Fundamentals of Information Theory

Basic Concepts

Yayu Gao
School of Electronic Information and Communications
Huazhong University of Science and Technology
Email: yayugao@hust.edu.cn

Outline

- Model of communication systems
- How to characterize the information source?
- How much information a message contains?
- What is entropy?
- Joint and conditional entropy
- Relative entropy and mutual information
- Entropies in communications
- Chain Rules
- Jensen’s Inequality and Log Sum Inequality
- Entropy rate: from single-outcome to sequence-outcome
- What is a Markov source?
- Differential Entropy: from discrete to continuous

本节学习目标

1. Entropy rate 熵率
   - 写出定义与表达式
   - 说出物理意义
   - 计算马尔科夫信源熵率
2. Differential entropy 微分熵
   - 写出定义与表达式
   - 说出≥3条微分熵的性质
   - 写出均匀分布与正态分布的微分熵
   - 说出≥3条微分熵与熵之间的差异

重难点：

- 信源拓展：从单输出到序列+从离散到连续
- 概念拓展：熵率+微分熵
- 理解相关性与差异
- 计算：Markov source

11

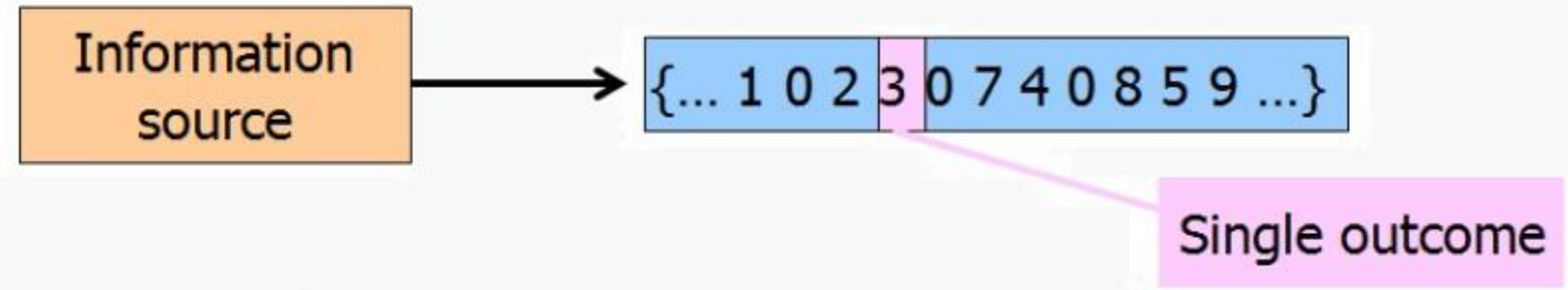
Entropy rate:

from single-outcome to sequence-outcome

Motivation Definition Theorem

Till now, we consider a discrete single-outcome source

- Outcome of the source:
  - **Single outcome**



- Model:

- Measure of information

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log [p(x)]$$

What if when things become complicated?

- We have a random output sequence $\{X_1, X_2, \ldots, X_n\}$

- If $\{X_i\}$ are *i.i.d.*, $H(X_1, X_2, \ldots, X_n) = nH(X)$.

- However, things are commonly related.

What if $\{X_i\}$ are not independent?

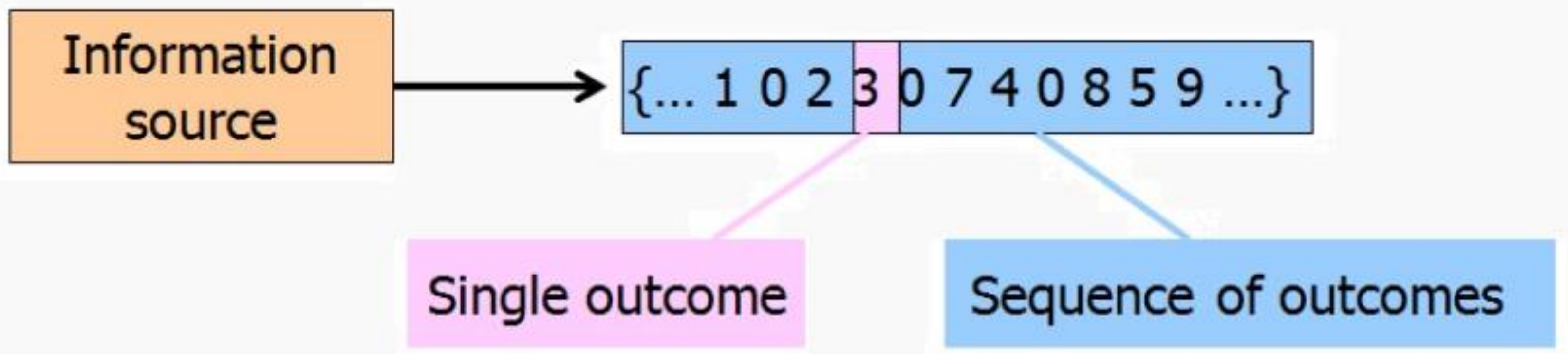
$$H(X_1, X_2, \ldots, X_n) \text{ vs } n \text{ ?}$$

Sources studied in our course

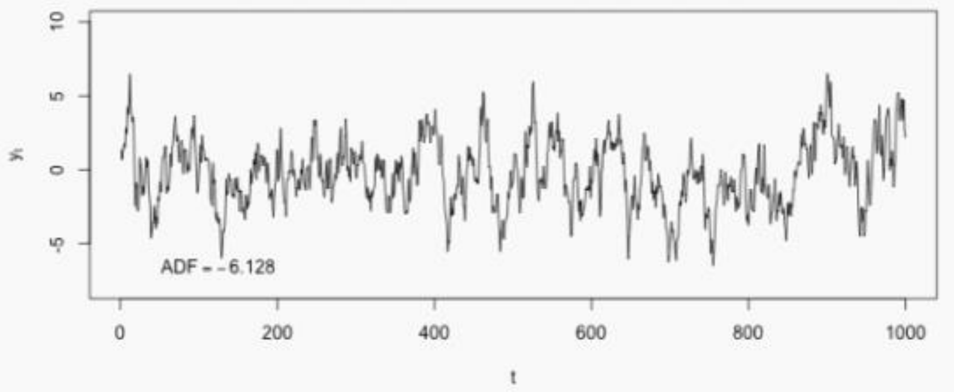
- We study the ideal sources with **good properties**, then use them to approximate real sources.
  - Discrete Source
    - Single Outcome Discrete Source
    - Outcome sequence Discrete Source
      - Discrete stationary memoryless source
      - Discrete stationary source with memory
  - Continuous source
    - Waveform source

Information source model

- Outcome of the source



- Model of discrete sequence source
  - Output is a discrete stochastic sequence
    - Sampled from continuous **stochastic process**



Sequence outcome source: system model

- Consider a sequence outcome source.

Information source $\rightarrow$ Discrete stochastic process $\{X_i\}$ at time n, $X_0, X_1, \ldots, X_{n-1}$

- Terms in this course
  - Sample space: $\mathcal{X}$
  - Random variable $(r.v.)$: X
  - Stochastic process: $X_i = X(t = i)$
  - Outcome of $\mathcal{X}$ or realization of X: x
  - Cardinality of set $\mathcal{X}$ (the number of elements): $|\mathcal{X}|$
- Joint probability mass function $(p.m.f.)$

$$Pr(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = p(x_0, x_1, \ldots, x_{n-1})$$

Revisit: What is a Stochastic Process?

- Tossing the Dice for once
  - It is a discrete **random variable**. (value 1-6)

- What if I toss the dice every hour in a day for 24 times?
  - The evolution in time is included.
  - It is now a **stochastic process**.
  - Discrete in both time and amplitude.

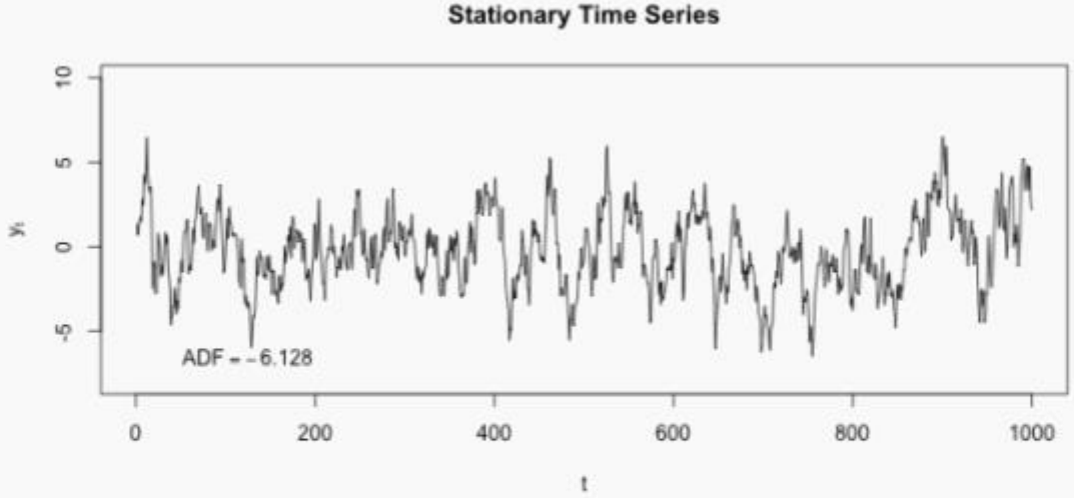
- A stochastic process is **a collection of random variables.**
- It describes **how a random event varies with time**.

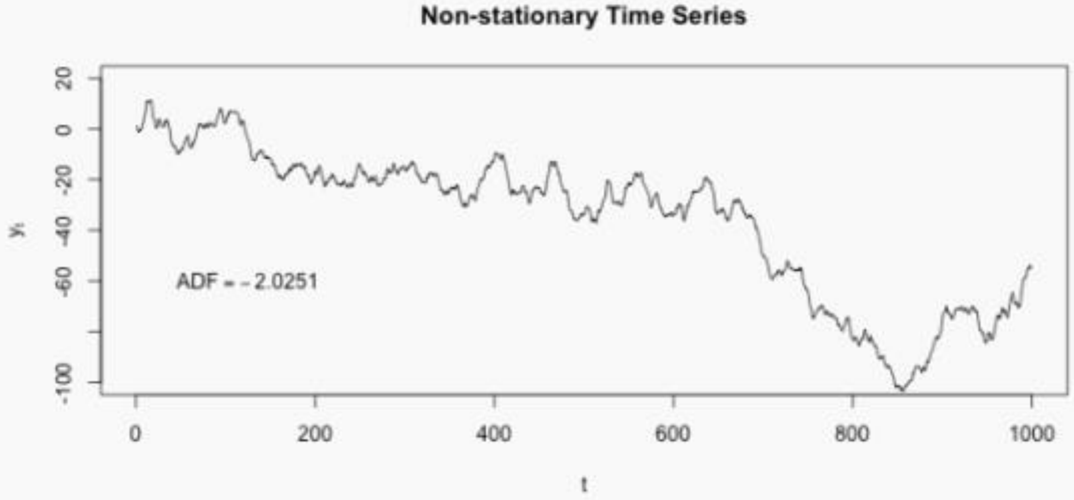
Revisit: Examples of Stochastic Processes

- Traffic on a highway during a day

- Stock price during a week

Revisit: Stationary processes





- A stochastic process $\{X(t), t \geq 0\}$ is said to be a stationary process if for all $n, s, t, \ldots, t_n$, the random vectors $X(t_1), \ldots X(t_n)$ and $X(t_1 + s), \ldots X(t_n + s)$ have the same joint distribution.
- In mathematics, a stationary process (or strict(ly) stationary process or strong(ly) stationary process) is a stochastic process whose joint probability distribution does not change when shifted in time or space.

How to measure the uncertainty of a stochastic process?

- Characterize uncertainty: entropy
  - Basic definition based on a random variable
  - Extension: entropy of a stochastic process

- Intuition
  - If a stochastic process is memoryless, its uncertainty should be the same as that of a random variable at given time *t*.
  - If a stochastic process has memory, the information carried by later messages is less than that carried by earlier messages.

- When the length of the sample sequence *n* approaches to infinity, **how does the entropy of the sequence grow with *n***?

Entropy rate: motivation

- In case of a stochastic process, the average entropy per symbol is defined as

$$H_n(\mathcal{X}) = \frac{H(X_1, X_2, \ldots, X_n)}{n}.$$

how does it grow with n?

$$\lim_{n\to\infty} H_n(\mathcal{X}) \to ?$$

Entropy rate: definition

- **Entropy rate**: The entropy rate of a stochastic processes $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} H_n(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}$$

when the limit exists.

per symbol entropy of the *n r.v.*'s

- **Conditional entropy rate**: We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$$

when the limit exists.

conditional entropy of the last *r.v.* given the past history

Entropy rate: example

#1 sequence of independent identical distributed (*i.i.d.*) random variables

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1)$$

#2 sequence of independent, but not identically distributed random variables

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} H(X_i)}{n} \to ?$$

For some distributions, $H(\mathcal{X})$ does not exist.

Entropy rate theorem

- **Theorem** For a stationary stochastic process, the limits of entropy rate and conditional entropy rate exist and are equal, i.e.

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

- Proof:
  - $H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$ has a limit $H'(\mathcal{X})$.
  - $H(\mathcal{X}) = H'(\mathcal{X})$.

Entropy rate theorem: proof

- **Theorem** For a stationary stochastic process, $H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$ is non-increasing in n and has a limit $H'(\mathcal{X})$.

- Proof:

$$H(X_{n+1}|X_n, X_{n-1}, \ldots, X_1) \leq H(X_{n+1}|X_n, X_{n-1}, \ldots, X_2) \qquad H(X|Y) \leq H(X)$$
$$= H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \qquad \text{Stationarity}$$

$H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$ is a decreasing sequence of nonnegative numbers.
It has a limit, $H'(\mathcal{X})$.

Entropy rate theorem: proof

- **Theorem** For a stationary stochastic process, the limits of entropy rate and conditional entropy rate exist and are equal, i.e.

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

Proof:

$$\begin{aligned}
H(\mathcal{X}) &= \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \qquad \text{Chain rule} \\
&= \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1) \quad \text{If } a_n \to a \text{ and } b_n = \frac{1}{n}\sum_{i=1}^{n} a_i, \text{ then } b_n \to a. \\
&= H'(\mathcal{X})
\end{aligned}$$

11

What is a Markov source?

What is a Markov process?

- Definition: A discrete stochastic process X_1, X_2, ... is said to be a Markov chain or a Markov process if

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1)$$
$$= \Pr(X_{n+1} = x_{n+1} | X_n = x_n)$$

Markov property: Given the present state, the future and past states are independent.

- The Markov process is time invariant if

$$\Pr(X_{n+1} = b | X_n = a) = \Pr(X_2 = b | X_1 = a)$$

- A time-invariant Markov chain can be characterized by its probability transition matrix $P = [P_{ij}]$

$$P_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$$

What is the entropy rate of a Markov process?

- The entropy rate of a typical case of stationary processes, Markov process, can be easily calculated.

$$
\begin{aligned}
H(\mathcal{X}) &= H'(\mathcal{X}) && \text{(Entropy rate theorem)} \\
&= \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \\
&= \lim_{n\to\infty} H(X_n|X_{n-1}) && \text{(Markovity)} \\
&= H(X_2|X_1) && \text{(Stationarity)}
\end{aligned}
$$

- If a stationary Markov distribution is μ_i and the transition matrix is P_{ij}, then

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$

m-th order Markov source

- A source is generating random outcomes: a_1, a_2, …, a_i, …
- The source has n possible outcomes.
- Let the **state** e_i be a sequence of m outcomes
- State space $E=\{e_1, e_2, \ldots, e_Q\}$, $Q=n^m$

- Example: Consider a binary source generating sequence ..01100011..
- Assume m=2.
- Then we have four possible states Q=4.
- e_1 =00, e_2 =01, e_3 =10, e_4=11

- What is a m-th order Markov source?

m-th order Markov source

If the output symbols and the state of source satisfying the following conditions, the source is called m-th order Markov source.

1. The outcome of source at this time point is only related to the current state of source

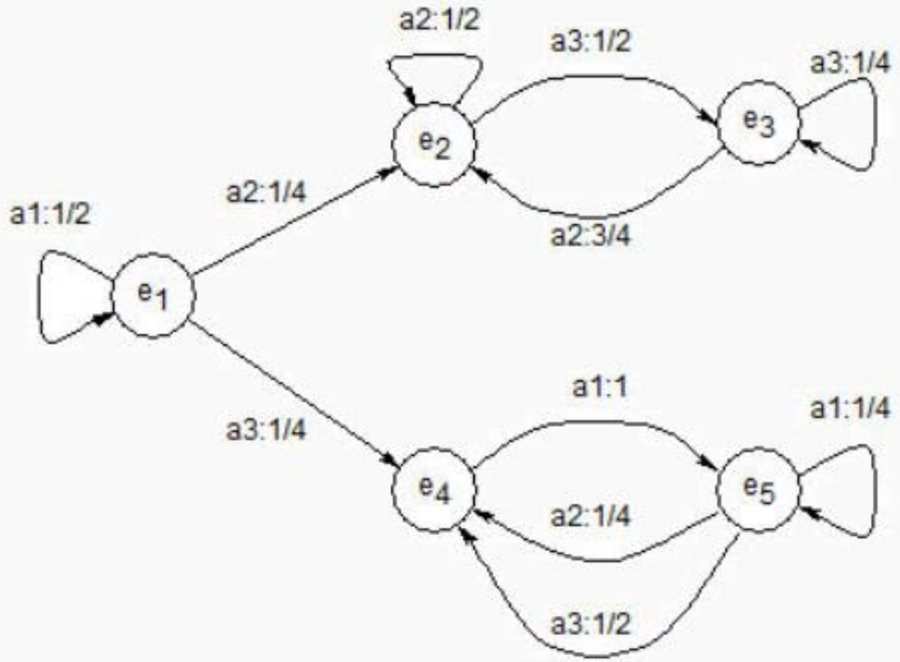
$$P\left(X_l = a_k | S_l = e_i, X_{l-1} = a_{k-1}, \ldots\right) = p_l\left(X = a_k | S = e_i\right)$$

2. The current state of source is only determined by the current outcome and the previous state

$$\boxed{P\left(S_l = e_j | X_{l-1} = a_{k-1}, S_{l-1} = e_i\right) = \begin{cases} 0, \\ 1. \end{cases}}$$

3. What is state e_i?
    - State represents a realization of the previous m output random variables.
    - e.g. $e_i = \{a_{k1}, a_{k2}, ..., a_{km}\}$

Markov source: example



$$A = \{a_1, a_2, a_3\}$$
$$E = \{e_1, e_2, e_3, e_4, e_5\}$$

$$\boxed{P(X_l = a_1 | S_l = e_1) = 1/2}$$
$$\boxed{P(X_l = a_2 | S_l = e_2) = 1/2}$$
$$\boxed{P(X_l = a_2 | S_l = e_1) = 1/4}$$

$$\boxed{P(S_l = e_2 | X_{l-1} = a_1, S_{l-1} = e_1) = 0}$$
$$\boxed{P(S_l = e_1 | X_{l-1} = a_1, S_{l-1} = e_1) = 1}$$
$$\boxed{P(S_l = e_4 | X_{l-1} = a_2, S_{l-1} = e_1) = 0}$$
$$\boxed{P(S_l = e_2 | X_{l-1} = a_2, S_{l-1} = e_1) = 1}$$

Entropy rate of Markov sources

Given the *m*-th order *n*-ary Markov source.

- m is the number of related previous outcomes.
- n is the number of elements in sample space.
- State space $S = \{e_i\}$, $i = 1, 2, \ldots, n^m$.
- Transition probability: $P_{ij} = p(e_j|e_i)$.
- Stationary probability: $\mu_j = \underset{l\to\infty}{p}(e_j)$.

Then, the entropy rate is

$$
\begin{aligned}
H(\mathcal{X}) = H'(\mathcal{X}) &= H(X_{m+1}|X_m, X_{m-1}, \ldots, X_1) \\
&= -\sum_{i=1}^{n^m}\sum_{j=1}^{n^m} p(e_i)p(e_j|e_i)\log p(e_j|e_i) \\
&= -\sum_{i,j} \mu_i P_{ij} \log P_{ij}.
\end{aligned}
$$

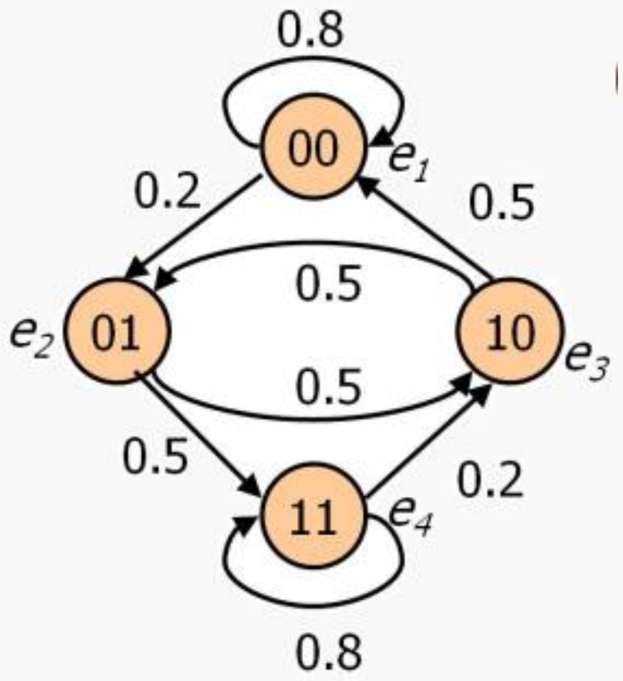
Markov source: example #1

A 2nd order and 2-ary Markov source.

$X = \{0, 1\}$.

Total $2^2 = 4$ states.

$S = \{e_1 = 00, e_2 = 01, e_3 = 10, e_4 = 11\}$.



Please calculate its entropy rate.

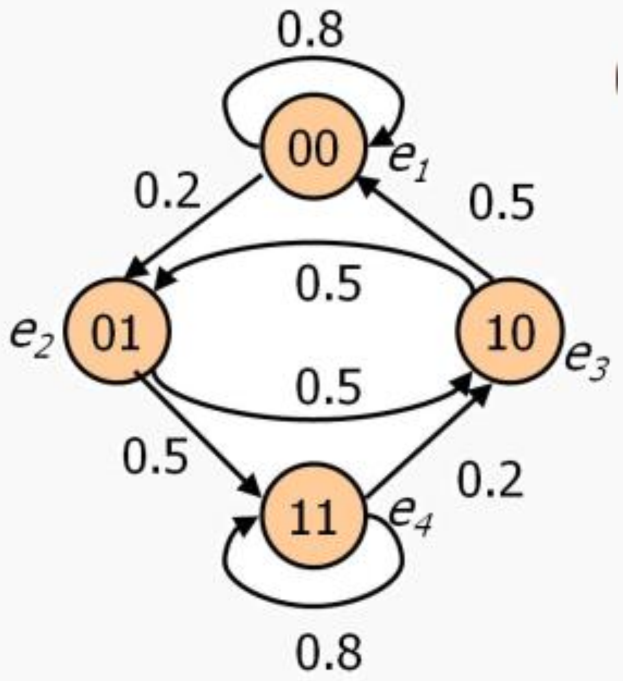
$$\boxed{H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.}$$

State probability

Transition probability

Markov source: solution #1

A 2nd order and 2-ary Markov source.
$X = \{0, 1\}$.
Total $2^2 = 4$ states.
$S = \{e_1 = 00, e_2 = 01, e_3 = 10, e_4 = 11\}$.



The transition probabilities are:

$$p(e_1|e_1) = 0.8, p(e_2|e_1) = 0.2,$$

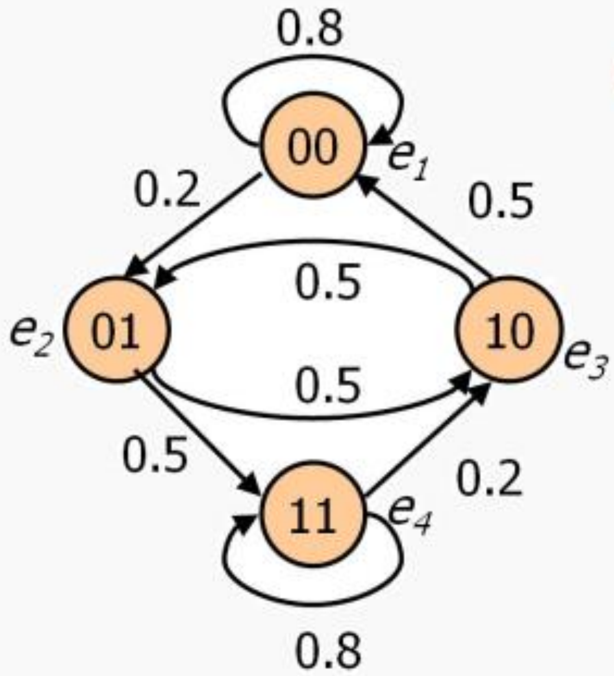
$$p(e_4|e_2) = 0.5, p(e_3|e_2) = 0.5,$$

$$p(e_1|e_3) = 0.5, p(e_2|e_3) = 0.5,$$

$$p(e_3|e_4) = 0.2, p(e_4|e_4) = 0.8.$$

Markov source: solution #1

The probability of each state:



$$\begin{cases} p(e_1) = 0.8p(e_1) + 0.5p(e_3), \\ p(e_2) = 0.2p(e_1) + 0.5p(e_3), \\ p(e_3) = 0.5p(e_2) + 0.2p(e_4), \\ p(e_4) = 0.5p(e_2) + 0.8p(e_4), \\ p(e_1) + p(e_2) + p(e_3) + p(e_4) = 1. \end{cases}$$

Thus,

$$p(e_1) = p(e_4) = \frac{5}{14}, p(e_2) = p(e_3) = \frac{2}{14}.$$

The entropy rate

$$H(\mathcal{X}) = -\sum_{i=1}^{4}\sum_{j=1}^{4} p(e_i)p(e_j|e_i)\log p(e_j|e_i) = 0.8 \text{ bits/symbol}.$$

Information Sources

Real discrete sources (most are un-stationary)

1. Assume stationary

   Stationary source, H_∞

严格来讲，大多是关联（记忆）长度为无穷大的多符号信源。

对实际信源，其所提供的信息量应该用 H_∞ 衡量。

但涉及到求解无穷维联合概率分布的问题。

将实际信源近似为 多符号信源 或 m 阶马尔可夫信源。

Information Sources

Real discrete sources (most are un-stationary)

1. Assume stationary

   Stationary source, H_∞

2. Assume limited memory

   m-th order Markov source, H_{m+1}

3. Assume no memory

   Stationary source without memory, $H_1 = H(X)$

4. Assume *i.i.d.*

   Extension of single outcome source, $H_0 = H(X)$

Information Sources: Markov sources

When we use Markov sources as an approximation， it is apparent that it is **more accurate for a larger m**. We then have

$$H_\infty \le H_{m+1} \triangleq H(X_{m+1}/X_1X_2\cdots X_m) \le \cdots \le H_{1+1} \triangleq \underbrace{H(X_2/X_1)}_{\text{1-order Markov source}}$$

$$\le \underbrace{H_{0+1} \triangleq H(X)}_{\text{0-order Markov source (no memory)}} \le \underbrace{H_0 = \log n}_{\text{Uniform-distributed source}}$$

Applications: Markov Models for Natural Language

Markov Models for Natural Language: Analysis 1

英语中包含26个英文字母，假设不区分大小写，并只有空格一个标点符号。

分析1：对英语信源，最粗略的近似可以如何进行处理？

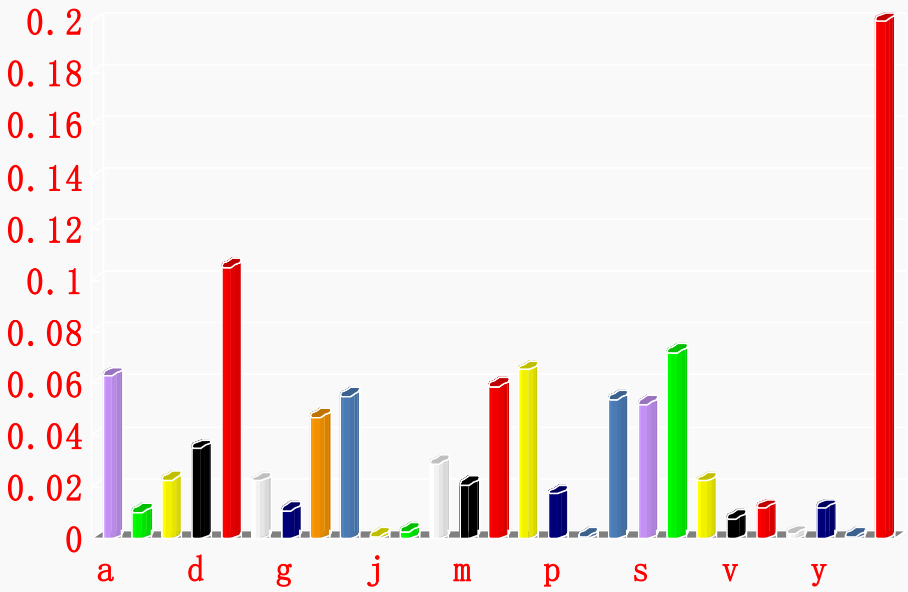
回答：假设认为前后符号间不相关，并且所有27个符号等概率分布。

$$H_0 = \log 27 \approx 4.76 \text{ 比特/符号}$$

为信源的最大熵

Markov Models for Natural Language: Question 1

实际英语信源，并非等概率分布



符号	概率	符号	概率	符号	概率
空格	0.2	S	0.052	Y, W	0.012
E	0.105	H	0.047	G	0.011
T	0.072	D	0.035	B	0.0105
O	0.0654	L	0.029	V	0.008
A	0.063	C	0.023	K	0.003
N	0.059	F, U	0.0225	X	0.002
I	0.055	M	0.021	J, Q	0.001
R	0.054	P	0.0175	Z	0.001

英文字母出现概率统计

Markov Models for Natural Language: Analysis 2

分析2：考虑英语符号概率分布，不考虑符号间依赖关系的情况下，平均符号熵等于多少?

$$H_{0+1} = -p(a)\cdot\log p(a) - p(b)\cdot\log p(b)$$
$$-\cdots - p(\text{空格})\cdot\log p(\text{空格})$$

$$\approx 4.03$$ **比特/符号**

问题：上述信源与实际情况近似到何种程度?

分析：按表的概率分布，随机选择英语字母得到一个信源输出序列为：

AI_NGAE_ITE_NNR_ASAEV_OTE_BAINTHA_HYROO_POER_SETRYGAIETRWCO_EHDUARU_EUEU_C_FT_NSREM_DIY_EESE_F_O_SRIS_R_UNNASHOR...

Markov Models for Natural Language: Analysis 3

分析3：考虑符号间依赖关系，可近似为马尔可夫信源。
1. 近似为一阶马尔可夫信源

前一个字母	后一个字母	条件概率
A	A	$P(A/A)$
	B	$P(B/A)$
	$\vdots$	$\vdots$
	空格	$P(\text{空格}/A)$
B	A	$P(A/B)$
	B	$P(B/B)$
	$\vdots$	$\vdots$
	空格	$P(\text{空格}/B)$

$$H_{1+1} = H(X_2/X_1)$$
$$= -\sum_{i=1}^{27}\sum_{j=1}^{27} p(x_i)\cdot p(x_j/x_i)\cdot \log p(x_j/x_i)$$
$$\approx 3.32 \text{ 比特/符号}$$

方法： 首字母可以任意选择。

首字母选定后，按条件概率选第二个字母。

第二个字母选定后，再按条件概率选第三个。

Markov Models for Natural Language: Analysis 3 (cont'd)

2. 类似地，近似为二阶马尔可夫信源。

$$H_{2+1} = H(X_3 / X_1 X_2) \approx 3.1 \text{ 比特/符号}$$

输出结果实例：

IANKS_CAN_OU_ANG_RLER_THTTED_OF_TO_SHOR_OF_TO_HAVEMEM_A_I_MAND_AND_BUT_WHISS_ITABLY_THERVEREER…

3. 类似地，可将英语信源近似为三阶、四阶…　。

$$\vdots$$

$$H_\infty \approx 1.4 \text{ 比特/符号}$$

依赖关系越多，及马尔科夫信源的阶数越高，输出的序列越接近实际情况。

Markov Models for Natural Language

$$H_\infty \approx 1.4 \le \cdots \le H_{2+1} \approx 3.1 \le H_{1+1} \approx 3.32 \le H_{0+1} \approx 4.03 \le H_0 \approx 4.76$$

上述结果，验证了随着阶数m的增加，符号相关性增加，熵值（平均每个符号所携带的信息量）会降低。

实际英语：

$$\underbrace{\text{Hello, My name is Lai. How are you}}_{\text{L个字符}}$$

问题：携带的信息量?

$$L \cdot H_\infty$$

Applications: Markov Models for Natural Language

- Question: **What is the entropy of natural language?**
- Shannon approximated the statistical structure of a piece of text using a simple mathematical model known as a Markov model.
- For example, with an input text a g g c g a g g g a g c g g c a g g g g . . .
  - Markov model with order 0: each letter is independently chosen.
  - However, there is a high correlation among successive letters in an English word or sentence. (as well as Chinese and other languages)

Can you establish a more refined statistical model for any given text using Markov chain?

Course Project!

Applications: Markov Models for Natural Language

Which language carries more information, Chinese or English?

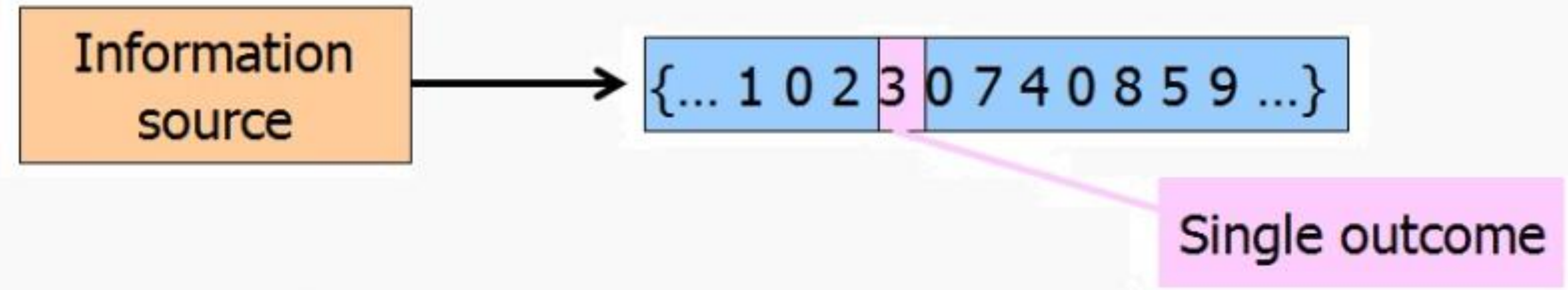
12

Differential Entropy: from discrete to continuous

1. Motivation
2. Definition
3. Properties
4. Examples
5. Maximum Entropy Theorems

So far, we consider discrete sources

- Outcome of the source:
  - **Discrete Single outcome**



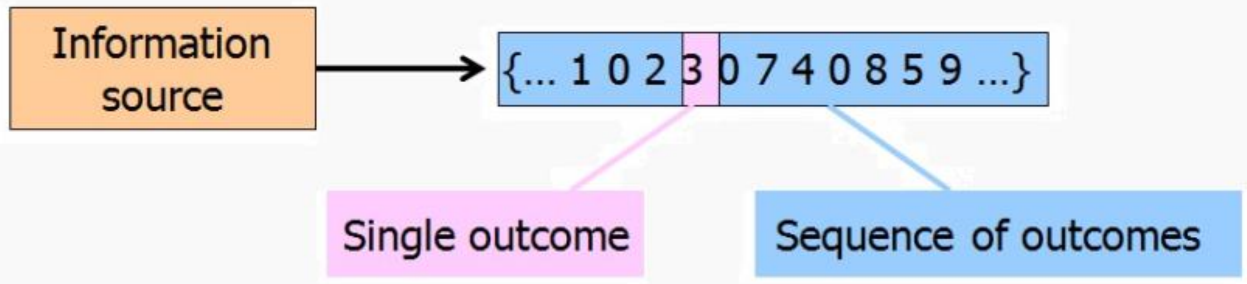
- Model:

- Measure of information: entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log [p(x)]$$

So far, we consider discrete sources

- Outcome of the source:
  - **Discrete sequence outcome**



- Model:

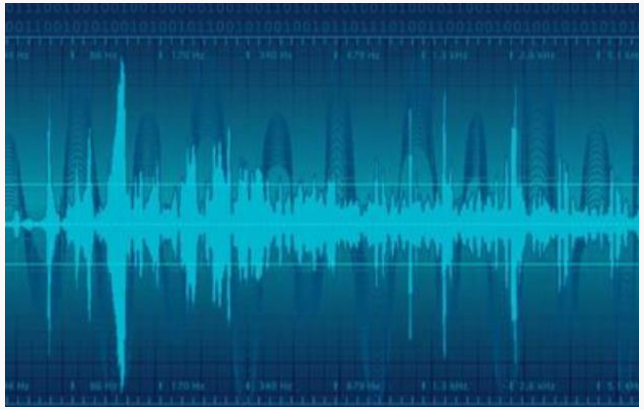
Discrete stochastic process $\{X_i\}$
at time n, $X_0, X_1, \ldots, X_{n-1}$

- Measure of information: entropy rate

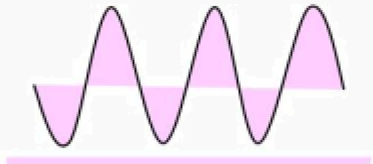
$$H(\mathcal{X}) = \lim_{n\to\infty} H_n(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}$$

How about continuous sources?

- In physical world, the output of sources are usually continuous.



Audio signal, Video signal…

Information source →  Continuous outcome

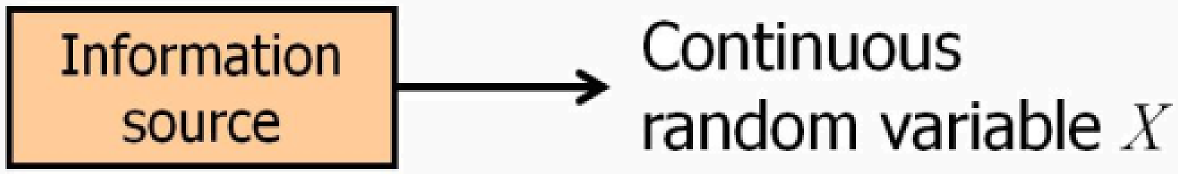
What is the information measure for continuous sources?

Sources studied in our course

- We study the ideal sources with **good properties**, then use them to approximate real sources.
  - Discrete Source
    - Single Outcome Discrete Source
    - Outcome sequence Discrete Source
      - Discrete stationary memoryless source
      - Discrete stationary source with memory
  - Continuous source
    - Waveform source

Continuous source: system model

- Consider a continuous source.



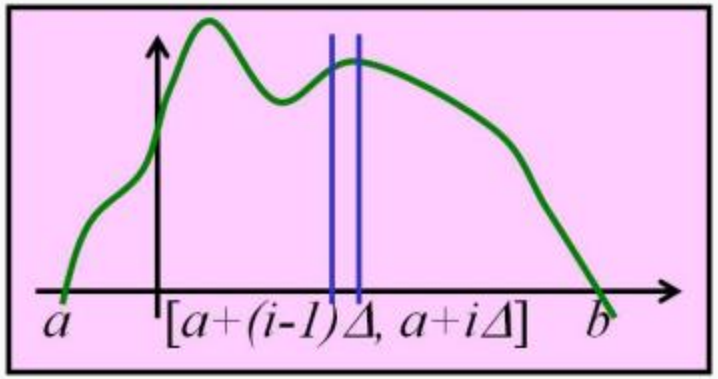
- Terms in this lecture
  - Sample space: $\mathcal{X}$
  - Random variable ($r.v.$): X
  - Outcome of $\mathcal{X}$ or realization of X : x
  - Cardinality of set $\mathcal{X}$ (the number of elements): $|\mathcal{X}|$
- Cumulative distribution function (c.d.f)
  - $F(x) = Pr(X \leq x)$
- Probability density function $(p.d.f)f(x)$

$$F(x) = \int_{-\infty}^{x} f(u)du \qquad f_x(x) = \frac{dF(x)}{dx}$$

How to measure the information of a continuous source?

- Generate a discrete source to simulate continuous source

$$\begin{bmatrix} R \\ f(x) \end{bmatrix}, \int_R f(x)dx = 1$$



Assume $p_i = P\left(a + (i-1)\Delta \le X \le a + i\Delta\right) = \int_{a+(i-1)\Delta}^{a+i\Delta} f(x)dx = \Delta f(x_i)$

$$\lim_{n\to\infty,\Delta\to 0} H(X) = -\lim_{n\to\infty,\Delta\to 0} \sum_{i=1}^{n} p_i \log p_i = -\lim_{n\to\infty,\Delta\to 0} \sum_{i=1}^{n} [\Delta f(x_i)] \log[\Delta f(x_i)]$$

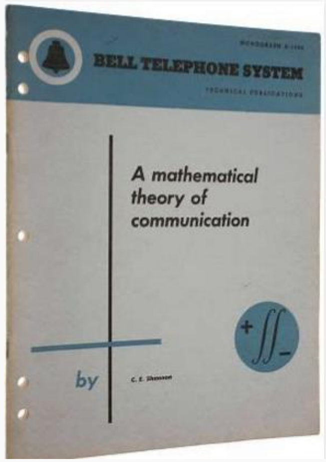
$$= -\lim_{n\to\infty,\Delta\to 0} \sum_{i=1}^{n} \Delta f(x_i) \log f(x_i) - \lim_{n\to\infty,\Delta\to 0} \left(\sum_{i=1}^{n} f(x_i)\, \Delta \log \Delta \right)$$

$$= -\int_a^b [f(x) \log f(x)]dx - \lim_{\Delta\to 0} \log \Delta$$

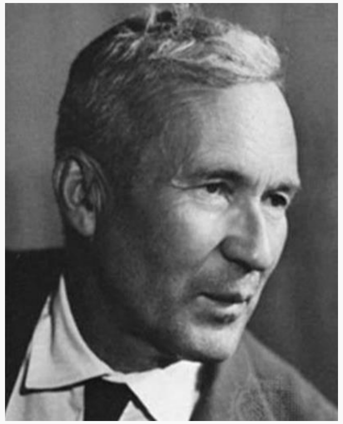
when $\Delta \to 0$, $H(X) \to \infty \Rightarrow$ the continuous entropy does not exist.

Differential entropy: a short history

- The concept of differential entropy was proposed first in his 1948 landmark paper by C. Shannon.



- The rigorous definition of differential entropy and mutual information of continuous variables were provided by Kolmogorov [2] and Pinsker [3].



A. Kolmogorov

M. S. Pinsker

[2] A Kolmogorov, “On the shannon theory of information transmission in the case of continuous signals," IRE Transactions on Information Theory, vol. 2, no. 4, pp. 102-108, Dec. 1956.
[3] M. S. Pinsker, “Information and stability of random variables and processes," Izd. Akad. Nauk, 1960, translated by A. Feinstein in 1964.

Differential entropy: definition

- A continuous random variable contains **infinite information**.

$$\lim_{n\to\infty,\Delta\to 0} H(X) = -\int_a^b [f(x)\log f(x)]dx - \lim_{\Delta\to 0}\log\Delta$$

- Define **differential entropy** as the information measure of a continuous random variable.

$$\boxed{h(X) = h(f) = -\int_S f(x)\log f(x)dx}$$

  - S is the support set of the $r.v.X$
  - $f(x)$ is the $p.d.f.$ of X
  - Since $h(X)$ only depends on the $p.d.f.$, it can also be marked as $h(X) = h(f)$

Differential entropy: remarks

- A continuous random variable contains **infinite information**.

$$\lim_{n\to\infty,\Delta\to 0} H(X) = -\int_a^b [f(x)\log f(x)]dx - \lim_{\Delta\to 0}\log\Delta$$

- Define **differential entropy** as the information measure of a continuous random variable.

$$\boxed{h(X) = h(f) = -\int_S f(x)\log f(x)dx}$$

- It is not the absolute entropy of a continuous source.
- It cannot represent the average uncertainty/information of the source.
- It is a relative value with the reference point $-\lim_{\Delta\to 0}\log\Delta$
- It represents the difference between former and later source entropy

Joint/conditional differential entropy

- Joint differential entropy

$$h(X_1, X_2, \ldots, X_n) = -\int_S f(x_1, x_2, \ldots, x_n) \log f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

- Conditional differential entropy

$$h(X|Y) = -\int f(x, y) \log f(x|y) dx dy$$
$$h(X, Y) = h(X) + h(Y|X)$$
$$h(X, Y) \leq h(X) + h(Y)$$

Relative entropy and mutual information

- Relative entropy

$$\boxed{D(f||g) = \int f \log\left(\frac{f}{g}\right)}$$

- Mutual information

$$\boxed{I(X;Y) = \int f(x,y) \log\left[\frac{f(x,y)}{f(x)f(y)}\right] dxdy}$$

Relative entropy and mutual information: relationship

$$\boxed{I(X;Y) = h(Y) - h(Y|X)}$$

Proof:

$$\begin{aligned}
I(X;Y) &= D(f(x,y)||f(x)f(y)) \\
&= \int\int f(x,y)\log\left[\frac{f(x,y)}{f(x)f(y)}\right] dxdy \\
&= \int\int f(x,y)\log\left[\frac{f(x,y)}{f(x)f(y)}\right] dxdy \\
&= h(X) - h(X|Y) \\
&= h(Y) - h(Y|X)
\end{aligned}$$

Differential entropy: properties

Non-negativity of relative entropy and its corollary

- $D(f||g)$

$$\boxed{\begin{aligned} &D(f||g) \geq 0 \\ &D(f||g) = 0 \iff f(x) = g(x) \text{ almost everywhere} \end{aligned}}$$

- $I(X;Y)$

$$\boxed{\begin{aligned} &I(X;Y) \geq 0 \\ &I(X;Y) = 0 \iff f(x,y) = f(x)f(y) \end{aligned}}$$

- $h(X|Y)$

$$\boxed{\begin{aligned} &h(X|Y) \leq h(X) \\ &h(X|Y) = h(X) \iff f(x,y) = f(x)f(y) \end{aligned}}$$

Differential entropy: properties

- Chain rule for differential entropy

$$\boxed{h(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} h(X_i | X_{i-1}, X_{i-2}, ..., X_1)}$$

- Independent bound

$$\boxed{h(X_1, X_2, ..., | X_n) \leq \sum_{i=1}^{n} h(X_i)}$$

- Translations and rotations

$$\boxed{\begin{aligned} &h(X + C) = h(X) \\ &h(aX) = h(X) + \log(|a|) \\ &h(\mathbf{AX}) = h(\mathbf{X}) + \log(|\mathbf{A}|) \end{aligned}}$$

- For discrete source, $H(aX)$ v.s. $H(X)$?

Differential entropy: properties

- Proof:

  Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

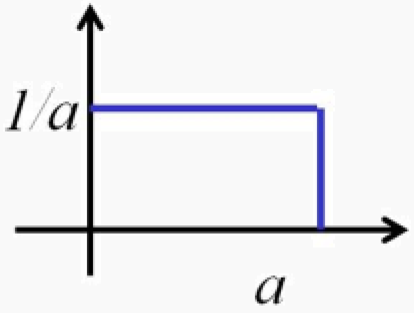
$$\begin{aligned} h(aX) &= -\int f_Y(y) \log f_Y(y) dy \\ &= -\int \frac{1}{|a|} f_X(\frac{y}{a}) \log \left(\frac{1}{|a|} f_X(\frac{y}{a}) \right) dy \\ &= -\int f_X(x) \log (f_X(x)) \, dx + \log |a| \\ &= h(X) + \log |a| \,. \end{aligned}$$

- Note that when $a > 1$, $\log |a| > 0$. This implies $h(aX) > h(X)$.
- The operation of aX physically extend X axis. The shape of the probability density function $f_X(x)$ actually is widened and lowered by $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$. Hence, the uncertainty of $H(aX)$ increases compared with $H(X)$.

Examples of continuous source

- Example #1: Uniform distribution

$$f(x) = \begin{cases} \frac{1}{a}, & 0 \leq x \leq a \\ 0, & \text{otherwise} \end{cases}$$



- Differential entropy

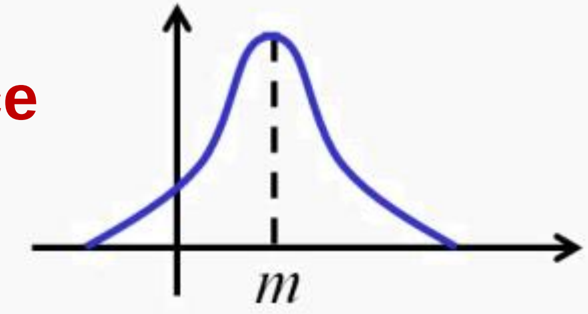
$$h(X) = -\int_S f(x) \log f(x) dx = -\int_0^a \frac{1}{a} \log\left(\frac{1}{a}\right) dx = \log(a)$$

- Comments
  - If $a < 1$, $\log(a) < 0$, thus differential entropy can be negative.

Examples of continuous source

- Example #2: normal distribution

Gaussian source



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

- Differential entropy

$$\begin{aligned}
h(X) &= -\int f(x) \ln f(x) dx \\
&= -\int f(x) \left[-\frac{(x-m)^2}{2\sigma^2} - \ln\left(\sqrt{2\pi\sigma^2}\right) \right] dx \\
&= \frac{1}{2\sigma^2} \int f(x)(x-m)^2 dx + \frac{1}{2} \ln\left(2\pi\sigma^2\right) \\
&= \frac{1}{2} \ln(e) + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} \ln\left(2\pi e\sigma^2\right) nats
\end{aligned}$$

Maximum entropy theorems for continuous source

- Uniform p.d.f differential entropy bound
  Uniform distribution maximizes the differential entropy over all distributions with the same range $[a, b]$.

$$\boxed{h(X) \leq \log \prod_{i=1} N(b_i - a_i)}$$

- Gaussian p.d.f differential entropy bound
  Multivariate normal distribution maximizes the differential entropy over all distributions with the same covariance.

$$\boxed{h(X_1, X_2, ..., X_n) \leq \frac{1}{2} \log(2\pi e)^n |K| bits,}$$

  where $|K|$ is the determinant of the covariance matrix K.

  $\Rightarrow$ Given continuous $r.v.$ X with mean m and variance σ^2, the differential entropy is maximized when it follows Gaussian distribution.

本节学习目标

1. 熵率
   - ➢写出定义与表达式
   - ➢说出物理意义
   - ➢计算马尔科夫信源熵率
2. 微分熵
   - ➢写出定义与表达式
   - ➢说出≥3条微分熵的性质
   - ➢写出均匀分布与正态分布的微分熵
   - ➢说出≥3条微分熵与熵之间的差异

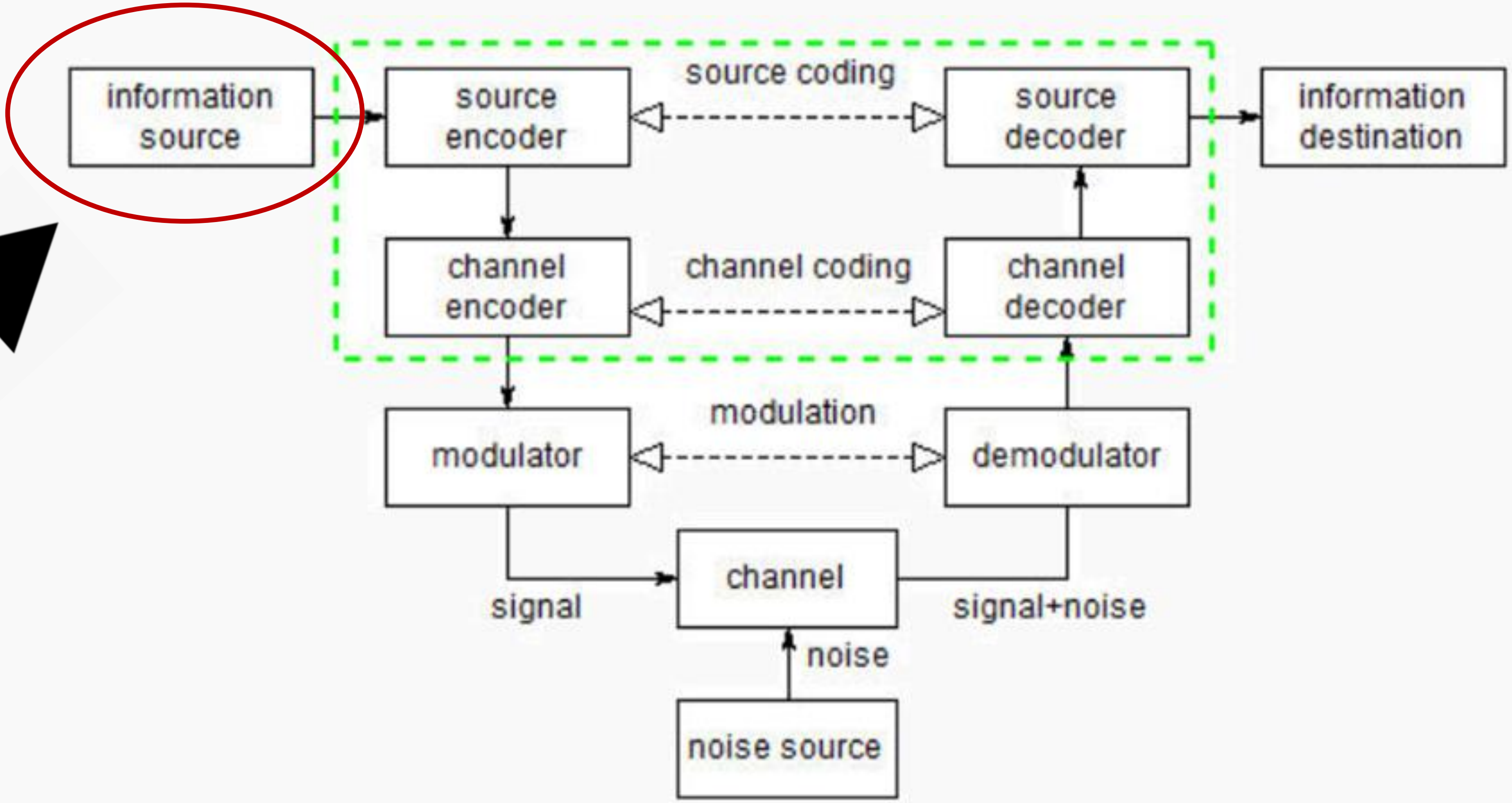
重难点：

- ➢ 信源拓展：从单输出到序列+从离散到连续
- ➢ 概念拓展：熵率+微分熵
- ➢ 理解相关性与差异
- ➢ 计算：Markov source

Summary of Chapter 2

Summary: Focus on the Object

- Model of Communication Systems



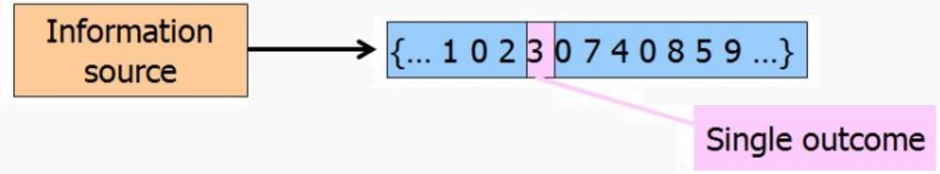
Summary: Ask the Key Question

How much information is transmitted?

How much information is lost?

Fundamental Question:
How to quantify information?

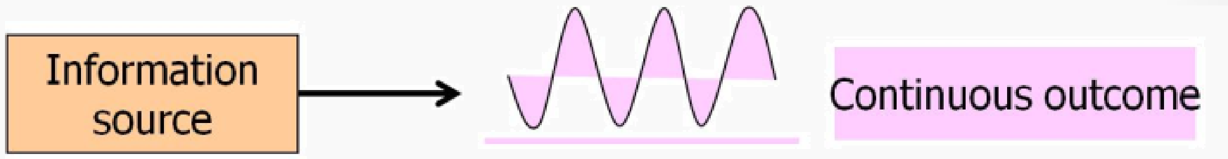
Summary: Gone with the Logic



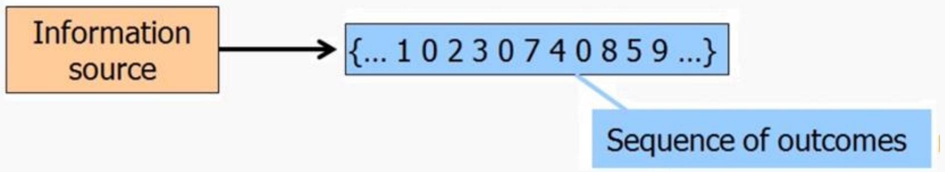
Discrete single outcome: Entropy

- ✓ Joint entropy
- ✓ Conditional entropy
- ✓ Relative entropy
- ✓ Mutual information

Discrete to Continuous: Differential entropy



Single to Sequence: Entropy rate



Start with simple examples

Extend to complex cases

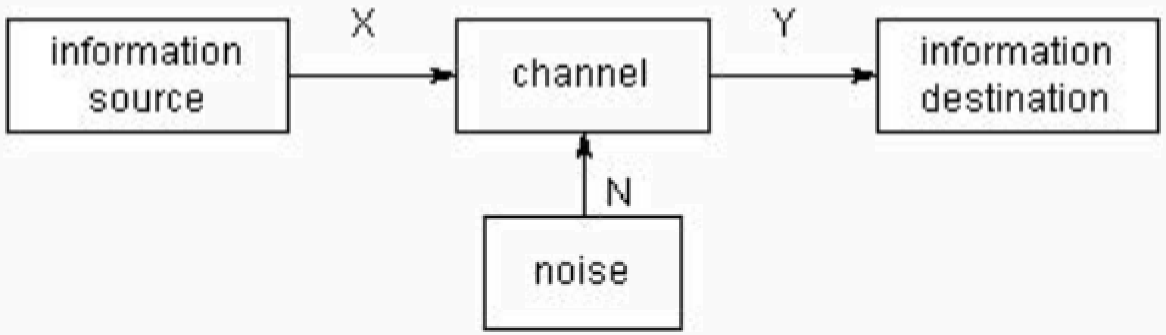
Summary: Concepts

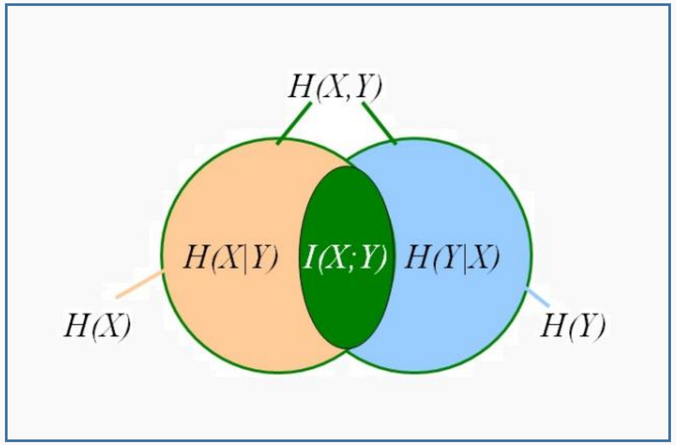
- *Self-Information $I(x)$*
  - Measure of uncertainty of single outcome
  - Non-negative
- *Entropy $H(X)$*
  - $H(X) = E_X[I(x)]$
  - Measure of uncertainty of information source
  - Non-negative
- *Relative entropy $D(p(x)||q(x))$*
  - Measure of similarity of distributions
  - Non-negative
- *Mutual information $I(X;Y)$*
  - $I(X;Y) = D[p(x,y)||p(x)p(y)] = E_{X,Y}[I(x;y)]$
  - Measure of similarity between joint and product $p.m.f.$'s
  - Special case of relative entropy (Non-negative)

Summary: Properties of Entropies

Non-negativity	$H(X) \geq 0$
Chain Rule	$H(X, Y) = H(X) + H(Y\|X)$
Uniform p.m.f. maximization	$H(X) \leq \log(\|X\|)$
Conditional reduction	$H(X\|Y) \leq H(X)$
Independence bound	$H(X_1, X_2, \ldots, X_n) \leq \sum_i H(X_i)$

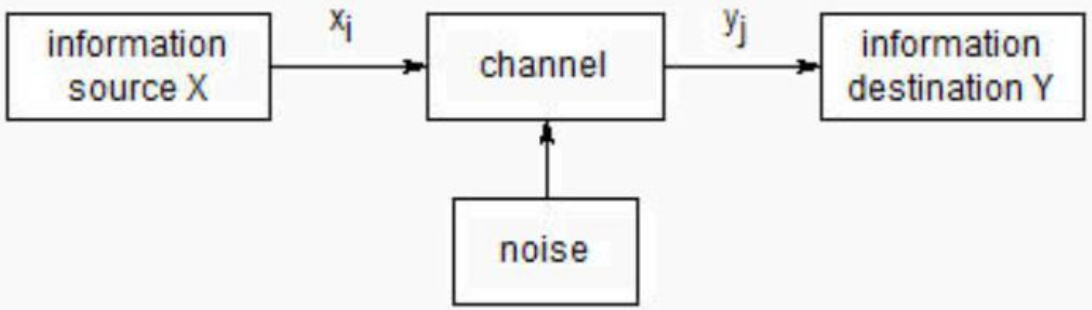
Summary: Physical meaning





- ✓ H(X|Y): loss entropy
- ✓ H(Y|X): noise entropy
- ✓ I(X;Y): information transmitted from source to destination

Summary: Physical meaning



- Mutual information of **realization at the micro-level**
  - $I(x_i; y_j) = \log\left[\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right] = \log\left[\frac{p(x_i|y_j)}{p(x_i)}\right] = \log\left[\frac{1}{p(x_i)}\right] - \log\left[\frac{1}{p(x_i|y_j)}\right]$
  - At destination: $I(x_i; y_j) = I(x_i) - I(x_i|y_j)$
  - At source: $I(y_j; x_i) = I(y_j) - I(y_j|x_i)$
  - From system: $I(x_i; y_j) = I(x_i) + I(y_j) - I(x_i, y_j)$

- Mutual information at the macro-level

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left[\frac{p(x,y)}{p(x)p(y)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y|x)\log\left[\frac{p(y|x)}{p(y)}\right]$$

My Homepage

Thank you!

Yayu Gao

School of Electronic Information and Communications
Huazhong University of Science and Technology
Email: yayugao@hust.edu.cn