INFORMATION THEORY & CODING

Dr. Rui Wang

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

Email: wang.r@sustech.edu.cn

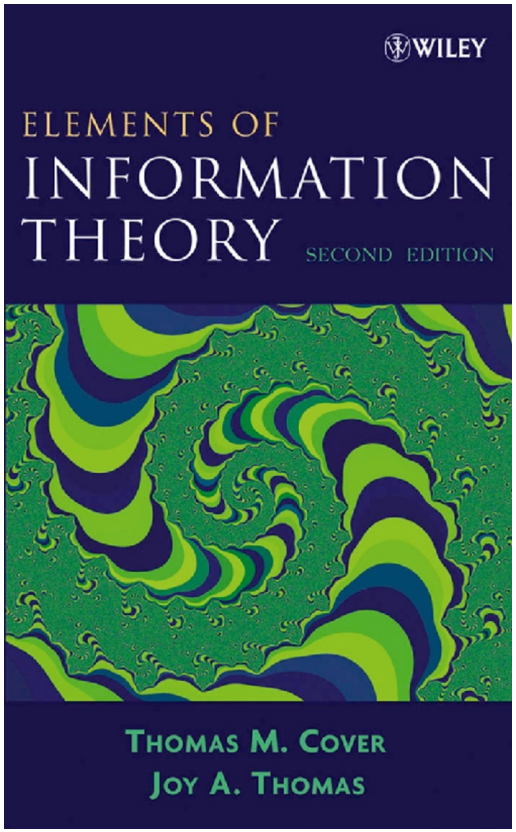
September 7, 2021

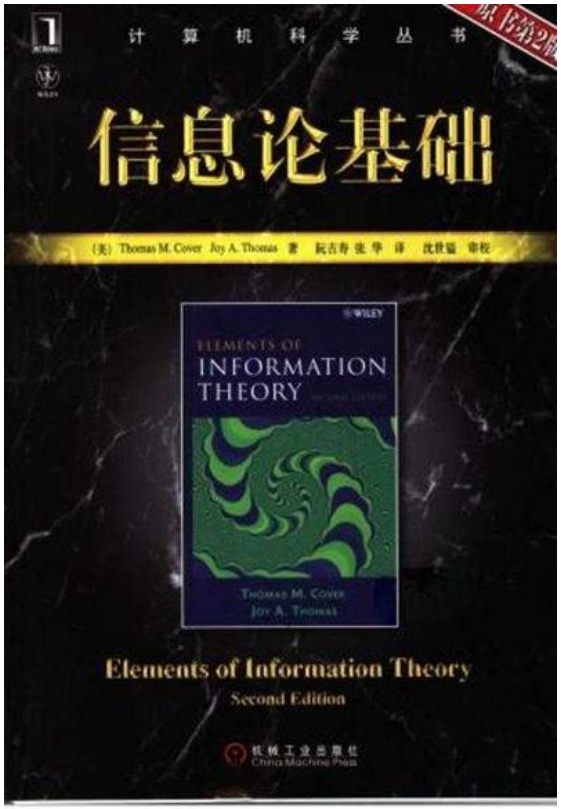
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Textbooks and References

Textbooks

- Thomas M. Cover, Joy A. Thomas, Elements of Information Theory, 2^{nd} Edition, Wiley-Interscience, 2006.



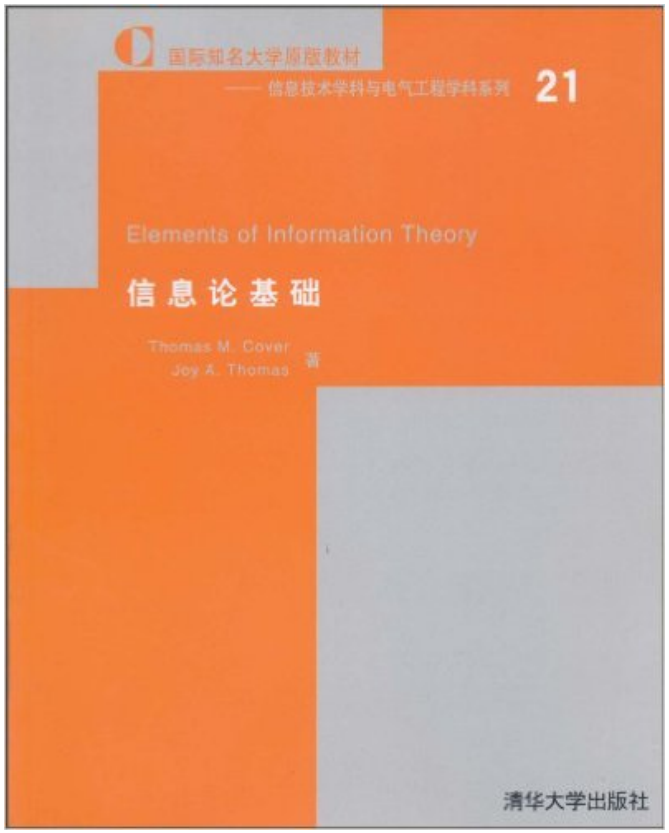


南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Textbooks and References

Textbooks

- Thomas M. Cover, Joy A. Thomas, Elements of Information Theory, 1^{st} Edition, Tsinghua Unviersity Press, 2003.



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Assessment

- Quiz: starts from the 3rd week, open book, around 10min, almost every week.
- Homework: starts from the 2nd week, every week, submits to BB.
- Project: report + Matlab simulation (if necessary).
- Final Exam.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Policy Reminders

Academic dishonesty consists of misrepresentation by deception or by other fraudulent means and can result in serious consequences, e.g. the grade of zero on an assignment, loss of credit with a notation on the transcript ("Grade of F assigned for academic dishonesty").

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Note to You

- These lecture notes are a perpetual work in progress. Please report any typo or other errors by email. Thanks!
- We try to prepare there lecture notes carefully, but they are NOT intended to replace the textbook.
- For more information, please refer to BB or eee.sustech.edu.cn/p/wangrui.
- Office hours and tutorials: discuss with TAs in QQ group (101147374)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History *

1877 - Showed that thermodynamic entropy is related to the statistical distribution of molecular configurations, with increasing entropy corresponding to increasing randomness.

$$S = k_B \log W \tag{1}$$

$$\text{where } W = N! \prod_i \frac{1}{N_i!}.$$

Ludwig Boltzman (1844-1906)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

1924 - Nyquist rate and reconstruction of bandlimited signals from their samples. Also stated formula $R = K \log m$, where R is the rate of transmission, K is a measure of the number of symbols per second and m is the number of message amplitudes available. Amount of information that can be transmitted is proportional to the product of bandwidth and time of transmission.

Harry Nyquist
(1889-1976)

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

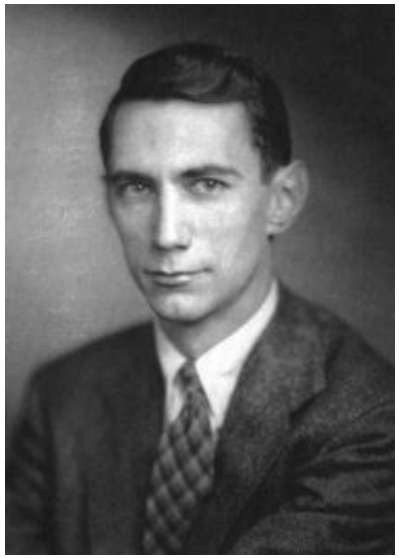
1928 - (inventor of the oscillator) - in the paper entitled "Transmission of Information" proposed formula $H = n\log s$, where H is the "information" of the message, s is the number of possible symbols, n is the length of the message in symbols.

Ralph V. L. Hartley
(1888-1970)

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

1938 - In his Master's thesis *A Symbolic Analysis of Relay and Switching Circuits at MIT*, he demonstrated that electrical application of Boolean algebra could construct and resolve any logical, numerical relationship.



Claude E. Shannon

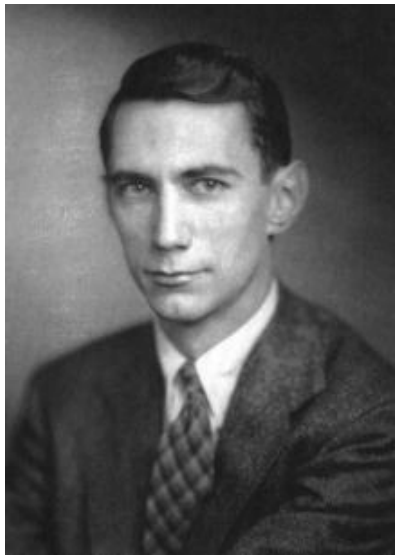
(Apr. 30, 1916 - Feb. 24, 2001)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

1938 - In his Master's thesis *A Symbolic Analysis of Relay and Switching Circuits at MIT*, he demonstrated that electrical application of Boolean algebra could construct and resolve any logical, numerical relationship.

"possibly the most important, and also the most famous, master's thesis of the century."



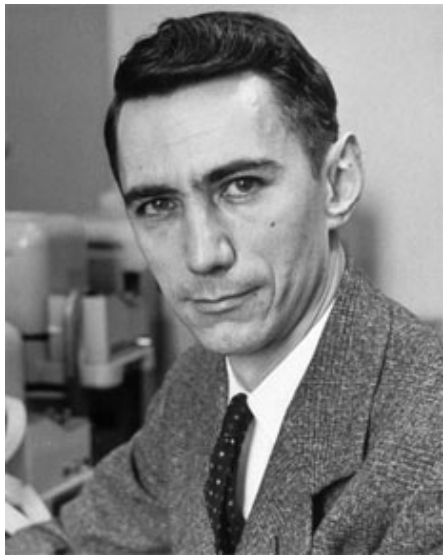
Claude E. Shannon

(Apr. 30, 1916 - Feb. 24, 2001)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

1948 - efficient source representation, reliable information transmission, digitalization - foundation of communication and information theory. Made the startling discovery that arbitrarily reliable communications are possible at non-zero rates. Prior to Shannon, it was believed that in order to get arbitrarily low probability of error, the transmission rate must go to zero. His paper *"A Mathematical Theory of Communications"* proved to be the foundation of modern communication theory.



Claude E. Shannon
(Apr. 30, 1916 - Feb. 24, 2001)

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History

A Mathematical Theory of Communication

By C. E. SHANNON

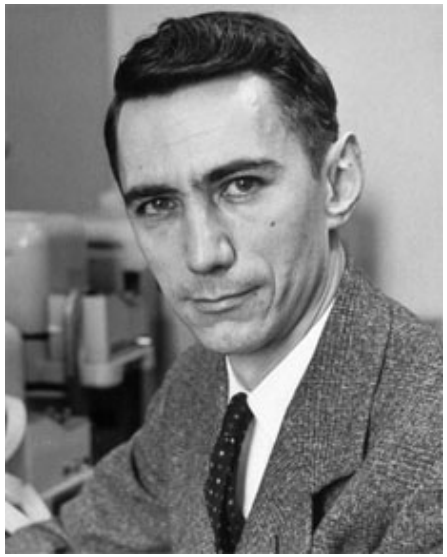
INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:



Claude E. Shannon

(Apr. 30, 1916 - Feb. 24, 2001)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Quotes

"What made possible, what induced the development of coding as a theory, and the development of very complicated codes, was Shannon's Theorem: he told you that it could be done, so people tried to do it." - Robert Fano

"Before 1948, there was only the fuzziest idea of a message was. There was some rudimentary understanding of how to transmit a waveform and process a received waveform, but there was essentially no understanding of how to turn a message into a transmitted waveform." - Robert Gallager

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Quotes

"To make the chance of error as small as you wish? Nobody had ever thought of that. How he got that insight, how he even came to believe such a thing, I don't know. But almost all modern communication engineering is based on that work." - Robert Fano

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Brief History (cont')

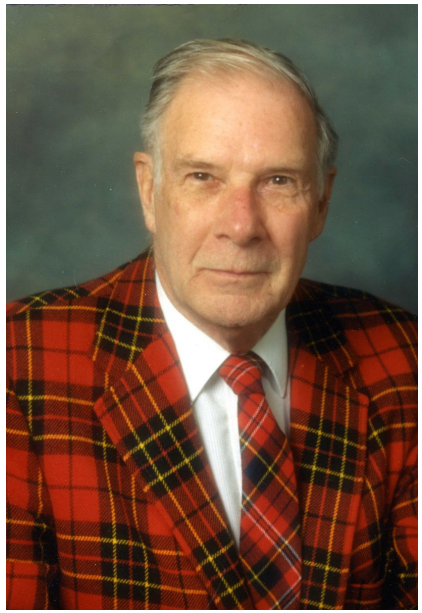
1950 R. Hamming - Developed a family of error-correcting codes

1952 D. Huffman - Efficient source encoding

1950-60's Muller, Reed, Solomon, Bose, Ray-Chaudhuri, Hocquenghem - Algebraic Codes

1970's Fano, Viterbi - Convolutional Codes

1990's Berrou, Glavieux, Gallager, Lin - Near capacity achieving coding schemes: Turbo Codes, Low-Density Parity Check Codes

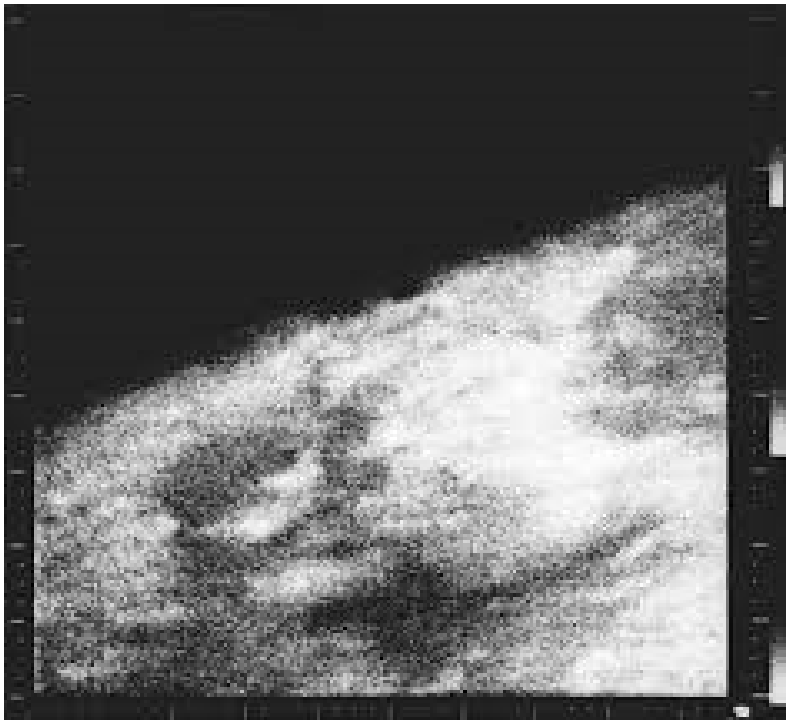


Richard W. Hamming (1915-1998)

2008 E. Arikan - First practical construction of codes achieving capacity for a wide array of channels: Polar Codes

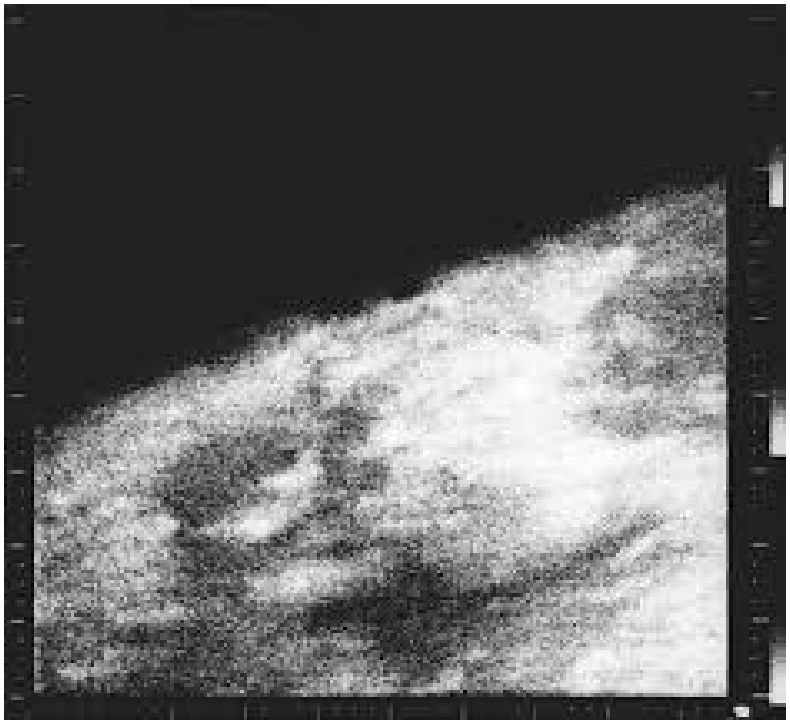
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

An example



Mars, Mariner IV, '64 using
no coding

An example

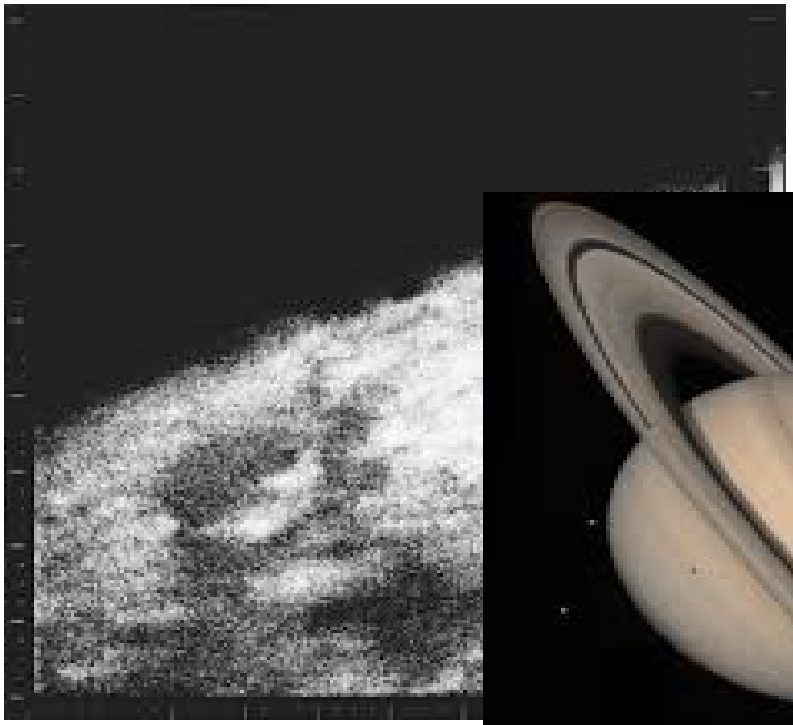


Mars, Mariner IV, '64 using
no coding

Mars, Mariner VI, '69 using
Reed-Muller coding

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

An example



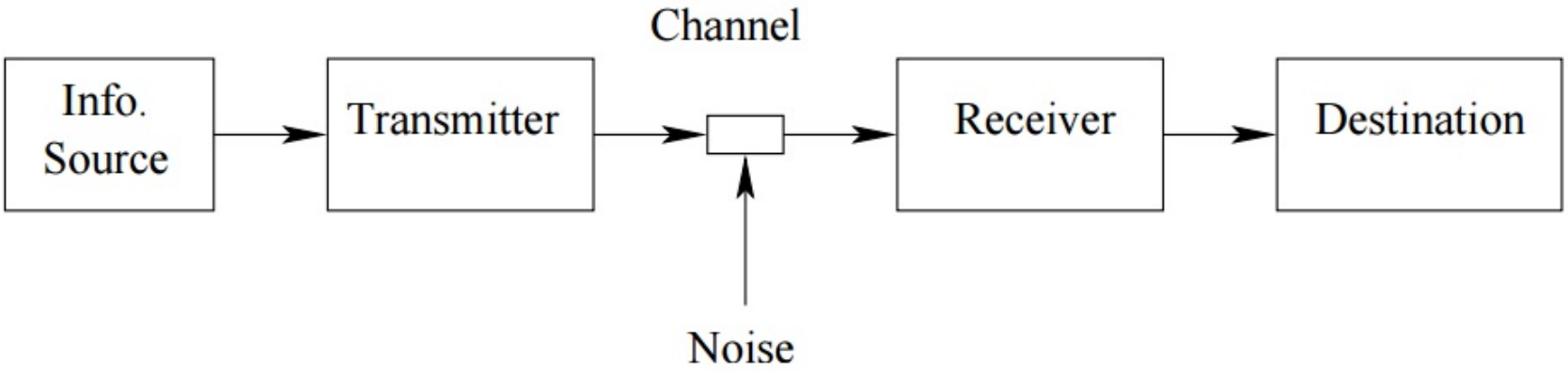
Mars, Mariner IV, no codin

Saturn, Voyager, '71 using Golay coding

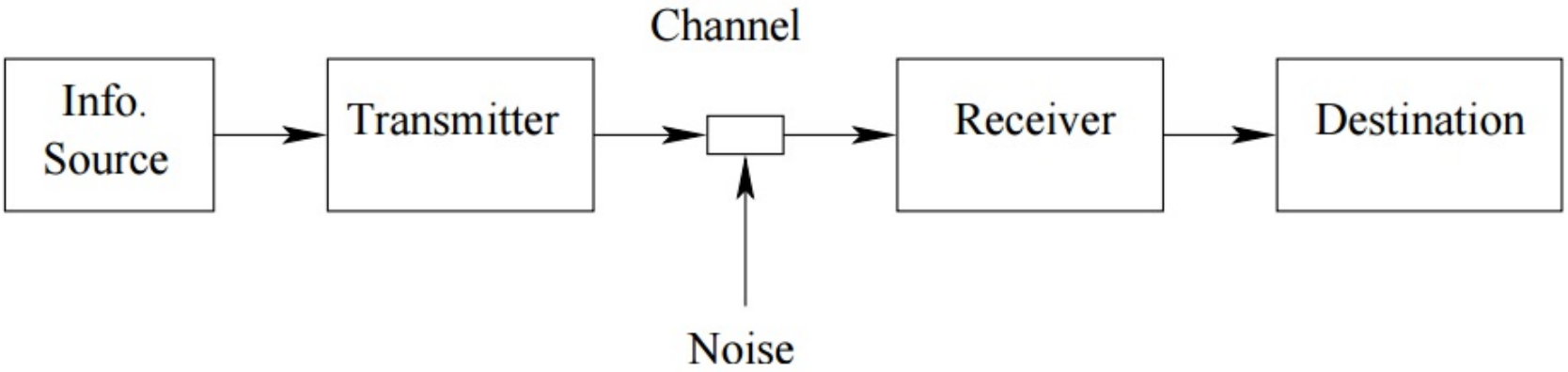
riner VI, '69 using Muller coding

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Communication System



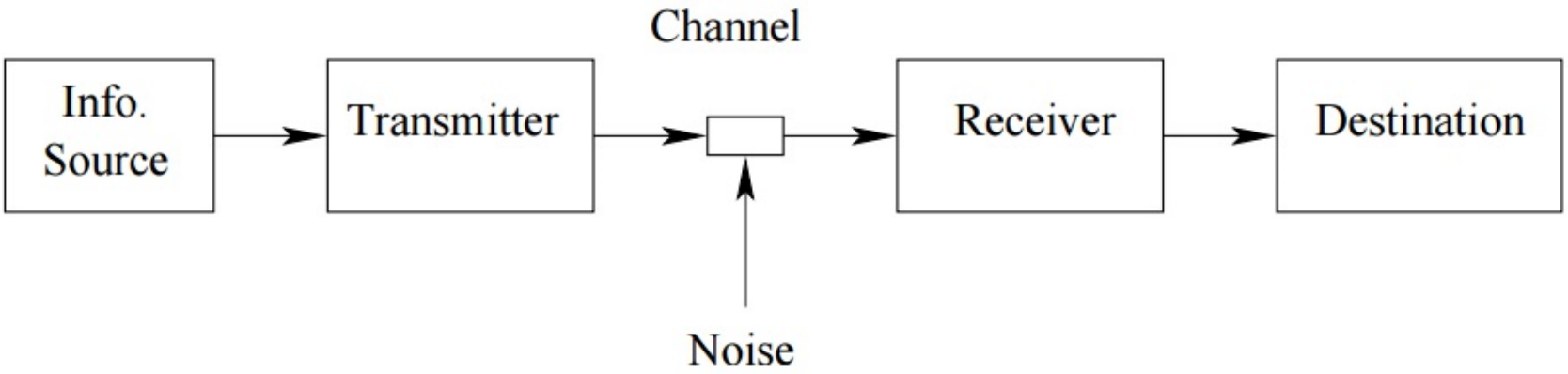
A Communication System



- Info. Source: any source of data we wish to transmit or store
- Transmitter: mapping data source to the channel alphabet in an efficient manner
- Receiver: mapping from channel to data to ensure "reliable" reception
- Destination: data sink

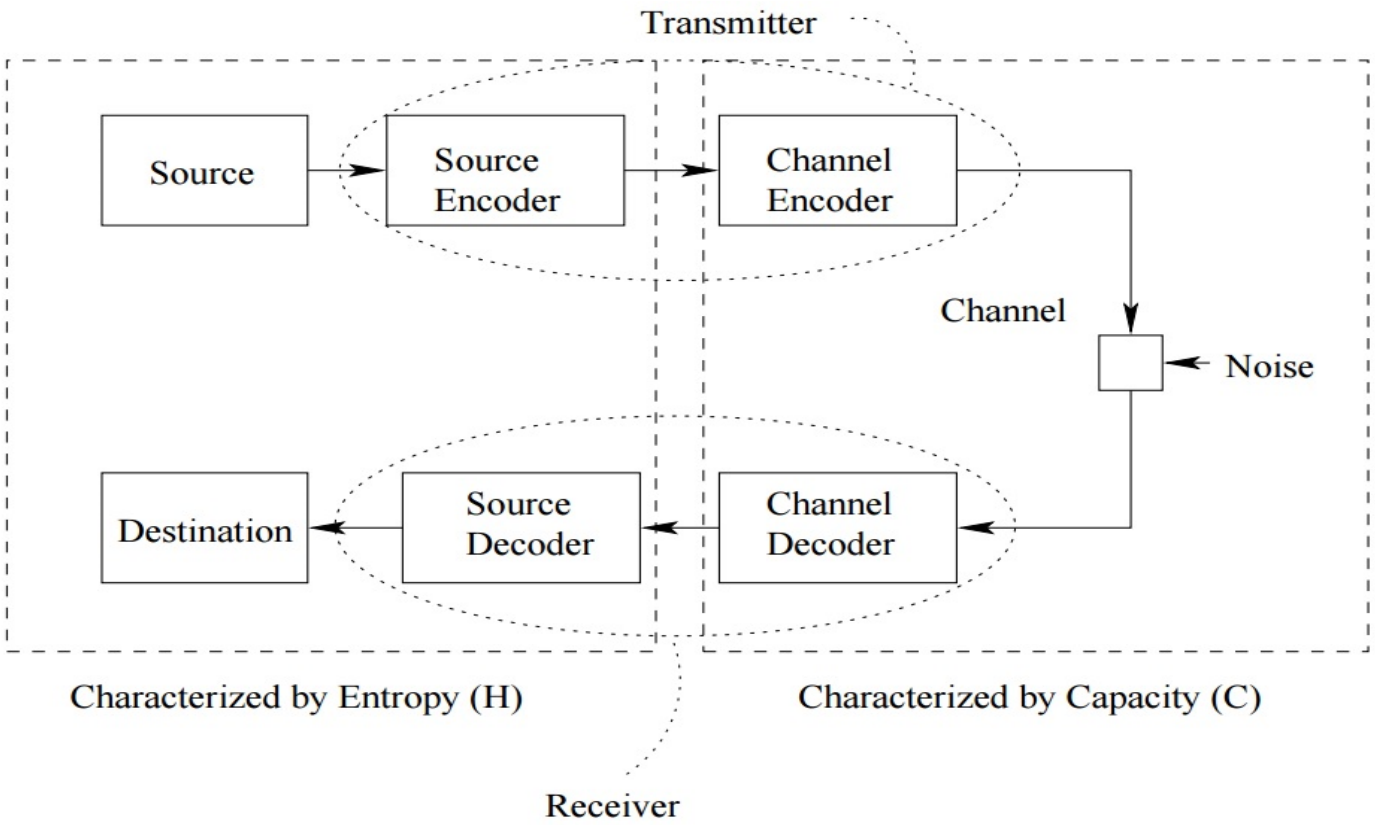
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Communication System



Question: Under what conditions can the output of the source be conveyed *reliably* to the destination? What is reliable? Low prob. of error? Low distortion?

An Expanded Communication System



What is the ultimate data compression (answer: the entropy H)? What is the ultimate transmission rate of communication (answer: channel capacity C)?

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Encoders

Source Encoder

- map from source to bits
- "matched" to the information source
- Goal: to get an *efficient* representation of the source (i.e., least number of bits per second, minimum distortion, etc.)

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Encoders

Source Encoder

- map from source to bits
- "matched" to the information source
- Goal: to get an *efficient* representation of the source (i.e., least number of bits per second, minimum distortion, etc.)

Channel Encoder

- map from bits to channel
- depends on channel available (channel model, bandwidth, noise, distortion, etc.) In communication theory, we work with hypothetical channels which in some way capture the essential features of the physical world.
- Goal: to get *reliable* communication

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

- Goal: To get an efficient representation (i.e., small number of bits) of the source on average.

Source Encoder: Examples

- Goal: To get an efficient representation (i.e., small number of bits) of the source on average.

Example 1: An urn contains 8 numbered balls. One ball is selected. How many binary symbols are required to represent the outcome?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

- Goal: To get an efficient representation (i.e., small number of bits) of the source on average.

Example 1: An urn contains 8 numbered balls. One ball is selected. How many binary symbols are required to represent the outcome?

Outcome	1	2	3	4	5	6	7	8
Representation	000	001	010	011	100	101	110	111

Answer: Require 3 bits to represent any given outcome.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

Example 2: Consider a horse race with 8 horses. It was determined that the probability of horse i winning is

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right) \tag{2}$$

Source Encoder: Examples

Example 2: Consider a horse race with 8 horses. It was determined that the probability of horse i winning is

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

Answer 1: Let's try the code of the previous example.

Outcome	Probability	Representation 1
0	$\frac{1}{2}$	000
1	$\frac{1}{4}$	001
2	$\frac{1}{8}$	010
3	$\frac{1}{16}$	011
4	$\frac{1}{64}$	100
5	$\frac{1}{64}$	101
6	$\frac{1}{64}$	110
7	$\frac{1}{64}$	111

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

Example 2: Consider a horse race with 8 horses. It was determined that the probability of horse i winning is

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

Answer 1: Let's try the code of the previous example.

Outcome	Probability	Representation 1
0	$\frac{1}{2}$	000
1	$\frac{1}{4}$	001
2	$\frac{1}{8}$	010
3	$\frac{1}{16}$	011
4	$\frac{1}{64}$	100
5	$\frac{1}{64}$	101
6	$\frac{1}{64}$	110
7	$\frac{1}{64}$	111

To represent a given outcome, the average number of bits is $\overline{\ell} = 3$.

Source Encoder: Examples

Example 2: Consider a horse race with 8 horses. It was determined that the probability of horse i winning is

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

Answer 2: What if we allow the length of each representation to vary amongst the outcomes, e.g., a Huffman code:

Outcome	Probability	Representation 2
0	$\frac{1}{2}$	0
1	$\frac{1}{4}$	10
2	$\frac{1}{8}$	110
3	$\frac{1}{16}$	1110
4	$\frac{1}{64}$	111100
5	$\frac{1}{64}$	111101
6	$\frac{1}{64}$	111110
7	$\frac{1}{64}$	111111

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

Example 2: Consider a horse race with 8 horses. It was determined that the probability of horse i winning is

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

Answer 2: What if we allow the length of each representation to vary amongst the outcomes, e.g., a Huffman code:

Outcome	Probability	Representation 2
0	$\frac{1}{2}$	0
1	$\frac{1}{4}$	10
2	$\frac{1}{8}$	110
3	$\frac{1}{16}$	1110
4	$\frac{1}{64}$	111100
5	$\frac{1}{64}$	111101
6	$\frac{1}{64}$	111110
7	$\frac{1}{64}$	111111

The average number of bits is

$$\begin{aligned}\bar{\ell} =& \frac{1}{2} + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 \\ &+ \frac{4}{64}\cdot 6 \\ =& 2\end{aligned}$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

Definition: The source **entropy**, $H(X)$ of a random variable X with a probability mass function $p(x)$, is defined as

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

As we will show later in the course, the most effcient representation has average codeword length $\bar{\ell}$ as

$$H(X) \leq \bar{\ell} < H(X) + 1$$

$$\Pr[\text{horse } i \text{ wins}] = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

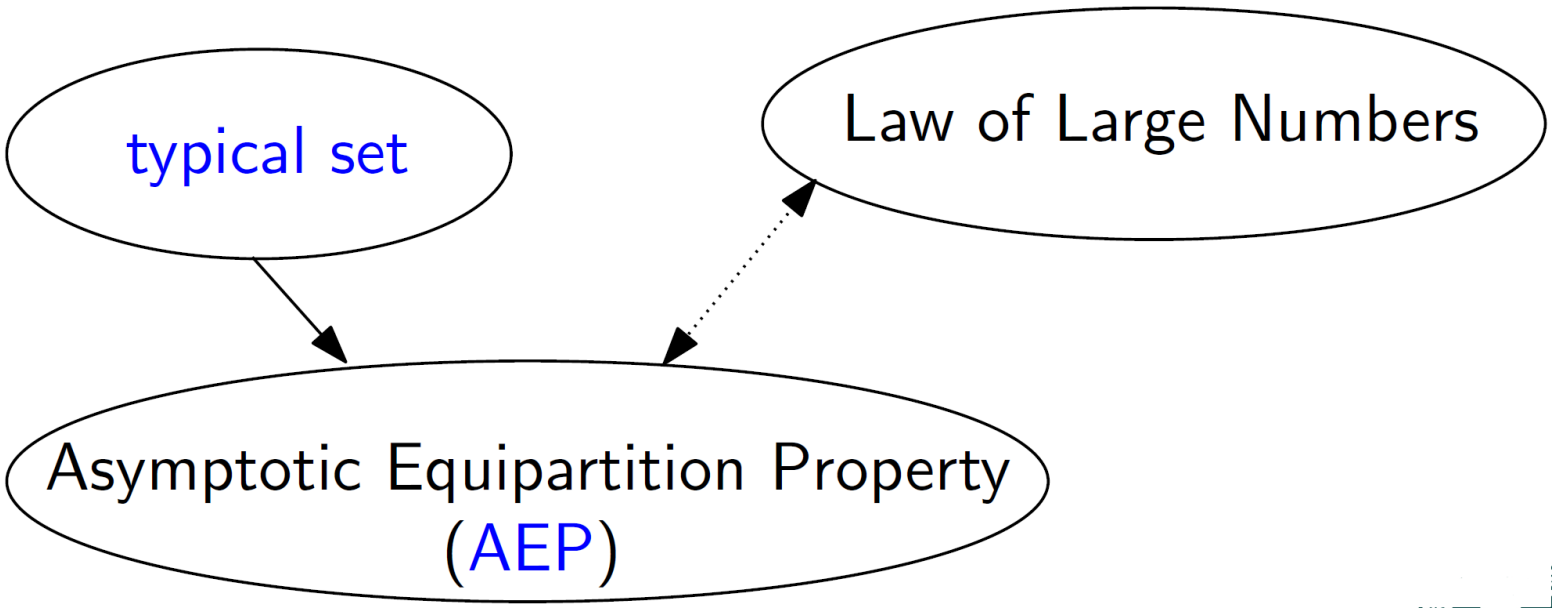
$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{16}\log 16 + \frac{4}{64}\log 64 = 2$$

The Huffman code is optimal!

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Source Encoder: Examples

- Information theory and coding deal with the "typical" or expected behavior of the source.
- Entropy is a measure of the *average* uncertainty associated with the source.



Channel Encoder

- Goal: To achieve an economical (high rate) and reliable (low probability of error) transmission of bits over a channel.

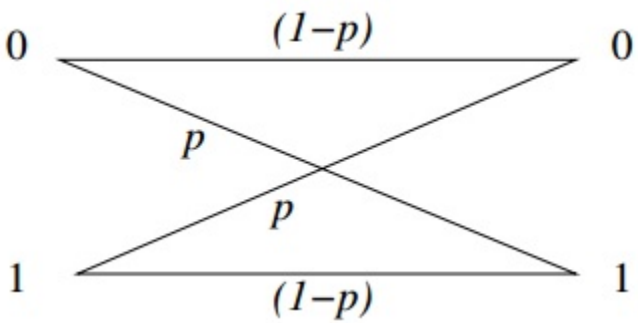
With a channel code we add *redundancy* to the transmitted data sequence which allows for the correction of errors that are introduced by the channel.

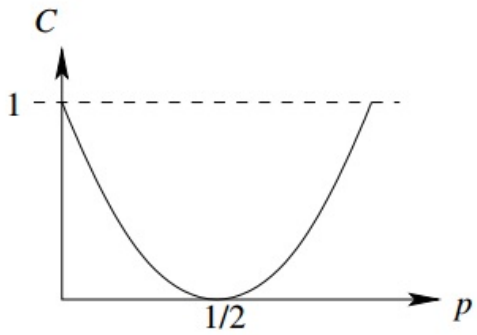
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Channel Encoder

Shannon's Channel Coding Theorem There is a quantity called the *capacity*, C, of a channel such that for every rate $R < C$ there exists a sequence of $\left(\underbrace{2^{nR}}_{\#\text{codewords}}, \underbrace{n}_{\#\text{chan. uses}} \right)$ codes such that $\Pr[error] \to 0$ as $n \to \infty$. Conversely, for any code, if $\Pr[error] \to 0$ as $n \to 1$ then $R \leq C$.

Example: binary Symmetric Channel





- Input channel alphabet = Output channel alphabet = $\{0, 1\}$
- Assume *independent* channel uses (i.e., memoryless)
- Channel randomly flips the bit *with probability p*
- For $p = 0$ or $p = 1$, $C = 1$ bits/channel use (*noiseless channel* or *inversion channel*)
- Worst case: p = 1/2, in which case the input and the output are *statistically independent* (C = 0)
- *Question*: How do we *devise codes* which perform well on this channel?

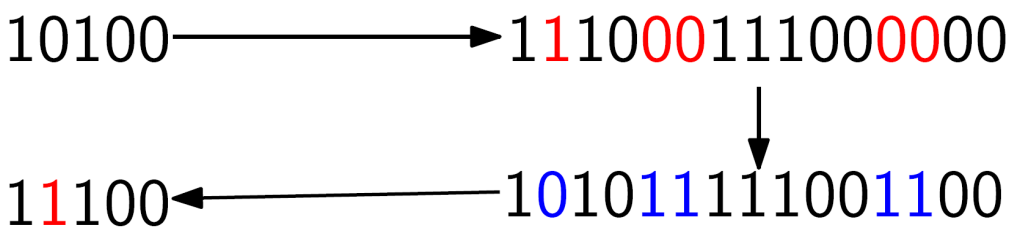
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Repetition Code

- In this code, we repeat one bit *odd times*. The code consists of two possible codewords:

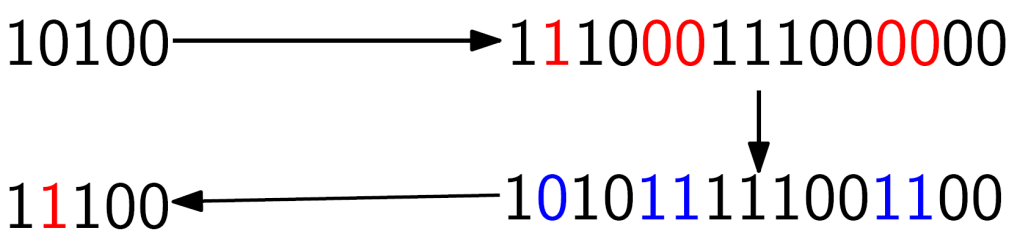
$$\mathcal{C} = \{000\cdots0, 111\cdots1\}$$

- Decoding by a *majority voting* scheme: if there are more 0's than 1's then declare 0, otherwise 1.
- Suppose that $R = 1/3$, i.e., the source output can be encoded before transmission by repeating each bit *three times*.
  Example:



Repetition Code

Example:



The *bit error probability* Pr_e is:

$$\begin{aligned}\mathrm{Pr}_e &= \mathrm{Pr}[\text{2channel errors}] + \mathrm{Pr}[\text{3channel errors}] \\ &= 3p^2(1-p) + p^3 \\ &= 3p^2 - 2p^3\end{aligned}$$

If $p \leq 1/2$, Pr_e is less than p. So, the repetition code **improves** the channel's reliability. And for **small** p, the improvement is dramatic.

Repetition Code

For $R = 1/3$, the *bit error probability* Pr_e is:

$$\mathrm{Pr}_e = 3p^2 - 2p^3.$$

For $R = 1/(2m+1)$, the *bit error probability* Pr_e is:

$$\begin{aligned}
\mathrm{Pr}_e &= \sum_{k=m+1}^{2m+1} \Pr[k \text{ errors out of } 2m+1 \text{ transmitted bits}] \\
&= \sum_{k=m+1}^{2m+1} \binom{2m+1}{k} p^k (1-p)^{2m+1-k} \\
&= \binom{2m+1}{m+1} p^{m+1} + \text{terms of higher degree in } p.
\end{aligned}$$

Thus, $\mathrm{Pr}_e \to 0$ as $m \to 1$. However, $R \to 0$! Repetition code is NOT efficient! Shannon demonstrated that there exist codes which are *capacity achieving* at non-zero rates.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **Discrete Random Variables**

  A *discrete random variable* is used to model a "random experiment" with a finite or countable number of possible outcomes. For example, the toss of a coin, the roll of a die, or the count of the number of telephone calls during a given time, etc.

  The *sample space* $\mathcal{S}$, of the experiment is the set of all possible outcomes and contains a finite or countable number of elements. Let $\mathcal{S} = \{\zeta_1, \zeta_2, \cdots\}$.

  An *event* is a subset of $\mathcal{S}$. Events consisting a single outcome are called *elementary events*.

- **Discrete Random Variables**

  Let X be a random variable with sample space $\mathcal{S}_X$. A *probability mass function (pmf)* for X is a mapping $p_X : \mathcal{S}_X \to [0, 1]$ from $\mathcal{S}_X$ to the closed unit interval [0, 1] satisfying

$$\sum_{x \in \mathcal{S}_X} p_X(x) = 1, \tag{3}$$

  where the number $p_X(x)$ is the *probability* that the outcome of the given random experiment is x, i.e., $p_X(x) = \Pr[X = x]$.

- **Discrete Random Variables**

  Let X be a random variable with sample space $\mathcal{S}_X$. A *probability mass function (pmf)* for X is a mapping $p_X : \mathcal{S}_X \to [0, 1]$ from $\mathcal{S}_X$ to the closed unit interval [0, 1] satisfying

$$\sum_{x \in \mathcal{S}_X} p_X(x) = 1, \tag{4}$$

  where the number $p_X(x)$ is the *probability* that the outcome of the given random experiment is x, i.e., $p_X(x) = \Pr[X = x]$.

  Every event $A \subseteq \mathcal{S}$ has a probability $p(A) \in [0, 1]$ satisfying the following:

  1. $p(A) \geq 0$
  2. $p(\mathcal{S}) = 1$
  3. for $A, B \subseteq \mathcal{S}$, $p(A \cup B) = p(A) + p(B)$ if $A \cap B = \emptyset$

- **Vector Random Variables**

  If the elements of sample space $\mathcal{S}_Z$ are vectors of real numbers, then Z is a *(real) vector random variable*.

  Suppose Z is a vector random variable with $Z = (X, Y)$, where X and Y are both discrete random variables. In its sample space, each elements has two components, i.e.,
  $\mathcal{S}_z = \{z_1, z_2, \cdots\} = \{(x_1, y_1), (x_2, y_2), \cdots\}$.

  The *projection* of $\mathcal{S}_Z$ on its first coordinate is

  $$\mathcal{S}_X = \{x | \forall (x, y) \in \mathcal{S}_Z\}.$$

  Example: If $Z = (X, Y)$ and $\mathcal{S}_Z = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$, then $\mathcal{S}_X = \mathcal{S}_Y = \{0, 1\}$.

- **Vector Random Variables**

  The *pmf* of a vector random variable $Z = (X, Y)$ is also called the *joint pmf* of X and Y, and is denoted by

$$p_Z(x, y) = p_{X,Y}(x, y) = \Pr(X = x, Y = y),$$

  where the comma in the last equation denotes a logical 'AND' operation.

- **Vector Random Variables**

  The *pmf* of a vector random variable $Z = (X, Y)$ is also called the *joint pmf* of X and Y, and is denoted by

  $$p_Z(x, y) = p_{X,Y}(x, y) = \Pr(X = x, Y = y),$$

  where the comma in the last equation denotes a logical 'AND' operation.

  From $p_{X,Y}(x, y)$, we can find marginal pmf $p_X(x)$ as

  $$p_X(x) \equiv \Pr(X = x) = \sum_{y \in \mathcal{S}_Y} p_{X,Y}(x, y);$$

  and similarly,

  $$p_Y(y) \equiv \Pr(Y = y) = \sum_{x \in \mathcal{S}_X} p_{X,Y}(x, y); \quad (5)$$

Conditional Probability

- Let A and B be events, with $\Pr[A] > 0$. The *conditional probability* of B given that A *occured* is

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}.$$

Conditional Probability

- Let A and B be events, with $\Pr[A] > 0$. The *conditional probability* of B given that A *occured* is

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}.$$

Thus, $\Pr[A|A] = 1$, and $\Pr[B|A] = 0$ if $A \cap B = \emptyset$.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Conditional Probability

- Let A and B be events, with $\Pr[A] > 0$. The *conditional probability* of B given that A *occured* is

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

Thus, $\Pr[A|A] = 1$, and $\Pr[B|A] = 0$ if $A \cap B = \emptyset$.

If $Z = (X, Y)$ and $p_X(x_k) > 0$, then

$$\begin{aligned} p_{Y|X}(y_j|x_k) &= \Pr[Y = y_j|X = x_k] \\ &= \frac{\Pr[X = x_k, Y = y_j]}{\Pr[X = x_k]} \\ &= \frac{p_{X,Y}(x_k, y_j)}{p_X(x_k)}. \end{aligned}$$

Conditional Probability

- If $Z = (X, Y)$ and $p_X(x_k) > 0$, then

$$p_{Y|X}(y_j|x_k) = \frac{p_{X,Y}(x_k, y_j)}{p_X(x_k)}.$$

Then random variables X and Y are *independent* if

$$\forall (x, y) \in \mathcal{S}_{X,Y},\ p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Conditional Probability

- If $Z = (X, Y)$ and $p_X(x_k) > 0$, then

$$p_{Y|X}(y_j|x_k) = \frac{p_{X,Y}(x_k, y_j)}{p_X(x_k)}.$$

Then random variables X and Y are *independent* if

$$\forall (x, y) \in \mathcal{S}_{X,Y}, \; p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

If X and Y are *independent*, then

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x),$$

and

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y),$$

Expected Value

- If X is a random variable, the *expected value* (or mean) of X, denoted by $E[X]$, is

$$E[X] = \sum_{x \in \mathcal{S}_X} x p_X(x).$$

Then *expected value* of the random variable $f(X)$ is

$$E[f(X)] = \sum_{x \in \mathcal{S}_X} f(x) p_X(x).$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Expected Value

- If X is a random variable, the *expected value* (or mean) of X, denoted by $E[X]$, is

$$E[X] = \sum_{x \in \mathcal{S}_X} x p_X(x).$$

Then *expected value* of the random variable $f(X)$ is

$$E[f(X)] = \sum_{x \in \mathcal{S}_X} f(x) p_X(x).$$

In particular, $E[X^n]$ is the *n-th moment* of X. The *variance* of X is

$$VAR[X] = E[X^2] - E[X]^2.$$