

INFORMATION THEORY & CODING

Dr. Rui Wang

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

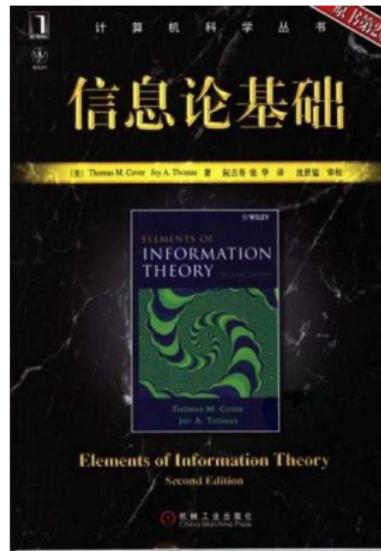
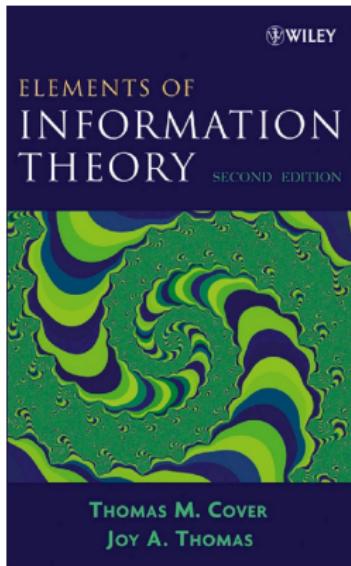
Email: wang.r@sustech.edu.cn

September 11, 2023



Textbooks and References

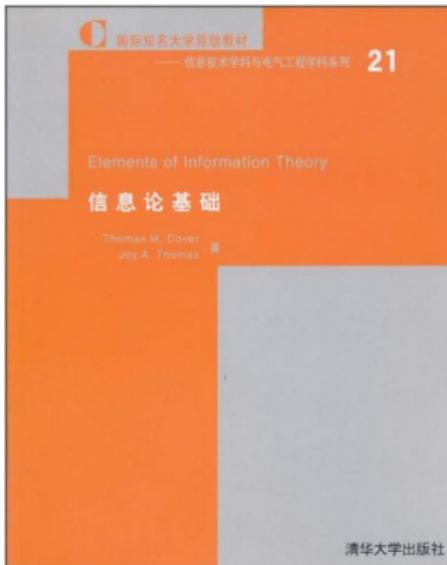
- Thomas M. Cover, Joy A. Thomas, **Elements of Information Theory**, 2nd Edition, Wiley-Interscience, 2006.



Textbooks and References

Textbooks

- Thomas M. Cover, Joy A. Thomas, **Elements of Information Theory**, 1st Edition, Tsinghua University Press, 2003.



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**.



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”).



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.
- Office hours and tutorials: discuss with TAs in QQ group



A Brief History *

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

$$S = k_B \log W \quad (1)$$

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



Ludwig Boltzmann
(1844-1906)

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)



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)



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)



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."



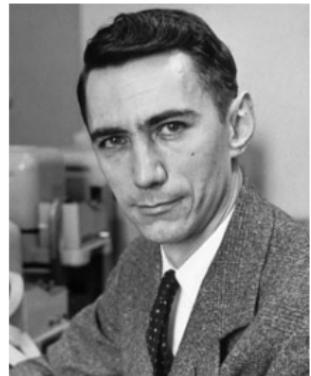
Claude E. Shannon

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



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)



A Brief History

A Mathematical Theory of Communication

By C. E. SHANNON

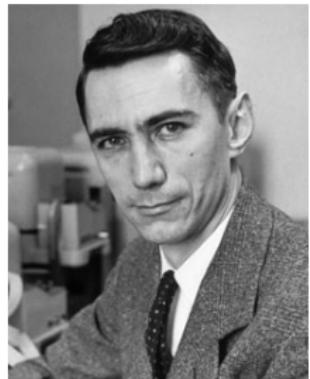
INTRODUCTION

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 Nyquist¹ and Hartley² 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)

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



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



A Brief History (cont')

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

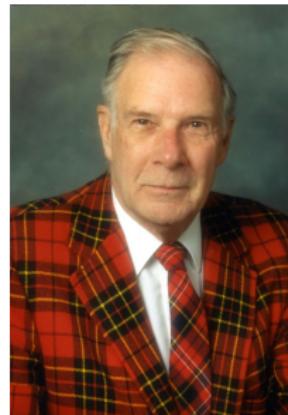
1952 D. Huffman - Efficient source encoding

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

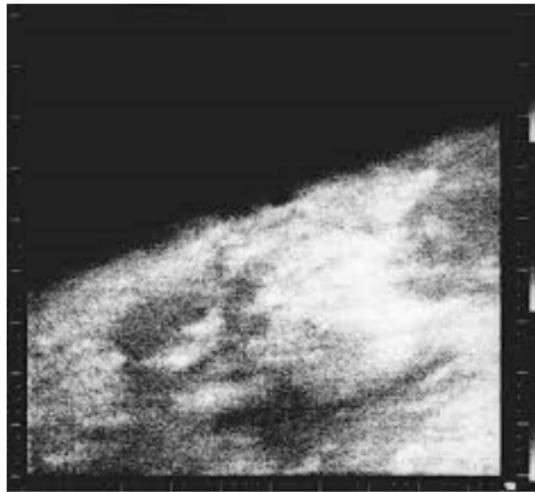
2008 E. Arikan - First practical construction of codes achieving capacity for a wide array of channels: Polar Codes



Richard W. Hamming
(1915-1998)

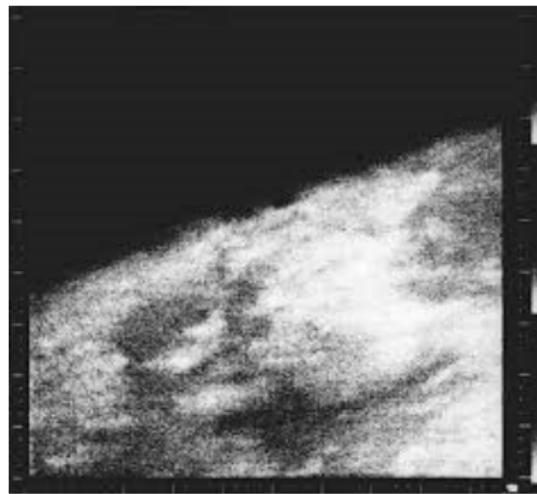


An example



Mars, Mariner IV, '64 using
no coding

An example

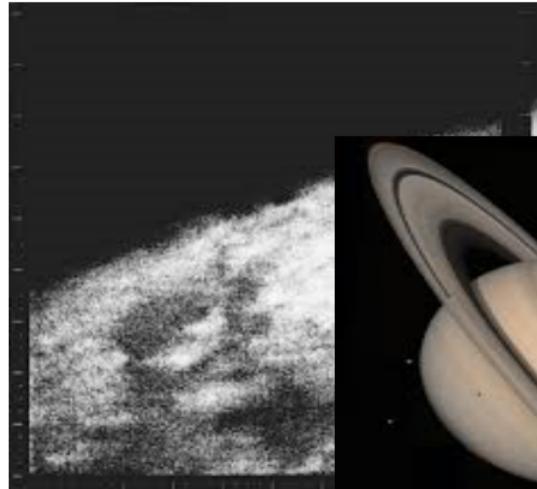


Mars, Mariner IV, '64 using
no coding



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

An example



Mars, Mariner IV,
no coding



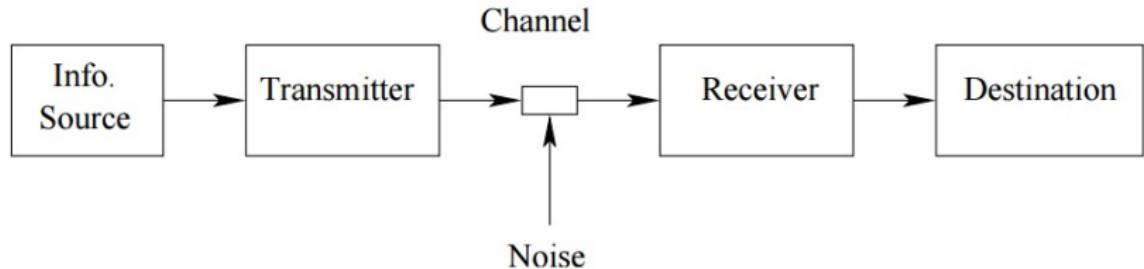
Saturn, Voyager, '71 using
Golay coding



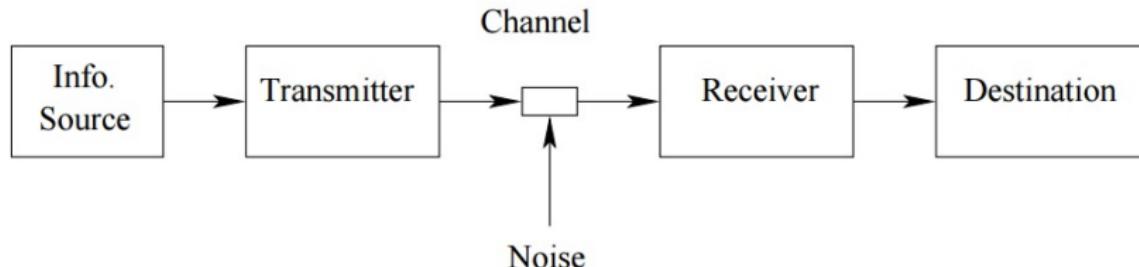
Mariner VI, '69 using
Muller coding



A Communication System

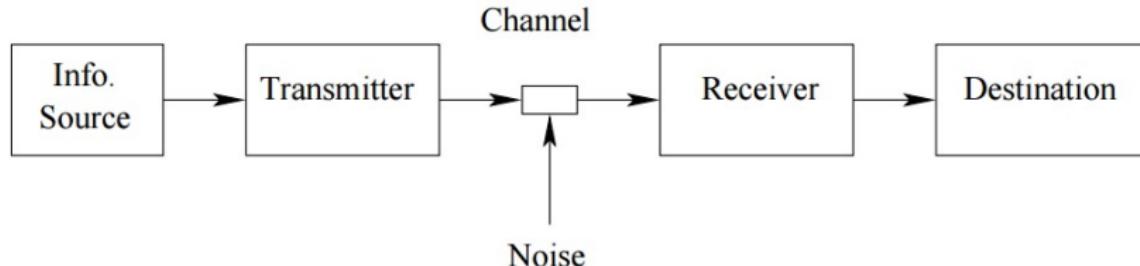


A Communication System



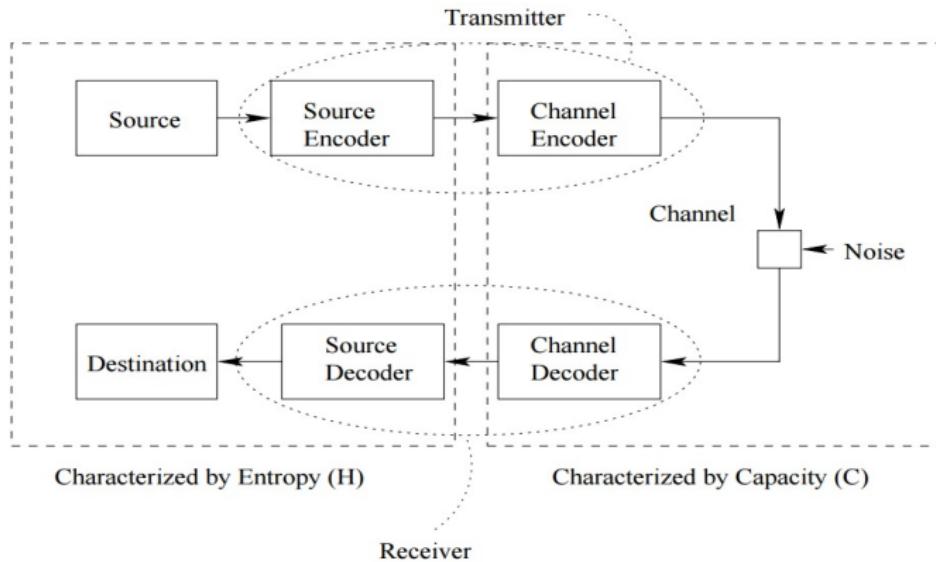
- **Info. Source:** any source of data we wish to transmit or store
- **Transmitter:** mapping data source to the channel alphabet in an efficient manner
- **Receiver:** mapping from channel to data to ensure “reliable” reception
- **Destination:** data sink

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**)?

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.)



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



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?



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.



Source Encoder: Examples

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

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



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



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 $\bar{\ell} = 3$.



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



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

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



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 efficient** representation has average codeword length \bar{l} as

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

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

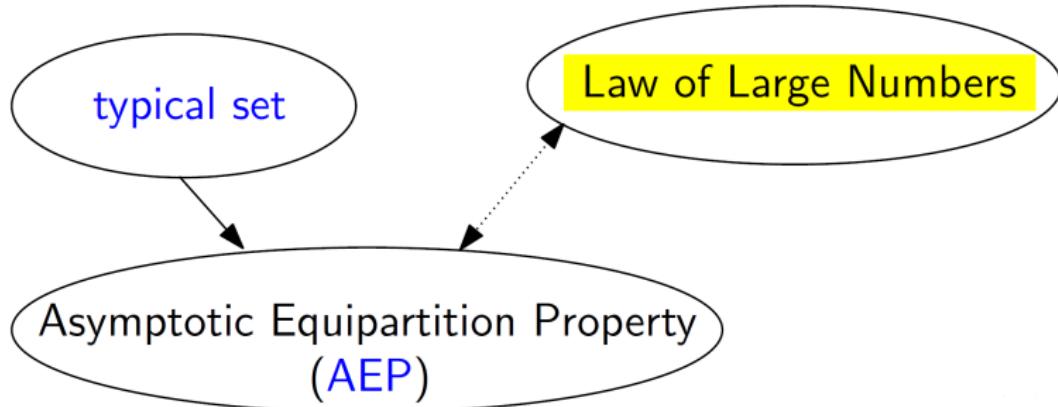
$$H(X) = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{16} \log 16 + \frac{4}{64} \log 64 = 2$$

The Huffman code is **optimal**!



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.



Channel Encoder

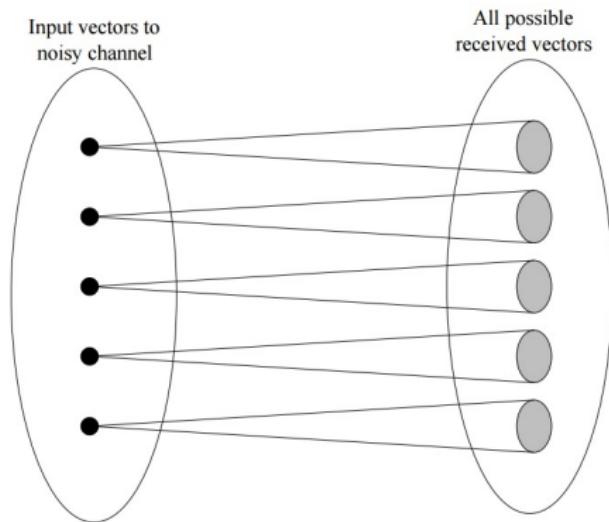
- Goal: To achieve an economical (**high rate**) and reliable (**low probability of error**) transmission of bits over a channel.

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.

Example:



Channel Encoder



Each transmitted codeword is corrupted by the channel. Each codeword **corresponds to a set of possible received vectors.**

Specify a set of codewords so that at the receiver it is possible to **distinguish which element was sent with high-probability.**

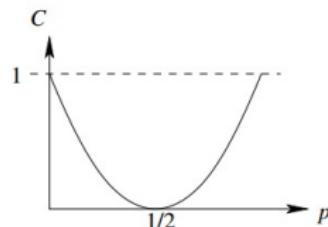
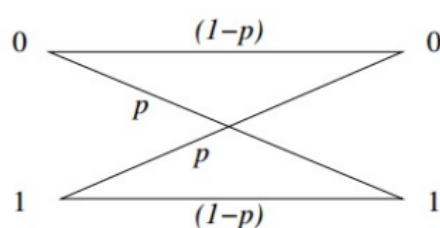
The *channel coding theorem* tells us the **maximum** number of such codewords we can define and still maintain completely distinguishable outputs.

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 ($\underbrace{2^{nR}}_{\text{\#codewords}}$, $\underbrace{n}_{\text{\#chan. uses}}$) codes such that $\Pr[\text{error}] \rightarrow 0$ as $n \rightarrow \infty$. Conversely, for any code, if $\Pr[\text{error}] \rightarrow 0$ as $n \rightarrow \infty$ 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?

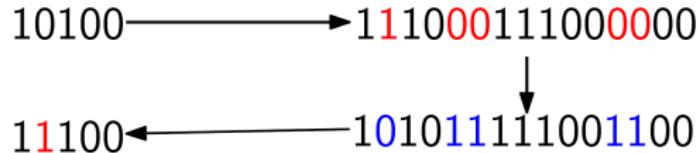
Repetition Code

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

$$\mathcal{C} = \{000\cdots 0, 111\cdots 1\}$$

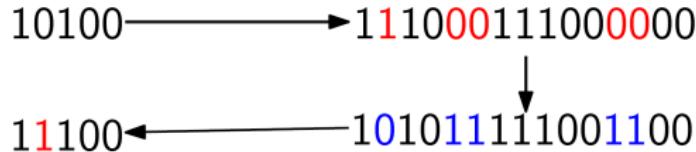
- Decoding by a *majority voting* scheme: if there are more 0's than 1's then declare 0, otherwise 1.
- Suppose that $R = 1/3$, i.e., the source output can be encoded before transmission by repeating each bit *three times*.

Example:



Repetition Code

Example:



The *bit error probability* \Pr_e is:

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

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:

$$Pr_e = 3p^2 - 2p^3.$$

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

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

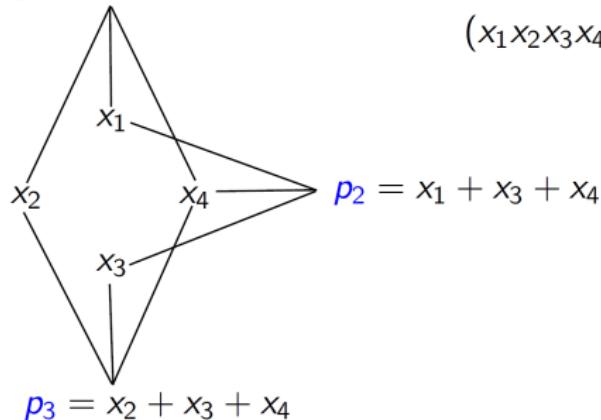
Thus, $Pr_e \rightarrow 0$ as $m \rightarrow 1$. However, $R \rightarrow 0$! Repetition code is NOT efficient! Shannon demonstrated that there exist codes which are *capacity achieving* at non-zero rates.



Hamming Code

$$p_1 = x_1 + x_2 + x_4$$

$$(x_1 x_2 x_3 x_4 p_1 p_2 p_3)$$



The $(7, 4)$ Hamming code can correct 1 bit error with Rate $R = 4/7$. This code is **much better** than repetition code.

Hamming codes can be computed in *linear algebra* through *matrices*. This will be explained later in this course.



Review of Probability Theory

• 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 $S = \zeta_1, \zeta_2, \dots$.

An *event* is a subset of S . Events consisting a **single** outcome are called *elementary events*.



Review of Probability Theory

• 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 $p_X : \mathcal{S}_X \rightarrow [0, 1]$ from \mathcal{S}_X to the closed unit interval $[0, 1]$ satisfying

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

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



Review of Probability Theory

• 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 $p_X : \mathcal{S}_X \rightarrow [0, 1]$ from \mathcal{S}_X to the closed unit interval $[0, 1]$ satisfying

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

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

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



Review of Probability Theory

- **Discrete Random Variables**

Example: A fair coin is tossed N times, and A is the event that an even number of heads occurs. What is $\Pr[A]$?



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Review of Probability Theory

• Discrete Random Variables

Example: A fair coin is tossed N times, and A is the event that an even number of heads occurs. What is $\Pr[A]$?

$$\begin{aligned}\Pr[A] &= \sum_{k=0, k \text{ even}}^N \Pr[\text{exactly } k \text{ heads occur}] \\ &= \sum_{k=0, k \text{ even}}^N \binom{N}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{N-k} \\ &= \frac{1}{2^N} \sum_{k=0, k \text{ even}}^N \binom{N}{k} \\ &= \frac{1}{2}.\end{aligned}$$



Review of Probability Theory

• Vector Random Variables

If the elements of \mathcal{S}_X are vectors of real numbers, then X is a *(real) vector random variable*.

Suppose Z is a vector random variable with a sample space in which each elements has *two* components (X, Y) , i.e.,
 $\mathcal{Z} = \{z_1, z_2, \dots\} = \{(x_1, y_1), (x_2, y_2), \dots\}$.

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

$$\mathcal{S}_X = \{x : \text{for some } y, (x, y) \in \mathcal{S}_Z\}.$$



Review of Probability Theory

• Vector Random Variables

If the elements of \mathcal{S}_X are vectors of real numbers, then X is a *(real) vector random variable*.

Suppose Z is a vector random variable with a sample space in which each elements has *two* components (X, Y) , i.e.,
 $\mathcal{Z} = \{z_1, z_2, \dots\} = \{(x_1, y_1), (x_2, y_2), \dots\}$.

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

$$\mathcal{S}_X = \{x : \text{for some } y, (x, y) \in \mathcal{S}_Z\}.$$

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



Review of Probability Theory

• 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.



Review of Probability Theory

• 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 $p_X(x)$ as

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

and similarly,

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



Conditional Probability

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

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

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



Conditional Probability

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

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



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

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

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

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



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 the second moment of $X - E[X]$, which can be computed as

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

