is

$$P_e^n = \Pr(\widehat{W} \neq 1) = Pr(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} \leq P(E_1^c) + \sum_{k=2}^{2^{nR}} Pr(E_k)$$

Now, by the joint AEP, when $n \to \infty$,

$$\Pr((X^n, Y^n) \in A_\epsilon^n) \geq 1 - \epsilon$$

$$\Pr((X^n, Y^n) \notin A_\epsilon^n) \leq \epsilon$$

What we didn't send happens to be jointly typical with received Y^n with probability $\leq 2^{-n(I(X;Y)-3\epsilon)}$.

$$\begin{aligned} P_e^n &\leq P(E_1^c) + \sum_{k=2}^{2^{nR}} Pr(E_k) \\ &\leq \epsilon + \sum_{k=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + 2^{-n((I(X;Y)-R+3\epsilon))} \end{aligned}$$

Thus if n is sufficient large and $R \leq I(X;Y)$, then $P_e^n \to 0$ for any $p(x)$. Because of uniform distribion, here the average error probability is the same as maximum error probability.
Choose $p(x)$ to maxmize $I(X;Y)$ we get channel capacity C, thus $R \leq C$ is achievable.

Question: So far we examine the *evarage* error probablity with W uniform. For general case with any distribution of W, what about the *maximum* error probabilty when satisfying $R \leq C$?

First examine error probability averaged over all the codebooks and all the codewords in each codebook. Because of the symmetry over all codebooks and codewords, this average error probability is the same as the probability of sending $W = 1$ as analyzed above.

$$\begin{aligned} P_e &= \sum \Pr(\text{codebook}) \frac{1}{M} \sum_{i=1}^{M} \lambda_i(\text{codeword}) \\ &= \sum \Pr(\text{codebook}) P_e^n(\text{codeword}) \\ &= \Pr(\text{error}|W = 1) \leq \epsilon \end{aligned}$$

Then

$$P_e \leq \epsilon \Rightarrow \exists P_e(\text{one good codebook}) \leq \epsilon.$$

Since the average error probabilty is less than ϵ, at least half of the codewords must have a maximal probability of error less than 2ϵ. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n.

Thus for codebooks generated randomly with average error probability almost zero, there is at least one good codebook with maximum error probability almost zero.

4.2 Converse

We want to show if $(2^{nR}, n)$ is achievable $(\lambda^{(n)} \to 0)$, then $R \leq C$.

4.2.1 For zero error probability

We will first prove that $P_e^n = 0$ implies that $R \leq C$. $\lambda^{(n)} = 0, Y^n \to W(message)$, which means when Y is received, we know exactly what is sent. Thus $H(W|Y^n) = 0$. Assume $W \in \{1, 2, \cdots, 2^{nR}\}$ with uniform probability distribution.

$$\begin{aligned}
H(W) &= H(W) - H(W|Y^n) \\
&= I(W; Y^n) \\
&\leq I(X^n; Y^n) \\
&= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum H(Y_i|X^n, Y^{i-1}) \\
&= H(Y^n) - \sum H(Y_i|X_i) \\
&\leq \sum H(Y_i) - \sum H(Y_i|X_i) \\
&= \sum I(X_i, Y_i) \leq nC
\end{aligned}$$

Where (15) → (16) comes from the data processing inequality; (18) → (19) comes from the assumption of memoryless channel without feedback; (19) → (20) comes from $H(Y_i|Y^{i-1}) \leq H(Y_i)$.

4.2.2 For vanishing (but non-zero) errror probability

Continue to prove for the case of vanishing (but non-zero) error.

Fano's inequality: If $X \to Y \to \hat{X}$ forms a Markov chain then

$$H(X|\widehat{X}) \leq H(P_e) + P_e \log |X| \tag{5}$$

where $P_e = \Pr(\hat{X} \neq X)$ is the probability of error for the estimator $\hat{X}$.

Proof. Define error event:

$$E = \begin{cases} 1, & \widehat{x} \neq x \\ 0, & \widehat{x} = x \end{cases}$$

We have:

$$H(X, E|\widehat{X}) = H(X|\widehat{X}) + H(E|X, \widehat{X}) = H(E|\widehat{X}) + H(X|\widehat{X}, E)$$

$H(E|X, \widehat{X})$ would be zero since we know both input and output.
Since $P_e = \Pr(\widehat{X} \neq X) = \Pr(E = 1)$, we can get

$$\begin{aligned} & H(E|\widehat{X}) \leq H(E) = H(P_e) \\ & H(X|\widehat{X}, E) = \Pr(E = 0)H(X|\widehat{X}, E = 0) + \Pr(E = 1)H(X|\widehat{X}, E = 1) \end{aligned}$$

Since $H(X|\widehat{X}, E = 0) = 0$,we have:

$$H(X|\widehat{X}, E) \leq H(X) \leq \log |X|$$

Putting all together

$$H(X|\widehat{X}) \leq H(P_e) + P_e \log |X|$$

From the Data Processing Inequality, we have:

$$\begin{aligned} & I(X;\widehat{X}) \leq I(X;Y) \\ \Rightarrow & H(X|Y) \leq H(X|\widehat{X}) \\ \Rightarrow & H(X|Y) \leq H(P_e) + P_e \log |X| \leq 1 + P_e \log |x| \end{aligned}$$

where the last inequality comes from $P_e = \Pr(\widehat{X} \neq X) = \Pr(E = 1)$ and E is a binary random variable. □

Converse to the channel coding theorem

Apply the above inequality to channel coding: $W \to X \to Y \to \widehat{W}$ where X is a codeword in a $(2^{nR}, n)$ code, and set $P_e = \Pr(\widehat{W} \neq W)$; we have:

$$H(W|\widehat{W}) \leq 1 + P_e \log |W| = 1 + P_e \cdot nR$$

where $W \sim \text{Uniform}\{1, 2, \ldots, 2^{nR}\}$, and we obtain

$$\begin{aligned} nR = H(W) &= H(W|\widehat{W}) + I(W;\widehat{W}) \\ &\leq 1 + P_e \cdot nR + I(X^n;Y^n) \\ &\leq 1 + P_e \cdot nR + nC \\ \Rightarrow R &\leq \frac{1}{n} + P_e \cdot R + C \end{aligned}$$

If $P_e \to 0$, then $R \leq C$.
Note also $P_e \geq (1 - \frac{C}{R} - \frac{1}{nR})$, thus if $R > C$ then P_e is bounded away from zero as $n \to \infty$.