Chapter 7

Channel Capacity

Peng-Hua Wang

Graduate Inst. of Comm. Engineering

National Taipei University
Chapter Outline

Chap. 7  Channel Capacity

7.1  Examples of Channel Capacity
7.2  Symmetric Channels
7.3  Properties of Channel Capacity
7.4  Preview of the Channel Coding Theorem
7.5  Definitions
7.6  Jointly Typical Sequences
7.7  Channel Coding Theorem
7.8  Zero-Error Codes
7.9  Fano’s Inequality and the Converse to the Coding Theorem
7.1 Examples of Channel Capacity
Channel Model

- Operational channel capacity: the bit number to represent the maximum number of distinguishable signals for $n$ uses of a communication channel.
  - In $n$ transmission, we can send $M$ signals without error, the channel capacity is $\log M/n$ bits per transmission.
- Information channel capacity: the maximum mutual information
- Operational channel capacity is equal to Information channel capacity.
  - Fundamental theory and central success of information theory.
Channel capacity

Definition 1 (Discrete Channel) A system consisting of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y|x)$.

Definition 2 (Channel capacity) The “information” channel capacity of a discrete memoryless channel is

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

where the maximum is taken over all possible input distribution $p(x)$.

- Operational definition of channel capacity: The highest rate in bits per channel use at which information can be sent.

- Shannon’s second theorem: The information channel capacity is equal to the operational channel capacity.
Example 1

Noiseless binary channel

\[ p(Y = 0) = p(X = 0) = \pi_0, \quad p(Y = 1) = p(X = 1) = \pi_1 = 1 - \pi_0 \]
\[ I(X; Y) = H(Y) - H(Y|X) = H(Y) \leq 1 \]
\[ \Rightarrow \pi_0 = \pi_1 = \frac{1}{2} \]
Example 2

Noisy channel with non-overlapping outputs

\[ I(X; Y) = H(Y) - H(Y|X) = H(Y) - \pi_0 H(p) - \pi_1 H(q) \]
\[ = H(\pi_0) = H(X) \leq 1 \]
Noisy Typewriter

FIGURE 7.4. Noisy Typewriter. $C = \log 13$ bits.
Noisy Typewriter

\[ I(X; Y) = H(Y) - H(Y|X) \]
\[ = H(Y) - \sum_x p(x) H(Y|X = x) \]
\[ = H(Y) - \sum_x p(x) H(\frac{1}{2}) \]
\[ = H(Y) - H(\frac{1}{2}) \]
\[ \leq \log 26 - 1 = \log 13 \]
\[ C = \max I(X; Y) = \log 13 \]
Binary Symmetric Channel (BSC)

**FIGURE 7.5.** Binary symmetric channel. $C = 1 - H(p)$ bits.
Binary Symmetric Channel (BSC)

\[
I(X;Y) = H(Y) - H(Y|X) \\
= H(Y) - \sum_x p(x)H(Y|X=x) \\
= H(Y) - \sum_x p(x)H(p) \\
= H(Y) - H(p) \\
\leq 1 - H(p) \\
C = \max I(X;Y) = 1 - H(p)
\]
Binary Erasure Channel

FIGURE 7.6. Binary erasure channel.
Binary Erasure Channel

\[ I(X; Y) \]
\[ = H(Y) - H(Y|X) \]
\[ = H(Y) - \sum_x p(x) H(Y|X = x) \]
\[ = H(Y) - \sum_x p(x) H(\alpha) \]
\[ = H(Y) - H(\alpha) \]

\[ H(Y) = (1 - \alpha)H(\pi_0) + H(\alpha) \]

\[ C = \max I(X; Y) = 1 - \alpha \]
7.3 Properties of Channel Capacity
Properties of Channel Capacity

- \( C \geq 0 \).
- \( C \leq \log |\mathcal{X}|. \)
- \( C \leq \log |\mathcal{Y}|. \)
- \( I(X; Y) \) is a continuous function of \( p(x) \),
- \( I(X; Y) \) is a concave function of \( p(x) \),
7.4 Preview of the Channel Coding Theorem
For each input \( n \)-sequence, there are approximately \( 2^{nH(Y|X)} \), possible \( Y \) sequences.

The total number of possible (typical) \( Y \) sequences is \( 2^{nH(Y)} \).

This set has to be divided into sets of size \( 2^{nH(Y|X)} \) corresponding to the different input \( X \) sequences.

The total number of disjoint sets is less than or equal to \( 2^{nH(Y)} / 2^{nH(Y|X)} = 2^n(H(Y) - H(Y|X)) = 2^nI(X;Y) \)

We can send at most \( 2^{nI(X;Y)} \) distinguishable sequences of length \( n \).
Example

- 6 typical sequences for $X^n$. 4 typical sequences for $Y^n$.
- 12 typical sequences for $(X^n, Y^n)$.
- For every $X^n$, we have
  \[ 2^{nH(X,Y)} / 2^{nH(X)} = 2^{nH(Y|X)} = 2 \] typical $Y^n$.

  e.g., for $X^n = 001100 \Rightarrow Y^n = 010100, 101011$. 
Example

Since we have $2^n H(Y) = 4$ typical $Y^n$ in total, how many typical $X^n$ can these typical $Y^n$ be assigned?

$$
2^n H(Y) / 2^n H(Y | X) = 2^n (H(Y) - H(Y | X)) = 2^n I(X;Y) = 2.
$$

Can we assign more typical $X^n$? No. For some $Y^n$ received, we can’t determine which $X^n$ is received. e.g., If we use 001100, 101101, and 101000 as codewords, we can’t determine which codeword is sent when we receive 101011.
7.5 Definitions
Communication Channel

- **Message** $W \in \{1, 2, \ldots, M\}$.
- **Encoder**: input $W$, output $X^n \equiv X^n(W) \in \mathcal{X}^n$
  - $n$ is the length of the signal. We then transmit the signal via the channel by using the channel $n$ times. Every time we send a symbol of the signal.
- **Channel**: input $X^n$, output $Y^n$ with distribution $p(y^n|x^n)$
- **Decoder**: input $Y^n$, output $\hat{W} = g(Y^n)$ where $g(Y^n)$ is a deterministic decoding rule.
- If $\hat{W} \neq W$, an error occurs.

**FIGURE 7.8.** Communication channel.
Definitions

Definition 3 (Discrete Channel) A discrete channel, denoted by 
\((\mathcal{X}, p(y|x), \mathcal{Y})\), consists of two finite sets \(\mathcal{X}\) and \(\mathcal{Y}\) and a collection of 
probability mass functions \(p(y|x)\).

- \(X\): input, \(Y\): output, for every input \(x \in \mathcal{X}\), \(\sum_y p(y|x) = 1\).

Definition 4 (Discrete Memoryless Channel, DMC) The \(n\)th 
extension of the discrete memoryless channel is the channel 
\((\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)\) where 
\(p(y_k|x^k, y^{k-1}) = p(y_k|x_k)\), \(k = 1, 2, \ldots, n\).

- Without feedback: \(p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})\)

- \(n\)th extension of DMC without feedback:

\[
p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i).
\]
Definitions

Definition 5 \((M, n)\) code] An \((M, n)\) code for the channel \((\mathcal{X}, p(y|x), \mathcal{Y})\) consists of the following:

1. An index set \(\{1, 2, \ldots, M\}\).

2. An encoding function \(X^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n\). The codewords are \(x^n(1), x^n(2), \ldots, x^n(M)\). The set of codewords is called the codebook.

3. A decoding function \(g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}\)
Definitions

Definition 6 (Conditional probability of error)

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

\[ = \sum_{y^n} p(y^n|x^n(i))I(g(y^n) \neq i) \]

\(I(\cdot)\) is the indicator function.
Definitions

Definition 7 (Maximal probability of error)

\[ \lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i \]

Definition 8 (Average probability of error)

\[ P_{e}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \]

The decoding error is

\[ \Pr(g(Y^n) \neq W) = \sum_{i=1}^{M} \Pr(W = i) \Pr(g(Y^n) \neq i | W = i) \]

If the index \( W \) is chosen uniformly from \( \{1, 2, \ldots, M\} \), then

\[ P_e^{(n)} = \Pr(g(Y^n) \neq W). \]
Definitions

**Definition 9 (Rate)** The rate $R$ of an $(M, n)$ code is

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

**Definition 10 (Achievable rate)** A rate $R$ is said to be achievable if there exists a $(\lceil 2^{nR} \rceil, n)$ code such that the maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$.

**Definition 11 (Channel capacity)** The capacity of a channel is the supremum of all achievable rates.
7.6 Jointly Typical Sequences
Definitions

Definition 12 (Jointly typical sequences) The set $A^{(n)}_\epsilon$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is defined by

$$A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \right.$$ \left.$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \right.$$ \left.$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \right\}$$

where

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

1. \(\Pr\left( (x^n, y^n) \in A^{(n)}_{\epsilon} \right) \rightarrow 1 \text{ as } n \rightarrow \infty.\)

2. \(\left| A^{(n)}_{\epsilon} \right| \leq 2^{n(H(X,Y) + \epsilon)}\)

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], then

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

Also, for sufficient large \(n\),

\[
\Pr\left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_{\epsilon} \right) \geq (1 - \epsilon)2^{-n(I(X;Y) + 3\epsilon)}.
\]
Joint AEP

Theorem 2 (Joint AEP) 1. \( \Pr \left( (x^n, y^n) \in A^{(n)}_\epsilon \right) \to 1 \text{ as } n \to \infty. \)

Proof. Given \( \epsilon > 0 \), define events \( A, B, C \) as

\[
A \triangleq \left\{ X^n : \left| -\frac{1}{n} \log p(X^n) - H(X) \right| \geq \epsilon \right\}
\]

\[
B \triangleq \left\{ Y^n : \left| -\frac{1}{n} \log p(Y^n) - H(Y) \right| \geq \epsilon \right\}
\]

\[
C \triangleq \left\{ (X^n, Y^n) : \left| -\frac{1}{n} \log p(X^n, Y^n) - H(X, Y) \right| \geq \epsilon \right\},
\]
Then, by weak law of large number, there exists \( n_1, n_2, n_3 \) such that,

\[
\Pr(A) < \frac{\epsilon}{3}, \quad \forall n > n_1, \quad \Pr(B) < \frac{\epsilon}{3}, \quad \forall n > n_2, \\
\Pr(C) < \frac{\epsilon}{3}, \quad \forall n > n_3.
\]

Thus,

\[
\Pr \left( (x^n, y^n) \in A_{\epsilon}^{(n)} \right) = \Pr(A^c \cap B^c \cap C^c) \\
= 1 - \Pr(A \cup B \cup C) \geq 1 - (\Pr(A) + \Pr(B) + \Pr(C)) \\
\geq 1 - \epsilon
\]

for all \( n > \max\{n_1, n_2, n_3\} \). \( \square \)
Joint AEP

Theorem 3 (Joint AEP) 2. \(|A^{(n)}_{\epsilon}| \leq 2^{n(H(X,Y)+\epsilon)}

Proof.

\[ 1 = \sum p(x^n, y^n) \geq \sum_{(x^n,y^n) \in A^{(n)}_{\epsilon}} p(x^n, y^n) \geq |A^{(n)}_{\epsilon}| 2^{-n(H(X,Y)+\epsilon)} \]

Thus,

\[ |A^{(n)}_{\epsilon}| \leq 2^{n(H(X,Y)+\epsilon)}. \]
Theorem 4 (Joint AEP) 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], then
\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_\epsilon \right) \leq 2^{-n(I(X;Y)-3\epsilon)}.
\]
Also, for sufficient large \(n\),
\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_\epsilon \right) \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)}.
\]
Joint AEP

Proof.

\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_\epsilon \right) = \sum_{(x^n, y^n) \in A^{(n)}_\epsilon} p(x^n)p(y^n) \\
\leq 2^n (H(X,Y) + \epsilon) 2^{-n(H(X) - \epsilon)} 2^{-n(H(Y) - \epsilon)} \\
= 2^{-n(I(X;Y) - 3\epsilon)}.
\]

For sufficient large \( n \), \( \Pr \left( A^{(n)}_\epsilon \right) \geq 1 - \epsilon \), and therefore

\[
1 - \epsilon \leq \sum_{(x^n, y^n) \in A^{(n)}_\epsilon} p(x^n, y^n) \leq |A^{(n)}_\epsilon| 2^{-n(H(X,Y) - \epsilon)}.
\]

and

\[
|A^{(n)}_\epsilon| \geq (1 - \epsilon) 2^{n(H(X,Y) - \epsilon)}
\]
Joint AEP

$$\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_{\epsilon} \right)$$

$$= \sum_{(x^n, y^n) \in A^{(n)}_{\epsilon}} p(x^n)p(y^n)$$

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

$$= (1 - \epsilon) 2^{-n(I(X;Y)+3\epsilon)}$$
Joint AEP: Conclusion

- There are about $2^n(H(X))$ typical $X$ sequences, and about $2^n(H(Y))$ typical $X$ sequences.
- There are about $2^n(H(X,Y))$ jointly typical sequences.
- Randomly chosen a pair of typical $X^n$ and typical $Y^n$, the probability that it is jointly typical is about $2^{-nI(X;Y)}$. 
7.7 Channel Coding Theorem
Channel Coding Theorem

**Theorem 5 (Channel coding theorem)** 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 $R \leq C$.

- We have to prove two parts.
  - $R < C \rightarrow$ achievable.
  - Achievable $\rightarrow R \leq C$.

- Main ideas.
  - Random encoding (random code)
  - Jointly typical decoding
Random Code

- Generate a \((2^{nR}, n)\) code at random according to the distribution \(p(x)\) (fixed). That is, the \(2^{nR}\) codewords have the distribution

\[
p(x^n) = \prod_{i=1}^{n} p(x_i)
\]

- A particular code \(C\) is the matrix with \(2^{nR}\) codewords as the row.

\[
C = \begin{bmatrix}
x_1(1) & x_2(1) & \cdots & x_n(1) \\
x_1(2) & x_2(2) & \cdots & x_n(2) \\
\vdots & \vdots & \ddots & \vdots \\
x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR})
\end{bmatrix}
\]

- The code \(C\) is revealed to both sender and receiver. Both sender and receiver are also assumed to know the channel transition matrix \(p(y|x)\) for the channel.
Random Code

- There are $\left(|\mathcal{X}|^n\right)^{2^nR}$ different codes.
- The probability of a particular code $\mathcal{C}$ is

$$Pr(\mathcal{C}) = \prod_{w=1}^{2^nR} \prod_{i=1}^n p(x_i(w))$$
A message $W$ is chosen according to a uniform distribution

$$\Pr[W = w] = 2^{-nR}, \quad w = 1, 2, \ldots, 2^{nR}.$$ 

The $w$th codeword $X^n(w)$, corresponding to the $w$th row of $C$, is sent over the channel.

The receiver receives a sequence $Y^n$ according to the distribution

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

That is, use the DMC channel for $n$ times.
The receiver declares that the message $\hat{W}$ was sent if
- $\left( X^n(\hat{W}), Y^n \right)$ is jointly typical.
- There is no other jointly typical pair for $Y^n$. That is, there is no other $W' \neq \hat{W}$ such that $W', Y^n$ is jointly typical.

If no such $\hat{W}$ exists or if there is more than one such, an error is declared ($\hat{W} = 0$).

There is decoding error if $\hat{W} \neq W$. Let $E$ be the event $[\hat{W} \neq W]$. 
Proof of $R < C \rightarrow$ Achievable

- The average probability of error averaged over all codewords in the codebook, and averaged over all codebooks.

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

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

- $P_e^{(n)}(C)$ is defined for jointly typical decoding.
- By the symmetry of the code construction, $\sum_C \Pr(C) \lambda_w(C)$ does not depend on $w$. 
Proof of $R < C \rightarrow \text{Achievable}$

Therefore,

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

$$= \sum_{C} \Pr(C) \lambda_w(C) \quad \text{for any } w$$

$$= \sum_{C} \Pr(C) \lambda_1(C) = \Pr(\mathcal{E}|W = 1)$$

Define $E_i = \{(X^n(i), Y^n) \text{ is jointly typical pair}\}$ for $i = 1, 2, \ldots, 2^{nR}$ where $Y^n$ is the channel output when the first codeword $X^n(1)$ was sent. Then decoding error is declared if

- $E_1^c$: The transmitted codeword and the received sequence are not jointly typical.

- $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$: a wrong codeword is jointly typical with the received sequence.
Proof of $R < C \rightarrow$ Achievable

- $Y^n$ is the channel output when the first codeword $X^n(1)$ was sent.
- $E_1^c$: The transmitted codeword and the received sequence are not jointly typical.
- $E_2, E_3, \ldots, E_{2^n R}$: wrong codewords that are jointly typical with the received sequence.
Proof of \( R < C \) → Achievable

The average error

\[
\Pr(\mathcal{E}|W = 1) = P(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^nR}|W = 1) \\
\leq P(E_1^c|W = 1) + \sum_{i=2}^{2^nR} P(E_i|W = 1)
\]

By AEP,

\[
P(E_1^c|W = 1) \leq \epsilon \quad \text{for } n \text{ sufficiently large}
\]

\[
P(E_i|W = 1) \leq 2^{-n(I(X;Y) - 3\epsilon)}
\]

\((Y^n \text{ and } X^n(1) \text{ are jointly typical.})\)
Proof of $R < C \rightarrow \text{Achievable}$

- We have

\[
\Pr(\mathcal{E}|W = 1) \leq \epsilon + (2^{nR} - 1)2^{-n(I(X;Y) - 3\epsilon)}
\leq \epsilon + 2^{nR}2^{-n(I(X;Y) - 3\epsilon)}
= \epsilon + 2^{-n(I(X;Y) - R - 3\epsilon)}
\]

- If $I(X;Y) - R - 3\epsilon > 0$, then $2^{-n(I(X;Y) - R - 3\epsilon)} < \epsilon$ for $n$ sufficiently large, and

\[
\Pr(\mathcal{E}|W = 1) \leq 2\epsilon.
\]

- So far, we prove that: for any $\epsilon$, if $R < I(X;Y)$ and $n$ sufficiently large, the average decoding error $\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) < 2\epsilon$.

- What do we need? If $R < C$, the maximum error probability $\lambda^{(n)} \rightarrow 0$.

(We are almost there. Almost...)

Peng-Hua Wang, April 16, 2012
Proof of $R < C \rightarrow$ Achievable, final part

- Choose $p(x)$ such that $I(X; Y)$ is maximum. That is, choose $p(x)$ such that $I(X; Y)$ achieve channel capacity $C$. Then the condition $R < I(X; Y) - 3\epsilon$ can be replaced by the achievability condition $R < C - 3\epsilon$.

- Since the average probability of error over codebooks is less than $2\epsilon$, there exists at least one codebook $C^*$ such that $Pr(E|C^*) < 2\epsilon$.
  - $C^*$ can be found by an exhaustive search over all codes.

- Since $W$ is chosen uniformly, we have

$$Pr(E|C^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*) \leq 2\epsilon$$

which implies that the maximal error probability of the better half codewords is less than $4\epsilon$.

- There are 10 students. Their average score is 40. Then the highest
Proof of $R < C \rightarrow$ Achievable, final part

- We throw away the worst half of the codewords in the best codebook $C^*$. The new code has a maximal probability of error less than $4\epsilon$. However, we construct a $(2^{nR/2}, n)$ or $(2^{n(R-1/n)}, n)$ code. The rate of the new code is $R - 1/n$.

- Summary. If $R - 1/n < C - 3\epsilon$ for any $\epsilon$, then $\lambda^{(n)} \leq 4\epsilon$ for $n$ sufficiently large.
7.8 Zero-Error Codes
No error $\rightarrow R \leq C$

- Assume that we have a $(2^{nR}, n)$ code with zero probability of error.
  - $W$ is determined by $Y^n$. $p(g(Y^n) = W) = 1$. $H(W|Y^n) = 0$.

- To obtain a strong bound, assume that $W$ is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)$$
$$\leq I(X^n; Y^n) \text{ (data processing ineq. } W \rightarrow X^n(W) \rightarrow Y^n)$$
$$\leq \sum_{i=1}^{n} I(X_i; Y_i) \quad \text{(See next page.)}$$
$$\leq nC \quad \text{(definition of channel capacity)}$$

- That is, no error $\rightarrow R \leq C$. 
Lemma 1 Let $Y^n$ be the result of passing $X^n$ through a discrete memoryless channel of capacity $C$. Then for all $p(x^n)$, $I(X^n; Y^n) \leq nC$.

Proof.

\[ I(X^n; Y^n) = H(Y^n) - H(Y^n | X^n) \]

\[ = H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y_1, \ldots, Y_{i-1}, X^n) \]

\[ = H(Y^n) - \sum_{i=1}^{n} H(Y_i | X_i) \quad \text{(definition of DMC)} \]

\[ \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) = \sum_{i=1}^{n} I(Y_i; X_i) \leq nC \]
7.9 Fano’s Inequality and the Converse to the Coding Theorem
Fano’s Inequality

Theorem 6 (Fano’s inequality) Let \( X \) and \( W \) have the same sample spaces \( \mathcal{X} = \{1, 2, \ldots, M\} \) and have the joint p.m.f. \( p(x, w) \). Let

\[
P_e = \Pr[X \neq W] = \sum_{x \in \mathcal{X}} \sum_{w \in \mathcal{X}, w \neq x} p(x, w).
\]

Then

\[
P_e \log(M - 1) + H(P_e) \geq H(X|W)
\]

where

\[
H(P_e) = -P_e \log P_e - (1 - P_e) \log(1 - P_e).
\]
Fano’s Inequality

Proof. We will prove that $H(X|W) - H(P_e) - P_e \log(M - 1) \leq 0$.

\[
H(X|W) = \sum_x \sum_w p(x, w) \log \frac{1}{p(x|w)}
\]

\[
= \sum_x \sum_{w \neq x} p(x, w) \log \frac{1}{p(x|w)}
\]

\[
+ \sum_x \sum_{w = x} p(x, w) \log \frac{1}{p(x|w)}
\]

\[
- P_e \log(M - 1) = \sum_x \sum_{w \neq x} p(x, w) \log \frac{1}{M - 1}
\]

\[
- H(P_e) = P_e \log P_e + (1 - P_e) \log(1 - P_e)
\]

\[
= \sum_x \sum_{w \neq x} p(x, w) \log P_e
\]

\[
+ \sum_x \sum_{w = x} p(x, w) \log(1 - P_e)
\]

Add the above three terms together.
Fano’s Inequality

Proof (cont.)

\[ H(X|W) - P_e \log(M - 1) - H(P_e) \]

\[ = \sum_x \sum_{w \neq x} p(x, w) \log \frac{P_e}{(M - 1)p(x|w)} + \sum_x \sum_{w = x} p(x, w) \log \frac{1 - P_e}{p(x|w)} \]

\[ \leq \log \left( \sum_x \sum_{w \neq x} p(x, w) \frac{P_e}{(M - 1)p(x|w)} + \sum_x \sum_{w = x} p(x, w) \frac{1 - P_e}{p(x|w)} \right) \]

\[ = \log \left( \frac{P_e}{M - 1} \sum_x \sum_{w \neq x} p(w) + (1 - P_e) \sum_x \sum_{w = x} p(w) \right) \]

\[ = \log[P_e + (1 - P_e)] = 0 \quad \Box \]
Fano’s Inequality

**Corollary 1**

1. \( P_e \log M + H(P_e) \geq H(X|W), \) \( P_e = \Pr[X \neq W] \)

2. \( 1 + P_e \log M \geq H(X|W), \) \( P_e = \Pr[X \neq W] \)

3. If \( X \rightarrow Y \rightarrow \hat{X} \) and \( P_e = \Pr[X \neq \hat{X}] \), then

   \[
   H(P_e) + P_e \log M \geq H(X|\hat{X}) \geq H(X|Y)
   \]

**Remark.**

1. \( H(X|W) \leq P_e \log(M - 1) + H(P_e) \leq P_e \log M + H(P_e). \)
2. \( H(X|W) \leq P_e \log(M - 1) + H(P_e) \leq P_e \log M + 1. \)
3. The second ineq. can be obtained by data processing ineq.
Data Processing Inequality

Lemma 2 (Data processing inequality) If $X \rightarrow Y \rightarrow Z$, then

$I(X; Z) \leq I(X; Y)$

Proof.

\[
\begin{align*}
I(X; Z) - I(X; Y) &= H(X) - H(X|Z) - [H(X) - H(X|Y)] = H(X|Y) - H(X|Z) \\
&= \sum_x \sum_y p(x, y) \log \frac{1}{p(x|y)} - \sum_x \sum_z p(x, z) \log \frac{1}{p(x|z)} \\
&= \sum_x \sum_y \sum_z p(x, y, z) \log \frac{1}{p(x|y)} - \sum_x \sum_y \sum_z p(x, y, z) \log \frac{1}{p(x|z)} \\
&\leq \log \left( \sum_x \sum_y \sum_z p(x, y, z) \frac{p(x|z)}{p(x|y)} \right) \quad \text{(by convexity of logarithm)}
\end{align*}
\]
Data Processing Inequality

Proof (cont.) Since $X \rightarrow Y \rightarrow Z$, we have

$$p(x, y, z) = p(x, y)p(z|x, y) = p(x, y)p(z|y) = \frac{p(x, y)p(y, z)}{p(y)}$$

and

$$p(x, y, z) \frac{p(x|z)}{p(x|y)} = \frac{p(x, y)p(y, z)}{p(y)} \times \frac{p(x, z)p(y)}{p(z)p(x, y)} = \frac{p(x, z)p(y, z)}{p(z)}$$

Therefore,

$$\sum_x \sum_y \sum_z p(x, y, z) \frac{p(x|z)}{p(x|y)} = \sum_x \sum_y \sum_z \frac{p(x, z)p(y, z)}{p(z)}$$

$$= \sum_x \sum_z \frac{p(x, z)}{p(z)} \sum_y p(y, z) = \sum_x \sum_z p(x, z) = 1 \quad \square$$
Data Processing Inequality (Summary)

Lemma 3 1. If $X \rightarrow Y \rightarrow Z$, then

\[
I(X; Z) \leq \begin{cases} 
I(X; Y) \\
I(Y; Z)
\end{cases},
\]

\[H(X|Y) \leq H(X|Z)\]

2. If $X \rightarrow Y \rightarrow Z \rightarrow W$, then

\[
I(X; Z) + I(Y; W) \leq I(X; W) + I(Y; Z),
\]

\[I(X; W) \leq I(Y; Z)\]
Achievable $\rightarrow R \leq C$

Theorem 7 (Converse to Channel coding theorem) Any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof.

- For a fixed encoding rule $X^n(W)$ and a fixed decoding rule $\hat{W} = g(Y^n)$, we have $W \rightarrow X^n(W) \rightarrow Y^n \rightarrow \hat{W}$.
- For each $n$, let $W$ be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$.
- Since $W$ has a uniform distribution,

$$
\Pr[W \neq \hat{W}] = P_e^{(n)} = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i.
$$
Achievable $\rightarrow R \leq C$

Proof (cont.)

$$nR = H(W) \quad (W \text{ is uniform distribution})$$

$$= H(W|\hat{W}) + I(W;\hat{W})$$

$$\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \quad \text{(Fano’s ineq.)}$$

$$\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) \quad \text{(data processing ineq.)}$$

$$\leq 1 + P_e^{(n)} nR + nC \quad \text{(lemma 7.9.2)}$$

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

That is, if $R > C$, the probability of error is large than a positive value for sufficiently large $n$. The error probability can’t achieve arbitrary small. □