INFORMATION THEORY & CODING

Channel Coding - 1

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

November 14, 2023

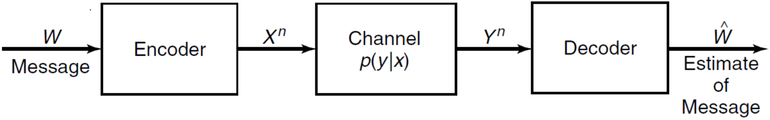
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Outline

- **Channel model**: conditional distribution
- **Channel capacity**: defined in a pure way of information theory, not operational
- **Channel coding & data rate**: operational indicator of channel

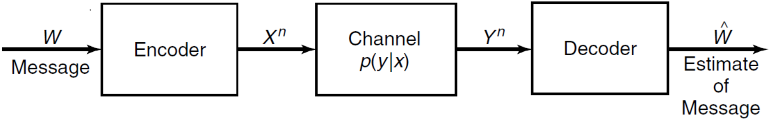
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Communication System Model



- $X^n = [X_1, X_2, \ldots, X_n]$
- $Y^n = [Y_1, Y_2, \ldots, Y_n]$
- Channel $p(y^n|x^n)$: probability of observing y^n given input input sequence x^n

Discrete memoryless channel (DMC)



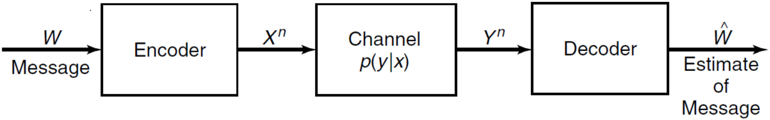
Definition

A discrete channel consists of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y^n|x^n)$ that expresses the probability of observing the output sequence y^n given that we send the sequence x^n.

Definition

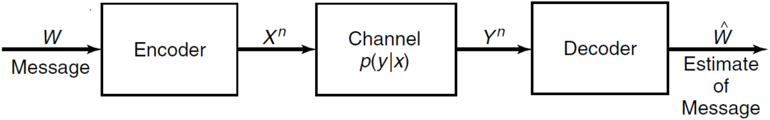
The channel is called memoryless if $p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$.

Communication System Model



- $X^n = [X_1, X_2, \ldots, X_n] \in \mathcal{X}^n$, $Y^n = [Y_1, Y_2, \ldots, Y_n] \in \mathcal{Y}^n$
  Channel $p(y^n|x^n)$: probability of observing y^n given input symbol x^n
  Memoryless: $p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$
- Messages are mapped into some sequence of the channel symbols. Output sequence is random but has a distribution that depends on the input sequences. Each possible input sequence may induce several possbile outputs, and hence inputs are confusable. Can we choose a *non-confusable* subset of input sequences?

Duality



- Data compression: we remove all the redundancy in the data to form the most compressed version possible.

- Data transmission: we add redundancy in a controlled manner to combat errors in the channel.

"Survivor"

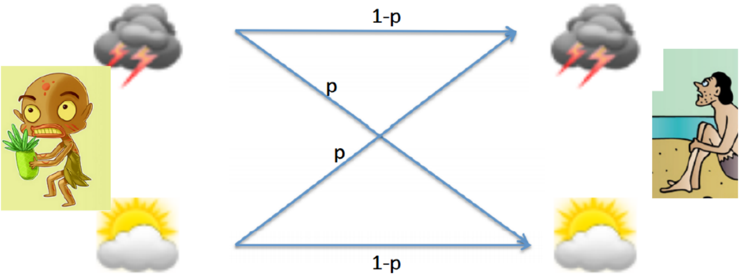
- You were deserted on a small island. You met a native and asked about the weather.
- True weather is a random variable X

$$X = \begin{cases} \text{rain} & \text{w.p. } \alpha, \\ \text{sunny} & \text{w.p. } 1-\alpha, \end{cases}$$

- Native knows tomorrow's weather perfectly, but only tells truth with probability $1-p$.
- Native's answer is a random variable $Y \in \{\text{rain, sunny}\}$.

"Survivor"

- How informative is the native's answer?



What is $I(X;Y)$?

- $I(X;Y) = H(X) - H(X|Y)$
- $H(X) = H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$
- $H(X|Y) = H(X|Y = \text{rain})p(\text{rain}) + H(X|Y = \text{sunny})p(\text{sunny})$
- $H(X|Y = \text{rain})$ is equal to $-\sum_{i\in\{\text{rain},\text{sunny}\}} p(X = i|Y = \text{rain}) \log p(X = i|Y = \text{rain})$. Note that

  $$p(X = \text{rain}|Y = \text{rain}) = \frac{p(X=\text{rain}|Y=\text{rain})p(X=\text{rain})}{p(Y=\text{rain})} = \frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)}$$

  Thus, $H(X|Y) = \alpha H\left(\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)}\right) + (1-\alpha)H\left(\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)}\right)$
- $I(X;Y) = H(\alpha) - \alpha H\left(\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)}\right) - (1-\alpha)H\left(\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)}\right)$

Special Cases

- $I(X;Y) = H(\alpha) - \alpha H\left(\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)}\right) - (1-\alpha)H\left(\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)}\right)$
- Always telling the truth: $p = 0$

$$I(X;Y) = H(\alpha) - \alpha H(1) - (1-\alpha)H(0) = H(\alpha) \leq 1 \text{ bit}$$

- Telling truth half of the time: $p = 1/2$

$$I(X;Y) = H(\alpha) - \alpha H(\alpha) - (1-\alpha)H(\alpha) = 0 \text{ bit}$$

- Fix p, maximize with respect to α, maximum achieved when $\alpha = 1/2$

$$\max_{\alpha} I(X;Y) = H(1/2) - \frac{1}{2}H(1-p) - \frac{1}{2}H(p) = 1 - H(P)$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Special Cases

- $I(X;Y) = H(\alpha) - \alpha H\left(\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)}\right) - (1-\alpha)H\left(\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)}\right)$
- Always telling the truth: $p = 0$

$$I(X;Y) = H(\alpha) - \alpha H(1) - (1-\alpha)H(0) = H(\alpha) \leq 1 \text{ bit}$$

- Telling truth half of the time: $p = 1/2$

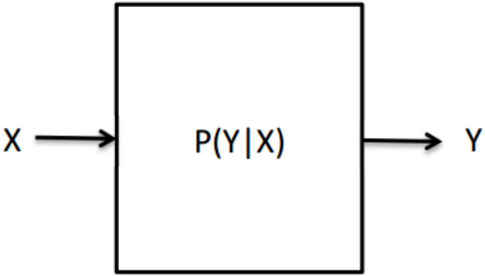
$$I(X;Y) = H(\alpha) - \alpha H(\alpha) - (1-\alpha)H(\alpha) = 0 \text{ bit}$$

- Fix p, maximize with respect to α, maximum achieved when $\alpha = 1/2$

$$\max_{\alpha} I(X;Y) = H(1/2) - \frac{1}{2}H(1-p) - \frac{1}{2}H(p) = 1 - H(P)$$

Definition ("Information" Channel Capacity)

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



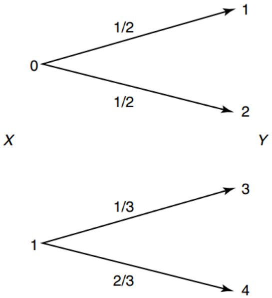
Examples

- Binary noiseless channel

$$C = \max I(X;Y) = \log 2 = 1 \text{ bits } \left(\text{with } p(x) = (\frac{1}{2}, \frac{1}{2})\right)$$

Examples

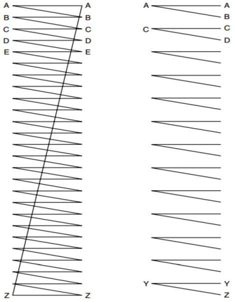
- Noisy channel with nonoverlapping outputs



$$C = \max I(X;Y) = \log 2 = 1 \text{ bits } \left(\text{with } p(x) = (\frac{1}{2}, \frac{1}{2})\right)$$

Examples

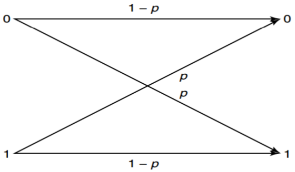
- Noisy typewriter



$$C = \max I(X;Y) = \log \frac{26}{2} = \log 13 \text{ bits } \left(\text{with } p(x) \text{ uniformly distributed}\right)$$

Examples

- Binary symmetric channel



CD-ROM read channel

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \{0,1\}} p(x) H(Y|X=x)$$

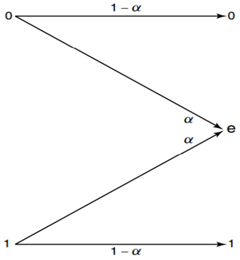
$$= H(Y) - \sum_{x \in \{0,1\}} p(x) H(p) = H(Y) - H(p) \leq 1 - H(p)$$

$$C = \max I(X;Y) = I - H(p) \text{ bits}$$

Examples

- Binary erasure channel

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} \Big(H(Y) - H(Y|X)\Big) \\
&= \max_{p(x)} H(Y) - H(\alpha)
\end{aligned}
$$



Let $\Pr[X = 1] = \pi$, then

$$H(Y) = H\Big((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha)\Big) = H(\alpha) + (1-\alpha)H(\pi)$$

Thus, $C = \max_{\pi}(1-\alpha)H(\pi) = 1-\alpha$ (with $\pi = \frac{1}{2}$)

Symmetric channel

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}.$$

All the rows of the transition matrix are permutations of each other and so are the columns. Let $\mathbf{r}$ be a row of the transition matrix.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathbf{r}) \leq \log|\mathcal{Y}| - H(\mathbf{r})$$

with equality if $\mathcal{Y}$ is uniformly distributed. If $p(x) = \frac{1}{|\mathcal{X}|}$, Y is also uniformly distributed:

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} p(y|x) = \frac{c}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|},$$

where c is the sum of the entries in one column.

Symmetric channel

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}.$$

All the rows of the transition matrix are permutations of each other and so are the columns. Let $\mathbf{r}$ be a row of the transition matrix.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathbf{r}) \le \log|\mathcal{Y}| - H(\mathbf{r})$$

with equality if $\mathcal{Y}$ is uniformly distributed. If $p(x) = \frac{1}{|\mathcal{X}|}$, Y is also uniformly distributed:

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} p(y|x) = \frac{c}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|},$$

where c is the sum of the entries in one column.

Fundamental question

- How fast can we transmit information over a channel?
- Suppose a source sends r messages per second, and the entropy of a message is H bits per message, information rate is $R = rH$ bits/second.
- Intuition: as R increases, error will increase.
- Surprisingly, Shannon showed error can approach to zero, as long as

$$R < C$$

INFORMATION THEORY & CODING

Channel Code - 2

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

November 21, 2023

Review

- **Channel capacity.** The logarithm of the number of distinguishable inputs is given by

$$C = \max_{p(x)} I(X;Y).$$

- **Examples**
  - Binary symmetric channel: $C = 1 - H(p)$
  - Binary erasure channel: $C = 1 - \alpha$
  - Symmetric channel: $C = \log|\mathcal{Y}| - H$ (row of trans. matrix)

Channel Code

Definition

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

1. An index set $\{1, 2, \ldots, M\}$ representing messages.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$. The set of codewords is called codebook.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

The rate R of an (M, n) code is

$$R = \frac{\log M}{n} \text{ bit per transmission}$$

On the other hand, we usually write

$$M = \lceil 2^{nR} \rceil$$

Channel Code

Definition

An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of :
1. An index set $\{1, 2, \ldots, M\}$ representing messages.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$. The set of codewords is called codebook.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

The rate R of an (M, n) code is

$$R = \frac{\log M}{n} \text{ bit per transmission}$$

On the other hand, we usually write

$$M = \lceil 2^{nR} \rceil$$

Performance Metric

- Conditional probability of error:

$$\lambda_i = \Pr[g(Y_n) \neq i | X^n = x^n(i)] = \sum_{y^n} p\Big(y^n | x^n(i)\Big) I\Big(g(y^n) \neq i\Big)$$

- Maximal probability of error: $\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$
- Decoding error probability: $\Pr[W \neq g(Y^n)] = \sum_i \lambda_i \Pr[W = i]$
- Arithmetic average probability of error:

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i, \quad P_e^{(n)} \leq \lambda^{(n)}$$

If W is uniformly distributed:

$$P_e^{(n)} = \Pr[W \neq g(Y^n)] \text{ Decoding error probability}$$

Achievable Rate

- A rate R is achievable,

  if there exists a sequence of codes with rate R and codeword length n, denoted as $(\lceil 2^{nR} \rceil, n)$, such that the maximal probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$.

  Recall that

  The rate R of an (M, n) code is

  $$R = \frac{\log M}{n} \text{ bit per transmission.}$$

- Joint typicality. Given two i.i.d. random variable sequences X^n and Y^n, the set of jointly typical sequences is

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

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

Joint AEP

- **Joint AEP** Let (X^n, Y^n) be the sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$, then:

1. $\Pr\left[(X^n, Y^n) \in A_\epsilon^{(n)}\right] \to 1$ as $n \to \infty$.
2. $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X,Y)+\epsilon)}$.
3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$\Pr\left[\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\epsilon^{(n)}\right] \le 2^{-n(I(X;Y)-3\epsilon)}.$$

Please refer to p196 for the proof (proof of Theorem 7.6.1)

Channel Coding Theorem

Theorem (Channel coding theorem)

For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.

Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Achievability: when $R < C$, there exists zero-error code.
Converse: zero-error codes must have $R \leq C$.

Random Codebook

- Generate a $(2^{nR}, n)$ code at random according to $p(x)$, where $p(x)$ is the capacity achieving distribution. The 2^{nR} are the rows of a matrix:

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix}.$$

  Each entry is generated i.i.d. according to $p(x)$.

- Encoding: map the message $w = \{1, 2, 3, \dots, 2^{nR}\}$ to codeword $[x_1(w), x_2(w), \dots, x_n(w)]$, i.e.

$$\mathcal{C} \to [x_1(w), x_2(w), \dots, x_n(w)] = x^n_{\mathcal{C}}(w), w = 1, 2, \dots, 2^{nR}$$

- We shall prove the average detection error probability (over all codebooks) tends to zero as n increase, which implies that there must exists one good codebook whose detection error probability tends to zero

Jointly Typical Decoding

- Decoding: finds the only $\hat{w}$ such that $(x^n_{\mathcal{C}}(\hat{w}), Y^n_{\mathcal{C}})$ is jointly typical.
- Decoding error: Suppose message 1 is sent to via codeword $x^n_{\mathcal{C}}(1)$ and $Y^n_{\mathcal{C}}$ is the received signal, the possible decoding error events include:
  - $(x^n_{\mathcal{C}}(1), Y^n_{\mathcal{C}})$ is not joint typical.
  - $(x^n_{\mathcal{C}}(i), Y^n_{\mathcal{C}})$ is joint typical $(i = 2, 3, \dots, 2^{nR})$.
- Idea of proof: According to joint AEP, since $x^n_{\mathcal{C}}(1)$ and $Y^n_{\mathcal{C}}$ are generated according to joint distribution $p(x^n, y^n)$, the chance of the first event is small. Moreover, since $Y^n_{\mathcal{C}}$ is generated independently of $x^n_{\mathcal{C}}(i)$, the total chance of the second event is also small.

Proof for achievability

- A message W is chosen according to a uniform distribution

$$\Pr[W = w] = 2^{-nR},$$

for $w = 1, 2, \ldots, 2^{nR}$. The w-th codeword $x_{\mathcal{C}}^n(w)$, corresponding to the w-th row of $\mathcal{C}$, is sent over the channel.

- The receiver receives a sequence $Y_{\mathcal{C}}^n$ according to the distribution according to the distribution

$$\Pr\left(y_{\mathcal{C}}^n | x_{\mathcal{C}}^n(w)\right) = \prod_{i=1}^{n} \Pr\left(y_{i,\mathcal{C}} | x_{i,\mathcal{C}}(w)\right),$$

and guesses which message was sent using jointly typical decoding.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Proof for achievability

- Let $\varepsilon = \{\hat{W}(Y^n) \neq W\}$ denote the error event, $\lambda_w(\mathcal{C})$ be the error probability of the w-th codeword of code $\mathcal{C}$. The average probability of error, over all codewords and all codebooks, is:

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

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}),$$

where $\sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C})$, $\forall w \neq 1$.

Proof for achievability

- Let $Y_{\mathcal{C}}^n$ be the received signal for $x_{\mathcal{C}}^n(1)$

$$e_i(\mathcal{C}) = \{(x_{\mathcal{C}}^n(i), Y_{\mathcal{C}}^n) \in A_\epsilon^{(n)}\}, i \in \{1, 2, \ldots, 2^{nR}\},$$

and $e_i^c(\mathcal{C}) =!e_i(C)$. Thus,

$$\Pr[\varepsilon] = \sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_1(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr\left[e_1^c(\mathcal{C}) \cup (\cup_{i=2}^{2^{nR}} e_i(\mathcal{C}))\middle| W = 1\right]$$

$$\leq \sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr[e_1^c(\mathcal{C})|W = 1] + \sum_{\mathcal{C}} \Pr(\mathcal{C}) \sum_{i=2}^{2^{nR}} \Pr[e_i(\mathcal{C})|W = 1]$$

$$= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr[e_1^c(\mathcal{C})|W = 1] + \sum_{i=2}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr[e_i(\mathcal{C})|W = 1]$$

Proof for achievability

$$\sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr[e_1^c(\mathcal{C})|W=1]$$

$$=\sum_{\mathcal{C}} \left(\prod_{i=1}^{2^{nR}} \Pr(x_{\mathcal{C}}^n(i))\right) \Pr[e_1^c(\mathcal{C})|W=1]$$

$$=\sum_{x_1^n} \sum_{\mathcal{C}:x_{\mathcal{C}}^n(1)=x_1^n} \prod_{i=1}^{2^{nR}} \Pr(x_{\mathcal{C}}^n(i)) \Pr(x_1^n \text{ and } Y^n \text{ are not joint typical}|W=1)$$

$$=\sum_{x_1^n} \Pr(x_1^n) \Pr(x_1^n \text{ and } Y^n \text{ are not joint typical}|W=1)$$

$$\times \sum_{\mathcal{C}:x_{\mathcal{C}}^n(1)=x_1^n} \prod_{i=2}^{2^{nR}} \Pr(x_{\mathcal{C}}^n(i))$$

$$=\sum_{x_1^n} \Pr(x_1^n) \Pr(x_1^n \text{ and } Y^n \text{ are not joint typical}|W=1)$$

$$=\Pr(X_1^n \text{ and } Y^n \text{ are not joint typical}|W=1) = \Pr(E_1^c|W=1)$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Proof for achievability

- Similarly,

$$\sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr[e_1(\mathcal{C})|W=1] = \Pr(X_i^n \text{ and } Y^n \text{ are joint typical}|W=1)$$
$$= \Pr(E_i|W=1)$$

- As a result,

$$\Pr[\varepsilon] \leq \Pr[E_1^c|W=1] + \sum_{i=2}^{2^{nR}} \Pr[E_i|W=1]$$

Proof for achievability

- By the joint AEP, $\Pr[E_1^c|W=1] \le \epsilon$ for n sufficiently large. By the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, so are Y^n and $X^n(i)$. Hence the probability that $X^n(i)$ and Y^n are jointly typical is $\le 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP.

$$
\begin{aligned}
\Pr[\varepsilon] &\le \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\
&= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\
&\le \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)} \\
&\le 2\epsilon \quad \text{for } R \le I(X;Y) - 4\epsilon \text{ and sufficiently large n}
\end{aligned}
$$

  Hence, if $R < I(X;Y)$, we can choose ϵ and n so that the average probability of error, over codebooks and codewords, is less than 2ϵ.

- Since $p(x)$ is the capacity achieving distribution, $R < I(X;Y)$ beacomes $R < C$.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Get rid of the average over codebooks. Since the average probability of error is $\leq 2\epsilon$, there exists at least one codebook $\mathcal{C}^*$ with a small average probability of error ($\Pr(\varepsilon|\mathcal{C}^*) \leq 2\epsilon$). Since we have chosen $\hat{W}$ according to a uniform distribution, we have

$$\Pr(\varepsilon|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*).$$

- Throw away the worst half of the codewords in the best codebook $\mathcal{C}^*$. We have $\Pr(\varepsilon|\mathcal{C}^*) \leq \frac{1}{2^{nR}} \sum \lambda_i(\mathcal{C}^*) \leq 2\epsilon$. This implies that at least half the indices i and their associated codewords $X^n(I)$ must have conditional probability of error $\lambda_i \leq 4\epsilon$. If we reindex the codewords, we have 2^{nR-1} codewords. The rate now is $R' = R - \frac{1}{n}$ with maximal probability of error $\lambda^{(n)} \leq 4\epsilon$.

Proof for the converse

- The index W is uniformly distributed on the set $\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$, and the sequence Y^n is related to W. From Y^n, we estimate the index W as $\hat{W} = g(Y^n)$. Thus, $W \to X^n(W) \to Y^n \to \hat{W}$ forms a Markov chain.

$$\text{Data processing inequality:} \quad I(W; \hat{W}) \leq I(X^n(W); Y^n)$$

Lemma (Fano's inequality)

For a discrete memoryless channel with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have

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

Proof for the converse

- The index W is uniformly distributed on the set $\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$, and the sequence Y^n is related to W. From Y^n, we estimate the index W as $\hat{W} = g(Y^n)$. Thus, $W \to X^n(W) \to Y^n \to \hat{W}$ forms a Markov chain.

$$\text{Data processing inequality:} \quad I(W; \hat{W}) \leq I(X^n(W); Y^n)$$

Lemma (Fano's inequality)

For a discrete memoryless channel with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have

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

Lemma

Let Y^n be the result of passing X^n through a discrete memoryless channel of capacity C. Then

$$I(X^n;Y^n) \le nC, \quad \textit{for all} \quad p(x^n).$$

Proof.

$$
\begin{aligned}
I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{memoryless} \\
&\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{independence bound} \\
&= \sum_{i=1}^{n} I(X_i|Y_i) \le nC
\end{aligned}
$$

Lemma

Let Y^n be the result of passing X^n through a discrete memoryless channel of capacity C. Then

$$I(X^n;Y^n) \leq nC, \quad \textit{for all} \quad p(x^n).$$

Proof.

$$\begin{aligned}
I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{memoryless} \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad \text{independence bound} \\
&= \sum_{i=1}^{n} I(X_i|Y_i) \leq nC
\end{aligned}$$

Proof for the converse

Proof.

Converse to channel coding theorem: Since W has a uniform distribution, we have

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \quad \text{Fano's inequality} \\ &\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) \quad \text{data-processing inequality} \\ &\leq 1 + P_e^{(n)} nR + nC \quad \text{Lemma 7.9.2} \end{aligned}$$

We obtain $R \leq \frac{1}{n(1+P_e^{(n)})} + \frac{C}{1+P_e^{(n)}} \to \frac{1}{n} + C$.

Letting $n \to \infty$, we have $R \leq C$.

□

Reading & Homework

- Reading: Chapter 7: 7.6-7.10
- Homework: Problems 7.15, 7.31.