Channel Capacity

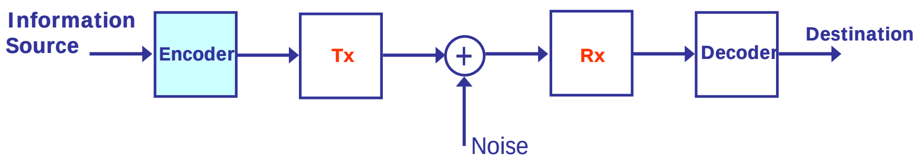
2009 2-2 Course

- Information Theory -

Tetsuo Asano and
Tad matsumoto
Email: {t-asano, matumoto}@jaist.ac.jp

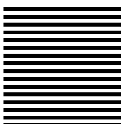
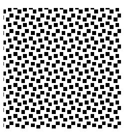
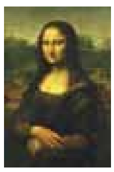
Japan Advanced Institute of Science and Technology
Asahidai 1-1, Nomi, Ishikawa 923-1292, Japan
http://www.jaist.ac.jp

In the last Chapters, we learned .....



How to effectively encode the source with the knowledge of source appearance probability.

What the theoretical asymptotic properties of the source encoding techniques are with the knowledge of the probability.

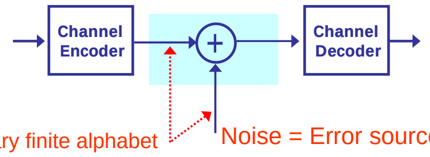
In this Chapters, we learn .....

Communication systems "use" channel as medium for information transfer.

- Then, question arises that what is the maximum capability of the channel.
- How can the communication systems use the channel's capability.
- What are theoretical asymptotic properties of the channel coding techniques for error protection, given the knowledge of the channel characteristics.

We skip the transmitter and receiver!!!!

Assumptions:



Binary/Non-binary finite alphabet

Noise = Error source

The transmitter and receivers are ignored (Channel output is directly connected to the channel decoder).

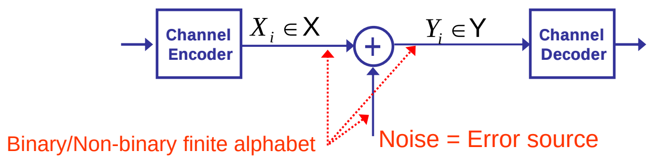
The both channel encoder output and noise take form of binary or non-binary *finite* alphabet.

Note: Those assumptions are to be eliminated in the next Chapter.

Outline

1. Channel Capacity
   - Definition
   - Some Examples
   - Some Properties
2. Random Coding
3. Channel Coding Theorem
   - Proof for Sufficiency
4. Fano's Inequality for Extension
   - Proof for Necessity

Channel Model: Review



Channel input sequence: $X_0, X_1, \ldots, X_i, \ldots$, $\quad X_i \in \mathcal{X}$

Finite Alphabet: $\mathcal{X}=\{x_1, x_2, x_3, \ldots, x_q\}$, where q is the alphabet size.

Channel output sequence: $Y_0, Y_1, \ldots, Y_i, \ldots$, $\quad Y_i \in \mathcal{Y}$

Finite Alphabet: $\mathcal{Y}=\{y_1, y_2, y_3, \ldots, y_r\}$, where r is the alphabet size.

Let the conditional joint probability of $Y_0=y_0, Y_1=y_1, \ldots, Y_{n-1}=y_{n-1}$, conditioned upon $X_0=x_0, X_1=x_1, \ldots, X_{n-1}=x_{n-1}$ be denoted as

$$P_{Y_0,Y_1,\ldots,Y_{n-1}|X_0,X_1,\ldots,X_{n-1}}\left(y_0, y_1, \ldots, y_{n-1} \middle| x_0, x_1, \ldots, x_{n-1}\right)$$

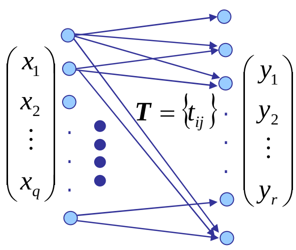
Channel Capacity (1)

Definition 8.1.0: Channel Matrix

The channel matrix $T=\{t_{ij}\}$ is given as a matrix of which entry is defined as the probability that the transmitted symbol x_i is received as y_j, as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix} = \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,q} \\ t_{2,1} & t_{1,1} & & t_{2,q} \\ \vdots & & \ddots & \vdots \\ t_{r,1} & t_{r,2} & \cdots & t_{r,q} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix}$$

or $\quad \boldsymbol{y} = \boldsymbol{T}\boldsymbol{x}$.



Definition 8.1.1: Channel Capacity

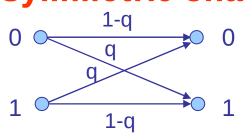
The information channel capacity C of memory-less channel is defined as:

$$C = \underbrace{\max}_{p(x)} I(X;Y)$$

where the maximization is taken over all possible input distributions $p(x)$.

Channel Capacity (2)

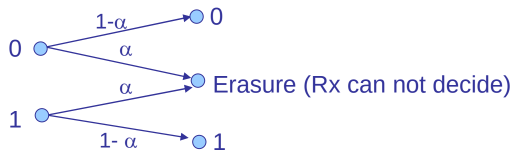
Example: Binary Symmetric Channel



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

where the equality holds if the input distribution is uniform. Therefore, we can conclude that $C = 1 - H(q)$ bits.

Example: Binary Erasure Channel



Channel Capacity (3)

Define the erasure event by E.
Since the probabilities of non-error and erasure events are distinctive,

$$H(Y) = H(Y,E) = H(E) + H(Y|E)$$

Denote that $Pr(X=1)=\pi$,

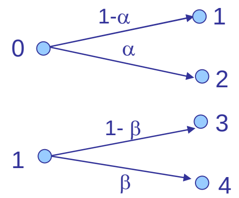
$$H(Y) = H((1-\pi)(1-\alpha), (1-\pi)\alpha + \pi\alpha, \pi(1-\alpha)) = H((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha))$$
$$= H(\alpha) + (1-\alpha)H(\pi)$$

Hence, $C = \underbrace{\max}_{p(x)} H(Y) - H(\alpha) = \underbrace{\max}_{\pi} (1-\alpha)H(\pi) + H(\alpha) - H(\alpha) = 1-\alpha$

with π=1/2.

Channel Capacity (4)

Example: Noisy Channel with no Overlapping Outputs



$$I(X;Y) = H(X) - H(X \mid Y) = H(X) - \underbrace{\sum_{y=1,2,3,4} p(y) H(X \mid Y = y)}_{=0}$$

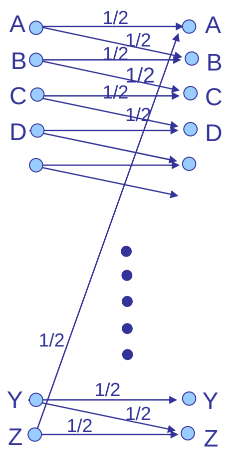
This is because the receiver can identify which of X=0 or X=1 was transmitted only by looking at the received symbol. Since the transmitted symbols are binary.

$$C = \max I(X;Y) = \max H(X) = 1 \quad bit$$

with $Pr(X=1)=Pr(X=0)=1/2$.

Channel Capacity (5)

Exercise: Noisy Type Writer



Calculate the capacity of this noisy type writer, assuming that all characters A-Z appears with equal probability.

Channel Capacity (6)

Definition 8.1.2: Symmetric Channel

The channel is called symmetric if the rows of the channel transition matrix are permutation of each other, and so are the columns.

Example:

A channel having the following transition matrix is symmetric.

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

This channel has the capacity:

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

where $\boldsymbol{r}$ is a row of the channel transition matrix.

The equality holds if the output distribution is uniform. However, if $p(x)$ is uniform,

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

and hence, $p(y)$ is also uniform.

Channel Capacity (7)

Property 8.1.1:

(1) The channel capacity is non negative, i.e., $C\ge 0$ because

$$C=\max I(X;Y)\ge I(X;Y)\ge 0$$

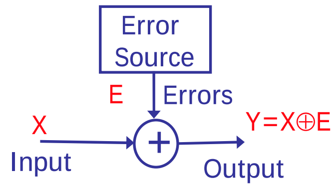
(2) $C\le \log|\mathcal{X}|$ because $C=\max I(X;Y)\le \max H(X)=\log|\mathcal{X}|$

(3) $C\le \log|\mathcal{Y}|$ because $C=\max I(X;Y)\le \max H(Y)=\log|\mathcal{Y}|$

(4) $I(X;Y)$ is a continuous and concave function of $p(x)$. (See Theorem 4.3.4)

Channel Capacity (8)

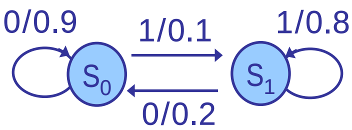
Burst Error Channel:



Since there is no chance that we change the appearance probability of the alphabet from thesource X, capacity is: C=1-H(E), where H(E)=H{p:Prob(E=1)=p}

with H(x)=-xlog_2x-(1-x)log_2(1-x).

Memory Channel



Stationary State Probabilities: p_0 and p_1

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \text{ with } p_0 + p_1 = 1$$

⇘ p_0=2/3 and p_1=1/3

Entropy of the Error Source: H(E)=(2/3)×H(0.1) +(1/3)×H(0.8)
= (2/3) ×0.4690+(1/3)×0.7219
=0.5532

Capacity: C=1-H(E)=1-0.5532=0.4467 (bits/symbol)

Channel Capacity (9)

Bit Error Rate: P_b=0.1×p_0+0.8×p_1=1/3

Memory-less Channel (P$_b$=1/3)

Capacity: C=1-H(E)=1-*H*(1/3)=0.0817 (bits/symbol)

Capacity with Memory Channel >>> Capacity with Memory-less Channel

Fading channel without Interleaving: Memory Channel

Interleaved channels: Memory-less Channel

Channel Coding Theorem: Preparation (1)

Assume that a message W to be transmitted over the channel is drawn from the index set {1,2, ..., M}, i.e., each message is indexed by a number from this set. The following summarizes how the communication system we analyze works:

(1) The message W is then encoded into a length n block of symbols, $(x_1, x_2, \dots, x_n)$. The encoded message is denoted as $X^n(W)$.
(2) The transmitted message $X^n(W)=(x_1, x_2, \dots, x_n)$ suffers from noise in the channel, and received as a random sequence $Y^n=(y_1, y_2, \dots, y_n)$ by the receiver.
(3) The channel has its transition matrix as described by a conditional probability $p(y^n|x^n)=p(y_1, y_2, \dots, y_n|x_1, x_2, \dots, x_n)$.
(4) The receiver aims to retrieve the transmitted message by an estimation rule $g(.)$, where $\hat{W}=g(Y^n)$.
(5) The receiver makes an error, if $\hat{W} \neq W$.

Definition 8.2.1: Communication System

The communication system that follows the rule described above is denoted as:

$$\left(X^n, p(y^n|x^n), Y^n\right)$$

Obviously, according to this definition, the channel is the n-th extension of the memory-less channel $\left(X, p(y|x), Y\right)$.

Channel Coding Theorem: Preparation (2)

Definition 8.2.2: Memory-less Channel

The channel according to Definition 8.2.1 is memory-less, if

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \quad k=1,2,\cdots,n$$

The transition function for the memory-less channel reduces to:

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

Definition 8.2.3: Channel Code

An (M, n) code is defined for the channel $\left(X, p(y|x), Y\right)$, where M is the number of the messages to be transmitted, and n is the length of the encoded sequence. The Roles of the encoder and decoder are defined as follows:

(1) Encoder maps the M messages to their corresponding sequences, $X^n(1)$, $X^n(2)$, ..., $X^n(M)$. The set of the code words is called code book.
(2) The decoder function $g(.)$ maps the received sequence of random variable on to the most likely message estimate $g(Y^n) \in \{1,2,\cdots,M\}$.

Channel Coding Theorem: Preparation (3)

Definition 8.2.3: Probability of Error

The conditional probability λ_i, given the message indexed by i has been transmitted, is given by:

$$\lambda_i = \Pr\{g(Y^n) \neq i | X^n = X^n(i)\} = \sum_{y^n \in Y^n} p(y^n | x^n(i)) \; I(g(y^n) \neq i)$$

where $I(.)$ is the indicator function, which takes value 1 if the argument is satisfied, and otherwise it takes value 0.

Definition 8.2.4: Maximal Probability of Error

The maximal error probability $\lambda^{(n)}$ for an (M, n) code is defined as:

$$\lambda^{(n)} = \underbrace{\max}_{i \in \{1,2,\cdots,M\}} \lambda_i$$

Definition 8.2.5: Average Probability of Error

The average error probability $P_e^{(n)}$ for an (M, n) code is defined as:

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

Channel Coding Theorem: Preparation (4)

Property 8.2.1: Average Probability of Error

The following properties hold:

(1) $P_e^{(n)} = \Pr(i \neq g(Y^n))$ if the message transmitted is chosen uniformly from the set.

(2) $P_e^{(n)} \leq \lambda^{(n)}$

Definition 8.2.6: Code Rate

The rate R of an (M, n) code is defined as: $R = \dfrac{\log M}{n}$

Random Coding (1)

Consider (2^{nR}, n) code. Generate a code *at random*, according to the probability distribution $p(x)$ of the symbols x. Since there are 2^{nR} code words randomly selected from the codebook C, which is denoted as:

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

where the argument of (.) indicates the code word index, and the subscript indicates the symbol index.

Since the appearance of the symbols is independent,

$$p(x^n) = \prod_{i=1}^{n} p\{x_i(w)\} \qquad for \quad w = 1,2,\ldots,2^{nR}$$

Since each code word is generated randomly, this codebook C itself is random variable. The probability of generating a particular code word is given by:

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

Random Coding (2)

The code C is known to the both transmitter and receiver. They are assumed to also know the channel transition matrix $p(y|x)$. A message W is chosen equi-probably. Therefore,

$$\Pr(W = w) = 2^{-nR}, \qquad w = 1,2,\ldots,2^{nR}$$

Assume that the w-th code word $X^n(w)$, corresponding to the w-th row of C, is transmitted.

A sequence received by the receiver, denoted as Y^n, follows the distribution:

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

The receiver guesses which of the 2^{nR} sequences is most likely to have been sent by comparing the conditional probabilities of the possible code words, and selects the one satisfying:

$$\hat{W} = \arg\max_{w} P(y^n | x^n(w))$$

If $\hat{W} \neq W$, there is a decoding error.

The codebook C is updated at every transmission timing. Therefore, this technique is called random coding.

Channel Coding Theorem (1)

Theorem 8.3.1: Channel Coding Theorem

(1) There exists a $(2^{nR}, n)$ rate R code such that the maximum error probability $\lambda^{(n)}$ can be made arbitrarily small, if the code rate is lower than the capacity $R<C$.

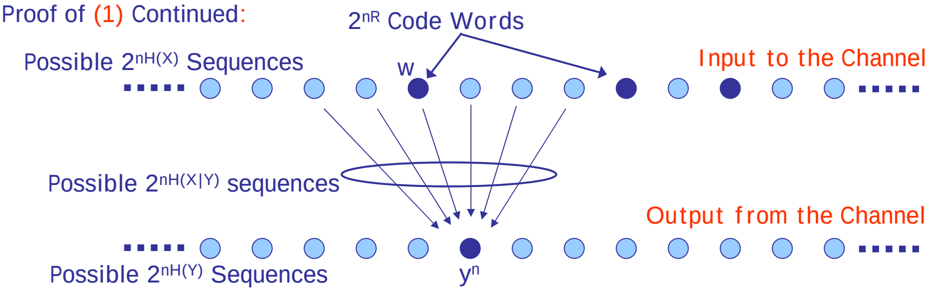
(2) Conversely, any $(2^{nR}, n)$ rate R code that can achieve arbitrarily small $\lambda^{(n)}$ must satisfy $R<C$.

Proof of (1):
The proof uses the properties of random coding. The detailed proof is far beyond the expected level of this course, and therefore a proof outline is described instead below:

Before receiver receives y^n, it has the only knowledge that there are $2^{nH(X)}$ "randomly selected" sequences to be sent from the transmitter. Then, receiver receives y^n. Receiver increases knowledge about the code word x^n by receiving the sequence y^n. However, still there remains uncertainty, which is averaged over all possible code words. The averaged uncertainty is expressed by the conditional entropy $H(X|Y)$.

Channel Coding Theorem (2)

Proof of (1) Continued:



This means that there are $2^{nH(X|Y)}$ candidate sequences, with which received sequence is most probably y^n, and the probability that the other code words are received in the form of the sequence y^n can be made arbitrarily small, if $R<C$, as shown below:

The probability that is selected sequence, selected from among the all possible length-n Sequences, is *a* code word of the rate R code, is $2^{nR}/2^{nH(X)}$, the average probability that any code word, other than w, is NOT selected is given by:

$$\bar{P}(g(y^n)=w|x^n(w)) = 1 - P(g(y^n)\neq w|x^n(w)) = \left[1-\frac{2^{nR}-1}{2^{nH(X)}}\right]^{2^{nH(X|Y)}} \approx \left[1-\frac{2^{nR}}{2^{nH(X)}}\right]^{2^{nH(X|Y)}}$$

Channel Coding Theorem (2)

Proof of (1) Continued:

$$\approx 1-2^{nH(X|Y)}2^{n(R-H(X))}=1-2^{-n\{(H(X)-H(X|Y)-R)\}}=1-2^{-n\{(C-R)\}}$$

where $C=H(X)-H(X|Y)$ because of the code randomness (maximization with respect to $p(x)$ doesn't have to be taken).

Since the probability described above indicates that the transmitted code word is decoded correctly, the error probability is given by

$$P(g(y^n)\neq w\,|\,x^n(w))\approx 2^{-n\{(C-R)\}}$$

Hence, the average error probability can be made arbitrarily small by making the code length n large enough, if $R<C$.

The probability that can be made arbitrarily small is averaged error probability, and therefore, roughly speaking, the best half of the code words have a maximal error probability less than $2^{-n(C-R)}$, and another half is higher. Throwing away the code words having error probability higher than $2^{-n(C-R)}$, and 2^{nR-1} code words remains. This means that the code rate is changed to: $R-1/n$, because $2^{nR-1}=2^{n(R-1/n)}$. This rate loss is negligible if n is large enough. Then, the maximal error probability satisfies:

$$\lambda^{(n)}\approx 2^{-n\{(C-R)\}}$$

Channel Coding Theorem (3)

Proof of (2):
The proof is comprised of the following 2 steps:
A) $P_e^{(n)}=0$ implies $R\leq C$
B) $P_e^{(n)}\rightarrow 0$ implies $R\leq C$

Proof of A):

Assume that we use $(2^{nR}, n)$ rate R code. There are code words that are selected equi-probably and sent to the receiver; Assume that the message W is to be sent. Then, the entropy of W is:

$$H(W)=nR=H(W|Y^n)+I(W;Y^n)$$

However, by the assumption that $g(Y^n)=W$, $H(W|Y^n)=0$. Hence,

$$H(W)=nR=I(W;Y^n)\underset{(a)}{\leq} I(X^n;Y^n)\underset{(b)}{\leq}\sum_{i=1}^{n}I(X;Y)\underset{(c)}{\leq} nC$$

(a): From the data processing inequality, where Markov chain $W\rightarrow X^n(W)\rightarrow Y^n$ holds.
(b): To be proven in the next slides (Theorem 8.4.1)
(c): By definition of Capacity

Hence, for any codes achieving $P_e^{(n)}=0$, $R\leq C$.

Fano's Inequality for Extension (1)

Proof of B) requires Fano's Inequality for Extension.

Let's define the event: $E = \begin{cases} 1, & if \quad W \neq \hat{W} \\ 0, & if \quad W = \hat{W} \end{cases}$ with $\hat{W} = g(Y^n)$

Then, by using the chain rule, $H(E,W|Y^n) = H(W|Y^n) + \underbrace{H(E|W,Y^n)}_{=0 \quad (a)}$

$$= \underbrace{H(E|Y^n)}_{H(E)\leq 1 \quad (b)} + H(W|E,Y^n)$$

(a): Because E is a function of W and Y^n.
(b): Because E is binary-valued.

Furthermore, $H(W|E,Y^n) = P(E=0)H(W|Y^n,E=0) + P(E=1)H(W|Y^n,E=1)$

$$\leq \Pr(\hat{W} = W)\times 0 + \Pr(\hat{W} \neq W)\log(|W|-1) \leq P_e^{(n)} nR$$

where $P_e^{(n)} = \Pr(\hat{W} \neq W)$

Combining all, we have $H(W|Y^n) \leq 1 + P_e^{(n)} nR$ This is Fano's inequality.

However, because $X^n(W)$ is a function of W, $H(X^n(W)|Y^n) \leq H(W|Y^n)$

Then, we have Fano's inequality for extension: $H(X^n|Y^n) \leq 1 + P_e^{(n)} nR$

Fano's Inequality for Extension (2)

Theorem 8.4.1: $I(X^n;Y^n) \leq nC, \quad for \quad any \quad p(x)$

Proof: $I(X^n;Y^n) = H(Y^n) - H(Y^n \mid X^n)$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,Y_2,\cdots,Y_{i-1},X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$$

because the channel is memory-less. However,

$$H(Y^n) \leq \sum_{i=1}^{n} H(Y_i)$$

Therefore, $I(X^n;Y^n) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i;Y_i) \leq nC$

With this result, the proof of A)-of-(2) of the channel coding theorem is completed.

Fano's Inequality for Extension (3)

Proof of B)

We now know that $H(W) = nR = H(W|Y^n) + I(W;Y^n)$

But we don't assume the fact $g(Y^n) = W$ is known, in this case. Hence, $H(W|Y^n) > 0$.

$$nR = H(W|Y^n) + I(W;Y^n) \le H(W|Y^n) + I(X^n(W);Y^n)$$

$$\le 1 + P_e^{(n)} nR + I(X^n(W);Y^n) \le 1 + P_e^{(n)} nR + nC$$

where we have used Fano's inequality for extension. Dividing the both sides by n,

$$R \le \underbrace{P_e^{(n)} R}_{\substack{\to 0 \quad n\to\infty \\ \text{by assumption}}} + \underbrace{\frac{1}{n}}_{\to 0 \quad n\to\infty} + C$$

Finally, we know that $R \le C$ has to be satisfied.

We also know that $P_e^{(n)} \ge 1 - \frac{C}{R} - \frac{1}{nR}$

Summary

We have visited.....

1. Channel Capacity
   - Definition
   - Some Examples
   - Some Properties
2. Random Coding
3. Channel Coding Theorem
   - Proof for Sufficiency
4. Fano's Inequality for Extension
   - Proof for Necessity