

- 0304 Entropy (1)
 - Entropy: Definition
 - Notation:
 - Examples:
 - Properties
 - More Entropies
 - Joint Entropy
 - Conditional Entropy
 - Chain Rule
 - Venn Diagram
 - Zero Entropy
- 0309 Entropy (2)
 - Relative Entropy
 - Relative Entropy is NOT Metric
 - Conditional Relative Entropy
 - Mutual Information
 - Mutual Information and Entropy
 - Conditional Mutual Information
 - Propositions About Information Quantities
 - Nonnegative Mutual Information
 - More Properties
- 0311 Entropy (3)
 - Independence Bound on Entropy
 - Markov Chain
 - Data Processing Inequality
 - $I(X;Y;Z)$
 - Information Diagram
 - 2RVs to more
 - Markov Chain
 - Examples: Use Info Diagram to Prove Inequalities
 - Practical Examples
 - Causality(因果推断)
 - Perfect Secrecy(完美安全模型)
 - Fano's Inequality: Estimation
 - Background
 - Proof
 - Convexity/ Concavity of Information Measures
- 0316 AEP
 - Law of Large Numbers
 - AEP (Asymptotic Equipartition Property)
 - Typical Set

- High Probability Set
- Data Compression
 - Problem Formulation
 - Procedure
 - Analysis
- 0318 Entropy Rate
 - Stochastic Process
 - Introduction
 - Stationary Process
 - Markov Chain
 - Stationary Distribution of MC
 - Entropy Rate
 - $H'(X)$
 - Cesaro Mean
 - Entropy Rate for Stationary Process
 - Entropy Rate for Markov Chain
 - Example: Random Walk
 - Second Law of Thermodynamics
 - Extension: Functions of Markov Chains
- 0323 Data Compression (1)
 - Example of Codes
 - Nonsingular Code
 - Prefix Code
 - Kraft Inequality
 - Extended Kraft Inequality
 - Optimal Codes
 - Problem Formulation
 - Solution
 - Bounds
 - Approach the limit
 - Wrong Code
 - Kraft Inequality For Uniquely Decodable Codes
 - Summary
- 0325 Data Compression (2)
 - Huffman coding
 - Algorithm
 - Extension
 - Canonical Codes
 - Optimality: Strategy
 - Shannon-Fano-Elias coding
 - Formulation
 - Example

- Optimality
- 0330 Data Compression (3)
 - Random Variable Generation
 - Introduction
 - Formulation
 - Properties
 - Algorithm
 - Example
 - Universal Source Coding
 - Minmax Redundancy
 - Redundancy and Capacity
 - Arithmetic Coding
 - Lempel-Ziv Coding: Introduction
 - Sliding Window
 - Tree-Structure
- 0401 Channel Capacity (1)
 - Noise in Information Transmission
 - Discrete Memoryless Channel
 - Channel Capacity
 - Properties Of Channel Capacity
 - Examples
 - Noiseless Binary Channel
 - Noisy Channel with Nonoverlapping Outputs
 - Noisy Typewriter
 - Example: Binary Symmetric Channel
 - Example: Binary Erasure Channel
 - Symmetric Channel
 - Computation Of Channel Capacity
- 0408 Channel Capacity (2)
 - Recall: Channel Model for Telegraph
 - Memory and Feedback
 - Definition
 - Analysis
 - Interpretation
 - Channel Model
 - Code
 - Probability of Error
 - Rate and Capacity
 - Joint Typicality
 - Intuition for Channel Capacity
- 0413 Channel Capacity (3)
 - Coverse Proof Special Case: Zero-Error Codes

- Coverse Proof: Channel Coding Theorem
- Achievability
 - Code Construction
 - Joint Decoding
- $\Pr(\mathcal{E}) \rightarrow 0$
 - $\Pr(\mathcal{E}) \rightarrow 0 \Rightarrow \lambda^{(n)} \rightarrow 0$
- (Introduction) Feedback Capacity
 - TODO:看一下证明
- (Introduction) Source-Channel Separation
 - Formal Problem
 - TODO: Theorem
- Error Correction Code
- Hamming Code
- 0415 Differential Entropy (1)
 - Differential Entropy
 - Definition
 - Example
 - $h(X)$: infinite information
 - $h(aX)$: Stretching Random Variable
 - Differential and Discrete Entropy
 - AEP For Continuous Random Variable
 - Joint and Conditional Differential Entropy
 - Entropy of Multivariate Normal Distribution
 - Covariance Matrix
 - Multivariate Normal Distribution
 - Entropy
 - Relative Entropy and Mutual Information
 - Mutual Information: Master Definition
- 0420 Differential Entropy (2)
 - Correlated Gaussian
 - Maximum Entropy with Constraints
 - Maximum Entropy
 - Hadamard's Inequality
 - Balanced Information Inequality
 - Han's Inequality
 - Information Heat
 - Heat Equation
 - Entropy and Fisher Information
 - Higher Order Derivatives of $h(Yt)$
 - EPI and FII
- 0422 Gaussian Channel
 - Gaussian Channel

- Energy Constraint
- Intuition
- Theorems
 - Definition
 - Code Construction
 - Generation of the codebook
 - Encoding
 - Decoding:
 - Probability of Error
 - Converse
- Parallel Gaussian Channel
 - Problem
 - Solution
- Worst Additive Noise
 - Problem
 - Entropy power inequality

0304 Entropy (1)

Entropy: Definition

Notation:

- Let X be a *discrete random variable* (离散随机变量) with alphabet (字母表/样本空间) \mathcal{X} and **probability mass function** $p(x) = Pr(X = x), x \in \mathcal{X}$. (abbrev. $p_{\mathcal{X}}(x)$)
- The entropy of X is defined by

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

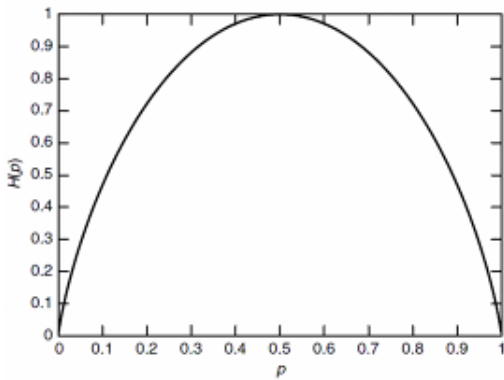
- Remark:
 - $0 \log 0 \rightarrow 0$
 - 我们有时也直接用概率分布函数表示字母表, 因为 $H(X)$ 仅与 $p(x)$ 有关
 - $H(X) \geq 0$
 - 当 X 均匀分布时, $H(X) = \log |\mathcal{X}|$, 在离散情况下熵是最大的
 - $H_b(X) = \log_b a H_a(X)$
 - 对数底取 e 时, entropy is measured in nats
 - 对数取 2 时, entropy is measured in bits
 - 本课程中, 主要讨论有限字母表的问题

Examples:

- Binary entropy function

$$\text{Let } X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

$$H(X) = -p \log p - (1 - p) \log(1 - p)$$



- 用期望的形式表达熵, 熵是随机变量 $\log \frac{1}{p(X)}$ 的期望

$$E_p g(X) = \sum_{x \in \mathcal{X}} g(x) p(x)$$

$$H(X) = E_p \log \frac{1}{p(X)}$$

Properties

Theorem For any discrete random variable X $0 \leq H(X) \leq \log |\mathcal{X}|$

Pf. 非负, trivial

Note. $f(x) = -x \log x$ is concave in x (by second-derivative). and that $\sum_X p(x) = 1$.

By applying the concavity of $f(x)$,

$$\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} -p(x) \log p(x) \leq -\frac{1}{|\mathcal{X}|} \log \frac{\sum_x p(x)}{|\mathcal{X}|} = \frac{1}{|\mathcal{X}|} \log |\mathcal{X}|$$

Recall: Concavity.

$$\sum_i p_i f(x_i) \leq f\left(\sum_i p_i x_i\right)$$

Lemma: 均匀分布最大化离散熵 equality holds iff $p(x) = 1/|\mathcal{X}|$

More Entropies

- 我们会定义更多熵, 条件熵/联合熵...
- 熵的定义只与概率密度有关, 和字母表的取值具体情况无关.
- 对多个随机变量, 我们可以定义
 - 联合分布 $p(x_i, x_j)$
 - 条件分布 $p(x_i | \dots)$
 - 都可以计算出熵
- 概率论中基本定律
 - Chain Rule $p(x_1, x_2, \dots, x_n) = p(x_n) p(x_{n-1} | x_n) \dots p(x_1 | x_2, \dots, x_{n-1})$
 - Bayesian Rule $p(y)p(x|y) = p(x)p(y|x)$
 - 这些基本准则的存在, 表明熵也可能存在特殊的结构

Joint Entropy

Facts: 多个随机变量的字母表可以组合成一个字母表

Definition The joint entropy $H(X, Y)$ of a pair of discrete random variable (X, Y) with joint distribution $p(x, y)$ is defined as

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

Propositions:

- $H(X, X) = H(X)$, 可以理解成本质上是同一件事, 只是多次实验而已
- $H(X, Y) = H(Y, X)$
- 联合熵也可以写成联合期望的形式

$$H(X_1, X_2, \dots, X_n) = - \sum p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n) = -E \log p(X_1, \dots, X_n)$$

Conditional Entropy

两种计算方式

- 先对fixed X 算条件熵, 再对所有条件熵加权求和.
 - Entropy for $p(Y|X = x)$

$$H(Y|X = x) = \sum_y -p(y|X = x) \log p(y|X = x) = -E \log p(Y|X = x)$$

- Entropy for $p(Y|X)$

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

- 也可以通过直接根据以下推导, 直接计算 $\log p(Y|X)$ 的期望.

$$\begin{aligned}
 H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) \\
 &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\
 &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\
 &= -E \log p(Y|X)
 \end{aligned}$$

Proposition $H(Y|X) \leq H(Y)$

直观理解, 条件熵(在X已知的情况下,Y的不确定度)会比原始系统的熵要低, 条件降低了系统的不确定度.

Remark Example中两个有意思的结论, 后续给出证明

$$\begin{aligned}
 &H(X|Y) \neq H(Y|X) \\
 &H(X|Y) + H(Y) = H(Y|X) + H(X) = H(X, Y)
 \end{aligned}$$

直观理解, 两件事的先后发生的不确定性不具有对称性, 两件事的不确定性之和可以理解为两件事(带条件)先后发生的不确定性之和.

Chain Rule

Recall: 概率论中, $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$. 因此我们有 $\log p(x, y) = \log p(x|y) + \log p(y) = \log p(y|x) + \log p(x)$.

考虑上节中定义的条件熵,

$$\begin{aligned}
 &E - \log p(x, y) \\
 &= E - \log p(x|y) + E - \log p(y) \\
 &= E - \log p(y|x) + E - \log p(x)
 \end{aligned}$$

Theorem: Chain Rule $H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$

Proposition

1. 如果X和Y独立, 那么 $H(X, Y) = H(X) + H(Y)$.
2. 如果X是关于Y的函数, 那么 $H(X, Y) = H(Y)$.
3. 贝叶斯公式: $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$

Pf. Note

$$p(x, z)p(y|x, z) = p(x, y, z) = p(z)p(x, y|z)$$

and that

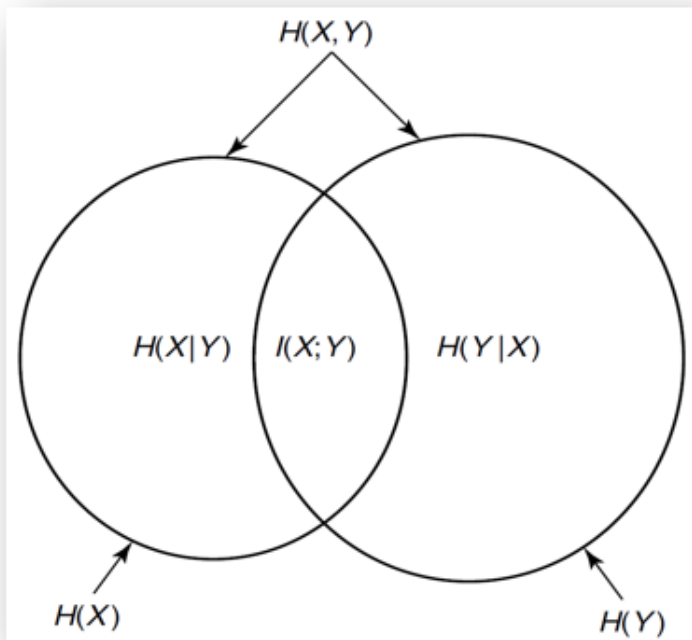
$$p(x, z) = p(x|z)p(z)$$

it follows that $p(x, y|z) = p(x|z)p(y|x, z)$.

Venn Diagram

我们如何高效地整理信息量的关系？

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



Zero Entropy

如果随机变量的条件熵为0, 那么Y是X的一个函数 (i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$).

Zero Entropy在网络分析/人工智能的推理上具有很大的应用. 当我们遇到条件关系时, 可以考虑用这种方式解决它.

Pf. By condition we have

$$H(Y|X) = \sum_x p(x) H(Y|X = x) = 0$$

Note that $p(x) > 0$, thus for any x , we have $H(Y|X = x) = 0$.

It follows that when x is determined, the distribution of Y is a single value.

0309 Entropy (2)

Relative Entropy

相对熵, 度量两种分布之间的距离 (K-L distance), 假定 $p(x)$, $q(x)$ 具有相同的维数 (over the same alphabet \mathcal{X}), 我们有

$$\begin{aligned} D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= E_p \log \frac{p(X)}{q(X)} \end{aligned}$$

注意, 仅对 p 求期望, 我们还有以下规定或性质

- $0 \log_0^0 = 0, 0 \log \frac{0}{q} = 0, p \log \frac{p}{0} = \infty$
- 若存在事件 x , 使得 $p(x) > 0$ 且 $q(x) = 0$, 那么 $D(p||q) = \infty$
- $D(p||q) \geq 0$
- $D(p||q) = E_p(-\log q(x)) - E_p(-\log p(x)) = E_p(-\log q(x)) - H(p)$, 相对熵可写作一个关于 p 分布的期望减去一个 p 的熵的形式

Relative Entropy is NOT Metric

A metric $d: X, Y \mapsto R^+$ between two distributions should satisfy

- $d(X, Y) \geq 0$
- $d(X, Y) \equiv d(Y, X)$
- $d(X, Y) = 0$ if and only if $X = Y$
- $d(X, Y) + d(Y, Z) \geq d(X, Z)$
- Euclidean distance is a metric
- KL distance is not a metric
 - $D(p||p) = 0$
 - but $D(p||q) \neq D(q||p)$
 - distance but not metric
- **Variational Distance**(差分/变分距离) between p and q is denoted as

$$V(p, q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

- Pinsker's Inequality: 相对熵是差分距离的一个上界

$$D(p||q) \geq \frac{1}{2 \ln 2} V^2(p, q)$$

Conditional Relative Entropy

条件相对熵 = 计算两种单个条件概率分布相对熵, 再对p取平均. 该计算方法也可改写成期望的形式

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)} \end{aligned}$$

性质: Chain Rule $D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

Proof. By Definition

$$\begin{aligned} D(p(x, y)||q(x, y)) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} = \sum_x \sum_y p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_x \sum_y p(x, y) \left(\log \frac{p(x)}{q(x)} + \log \frac{p(y|x)}{q(y|x)} \right) \end{aligned}$$

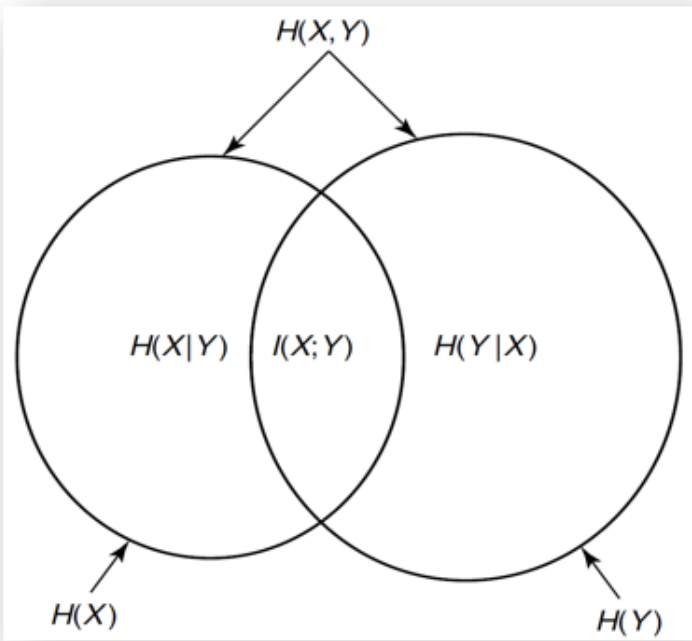
Mutual Information

$$\begin{aligned} I(X; Y) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= D(p(x, y)||p(x)p(y)) \\ &= E_{p(x,y)} \log \frac{p(X, Y)}{p(X)p(Y)} \end{aligned}$$

互信息的性质:

- $I(X; Y) = I(Y; X)$
- $I(X; X) = H(X)$ 随机变量的熵就是它本身互信息的值
- X 和 Y 独立, $I(X; Y) = 0$. 没有关联的随机变量互信息为0
- 注意Notation: $I(X; Y) H(X, Y)$
- 我们通常用关系式计算信息量

Mutual Information and Entropy



$$\begin{aligned}
 I(X; Y) &= H(X) - H(X|Y) \\
 I(X; Y) &= H(Y) - H(Y|X) \\
 I(X; Y) &= H(X) + H(Y) - H(X, Y) \\
 I(X; Y) &= I(Y; X) \\
 I(X; X) &= H(X)
 \end{aligned}$$

Note: 面积可能为负, 所以 $I(X; Y)$ 部分必不可少

Proof. Recall $p(X, Y) = p(X)p(Y|X) = p(Y)p(X|Y)$

To Prove $I(X; Y) = H(X) + H(Y) - H(X, Y)$, Since

$$\log \frac{p(X, Y)}{p(X)p(Y)} = -\log p(X) - \log p(Y) + \log p(X, Y)$$

By taking Expectation of $p(x, y)$ and using $E(X_1 + X_2) = E(X_1) + E(X_2)$, we get $I(X; Y) = H(X) + H(Y) - H(X, Y)$

Proposition:

If X and Y are independent, then

$$H(X, Y) = H(X) + H(Y)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

or vice versa.

Conditional Mutual Information

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$$

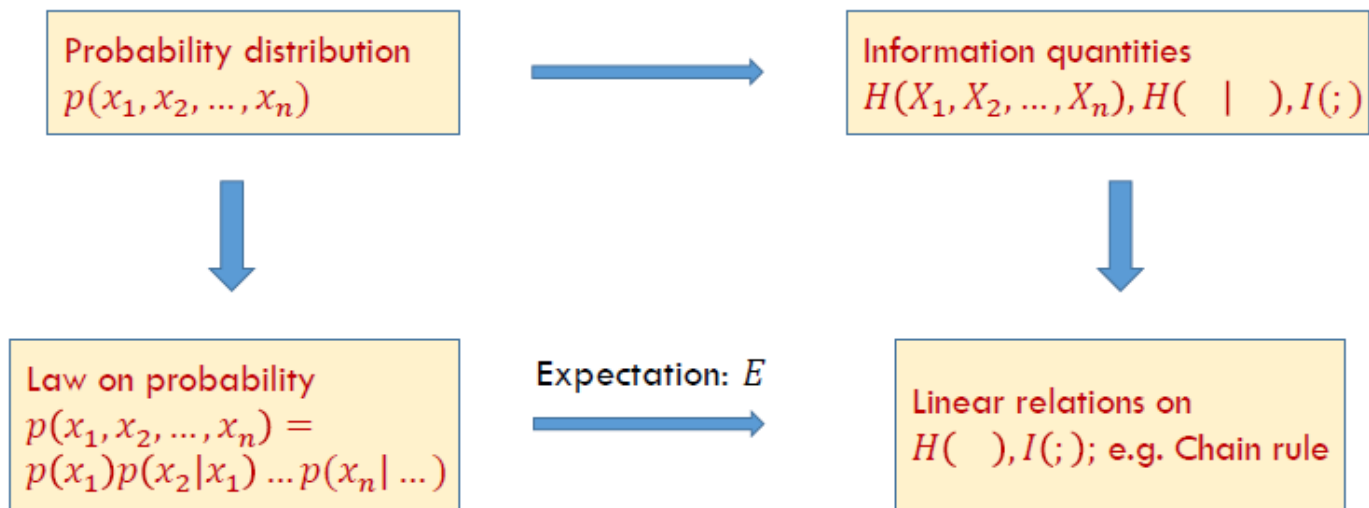
$$= E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}$$

Chain Rule is a decomposition

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

Proof. $I(X_1, X_2, \dots, X_n; Y) = H(X_1, \dots, X_n) - H(X_1, \dots, X_n | Y)$

Apply Chain Rule for entropy respectively, then bind the \sum together, rewrite in the mutual information form.



信息度量的好处是用单一值避免随着随机变量的增多,样本空间指数型的上升

Propositions About Information Quantities

Nonnegative Mutual Information

Information inequality Let $p(x), q(x), x \in X$ be two probability mass functions. Then

$$D(p||q) \geq 0$$

with equality iff $p(x) = q(x)$ for all x .

Proof.

- By convexity:

$$-D(p||q) = \sum p \log \frac{q}{p} \leq \log \sum p \frac{q}{p} = \log \sum q \leq \log 1 = 0$$

- Using $\log x \leq x - 1$ when $x > 0$

$$-D(p||q) = \sum p \log \frac{q}{p} \leq \sum p \left(\frac{q}{p} - 1 \right) = \sum q - \sum p \leq 0$$

$\sum q$ 可能取不到 1, 因为 $p = 0$ 而 $q \neq 0$ 时相对熵被定义为 0, 此时这一项 q 就被消去了. 实际参与运算的 q 之和 ≤ 1

More Properties

- $D(p||q) = 0$ iff $p(x) = q(x)$
- $I(X; Y) \geq 0$, with equality iff X and Y are independent
- $D(p(y|x)||q(y|x)) \geq 0$ with equality iff $p(y|x) = q(y|x)$ for all x and y such that $p(x) > 0$
- $I(X; Y|Z) \geq 0$ with equality iff X and Y are conditionally independent given Z .
- Let $u(x) = \frac{1}{|X|}$ be the uniform probability mass function over X , and let $p(x)$ be the probability mass function for X , Then

$$0 \leq D(p||u) = \log |\mathcal{X}| - H(X)$$

- (Conditioning reduces entropy) (Information can't hurt)

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent

hint: relation with $I(X; Y)$

0311 Entropy (3)

Independence Bound on Entropy

From intuition to math expression.

Let X_1, X_2, \dots, X_n be drawn according to $p(x_1, x_2, \dots, x_n)$. Then

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$$

with equality if and only if the X_i are independent.

Pf. by chain rule and conditioning reduces entropy

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \leq \sum_{i=1}^n H(X_i)$$

Markov Chain

$$p(x, y, z) = p(x)p(y|x)p(z|y) \text{ denoted as } X \rightarrow Y \rightarrow Z$$

$$\text{i.e. } p(z|y, x) = p(z|y)$$

Prop about markov chain

- $X \rightarrow Y \rightarrow Z$ iff X and Z are conditionally independent given Y .
- 时间可逆 $X \iff Y \iff Z$
 - an easy interpretation is that in the mutual information $I(X; Z|Y)$, X and Z can be switched.
- 仿射 if $Z = f(Y)$ then $X \rightarrow Y \rightarrow Z$.
- 体现在信息度量上, if $X \rightarrow Y \rightarrow Z$, then $I(X; Z|Y) = 0$ i.e. Y and Z are conditionally independent given Y .

Pf. from the probability formula we have

$$I(X; Z|Y) = E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}$$

Data Processing Inequality

马尔科夫系统的信息是如何演化的? 马尔可夫链实际上相当于数据分步处理的过程

Theorem : If $X \rightarrow Y \rightarrow Z$, Then $I(X; Y) \geq I(X; Z)$

信息处理得越多,信息丢失的越多

$$\text{Pf. } I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y)$$

Since $I(X; Z|Y) = 0$, we have

$$I(X; Z) + I(X; Y|Z) = I(X; Y)$$

- In particular, if $Z = g(Y)$, then $I(X; Y) \geq I(X; g(Y))$
- Collary: If $X \rightarrow Y \rightarrow Z$, $I(X; Y|Z) \leq I(X; Y)$, 对三个随机变量而言, 条件互信息不一定小于等于互信息(与条件熵不同)

$I(X; Y; Z)$

有关上面Remark的一个反例:

Assume X, Y are two independent random variables uniformly distributed on $\{0, 1\}$.

$$Z = X + Y \pmod{2}$$

We can find that $I(X; Y|Z) > I(X; Y)$.

从问题中, X, Y, Z 任意两个都能决定剩下一个, 分布都相同, 且两两相互独立(by def $p(X, Z) = p(X)p(Z)$).

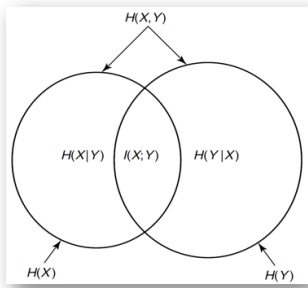
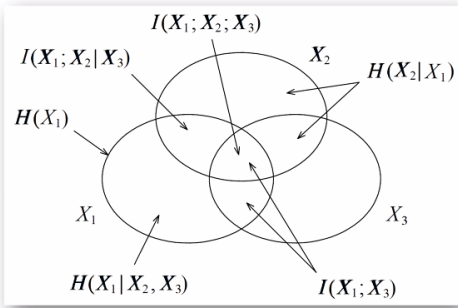
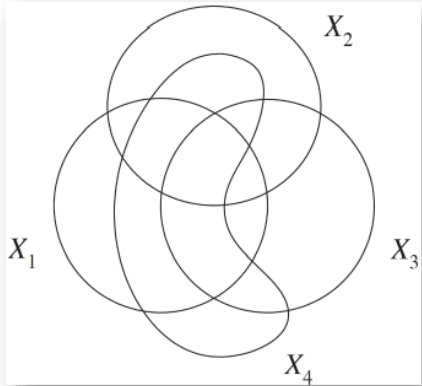
$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(X|Z) \\ &= H(X) \\ &= 1 \end{aligned}$$

$$1 = I(X; Y|Z) > I(X; Y) = 0$$

Intuition: 当你知道 Z 之后, X 和 Y 之间可以解出更多信息出来. 熵就不具备这一性质.
Denote: $I(X; Y; Z) = I(X; Y) - I(X; Y|Z)$ 仅仅是符号意义, 不具有互信息的信息, 因为它可能小于0.

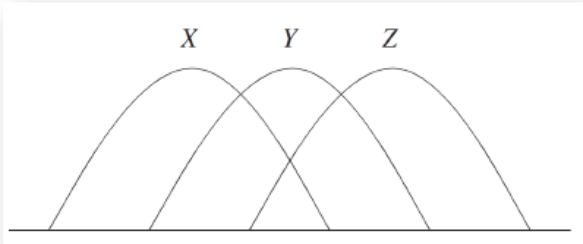
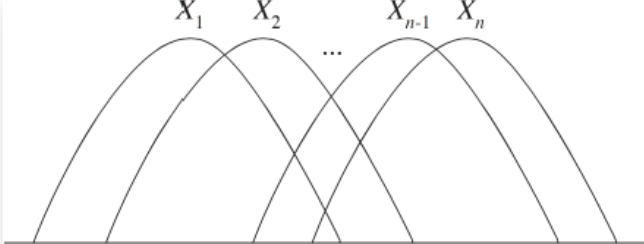
Information Diagram

2RVs to more

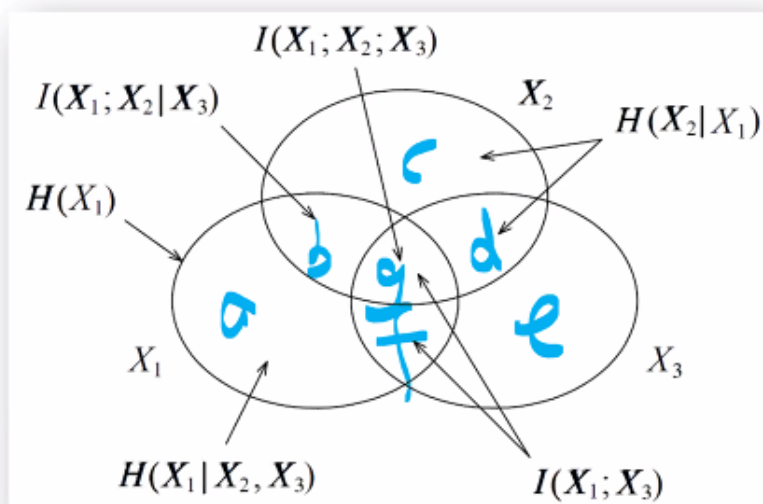
2RVs	3RVs	4RVs
		
Virtual Circles	Area are all nonnegative except $I(X; Y; Z)$	All areas can be expressed in combinations of (conditional) entropy/mutual info

Circles are not representing any entities, for independent variables, unintersected circles are not allowed, since values can be negative.
Only items like $H(X|Y)$, $I(X; Y|Z)$ are nonnegative
Reference: Ch. 3, Information Theory and Network Coding, R. W. Yeung

Markov Chain

$X \rightarrow Y \rightarrow Z$	$X_1 \rightarrow \dots \rightarrow X_n$
	
用半圆表示, 保证相互相交, 保证每块非负	更一般的情况, n个相互相交的半圆
共6块, 三者互信息= X与Z 的互信息	保证第一个和最后一个有明显相交

Examples: Use Info Diagram to Prove Inequalities



$$H(X, Y, Z) \leq \frac{H(X, Y) + H(Y, Z) + H(Z, X)}{2} \leq H(X) + H(Y) + H(Z)$$

$$H(X|Y, Z) + H(Y|X, Z) + H(Z|X, Y) \leq \frac{H(X, Y|Z) + H(Y, Z|X) + H(Z, X|Y)}{2} \leq H(X, Y, Z)$$

With graphical interpretation:

$$a + c + e \leq \frac{(a + b + c) + (c + d + e) + (a + f + e)}{2} \leq a + b + \dots + g$$

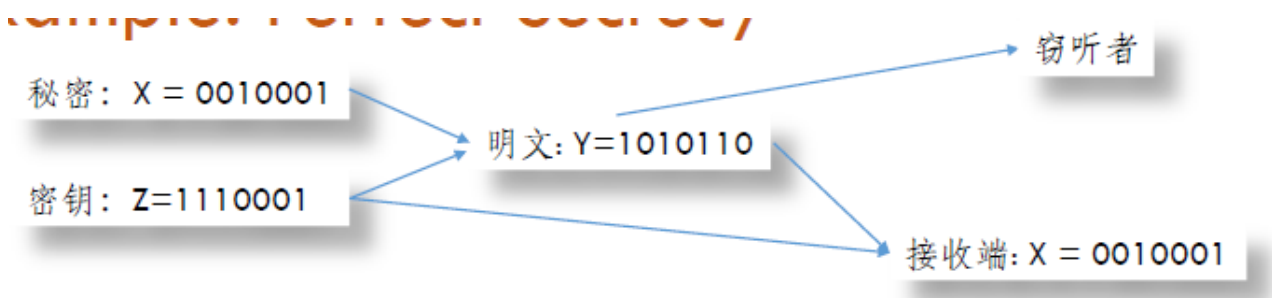
Note, some areas can be negative! Take signals into account

Practical Examples

Causality(因果推断)

我们将系统中的因素用图表示,计算信息量,写出信息量之间的相互关系,推导信息之间是否具有有一定关系
e.g. Given: $X \perp Y|Z$ and $X \perp Z$ and Prove: $X \perp Y$

Perfect Secrecy(完美安全模型)



明文由秘密和密钥生成: $H(Y|X, Z) = 0$

接收端可以通过明文和密钥生成: $H(X|Y, Z) = 0$

我们可以由此推出: $I(X; Y) \geq H(X) - H(Z)$

假设窃听器与秘密之间毫无关联 $I(X; Y) = 0$

那么如果要使 $I(X; Y) = 0$, 我们需要 $H(X) \leq H(Z)$, 即信息长度小于密钥长度.

Fano's Inequality: Estimation

Background

- Suppose that we wish to estimate a random variable X with a distribution $p(x)$.
- We observe a random variable Y that is related to X by the conditional distribution $p(y|x)$.
- From Y , we calculate a function $g(Y) = \hat{X}$, where \hat{X} is an estimate of X and takes on values in $\hat{\mathcal{X}}$.
 - We will not restrict the alphabet $\hat{\mathcal{X}}$ to be equal to \mathcal{X} , and we will also allow the function $g(Y)$ to be random.
- We wish to bound the probability that $\hat{X} \neq X$. We observe that $X \rightarrow Y \rightarrow \hat{X}$ forms a Markov chain. Define the probability of error
 $P_e = \Pr(\hat{X} \neq X)$
- When $H(X|Y)=0$, we know that $P_e = 0$. How about $H(X|Y)$, as $P_e \rightarrow 0$?

Theorem (Fano's Inequality) For any estimator \hat{X} such that $X \rightarrow Y \rightarrow \hat{X}$ with $P_e = \Pr(\hat{X} \neq X)$ we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

Or can be weakened to

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y) \text{ or } P_e \geq \frac{H(X|Y) - 1}{\log |x|}$$

后者是data-processing 不等式, 前者是法诺不等式的核心部分.

Proof

Define an error random variable

$$E = \begin{cases} 0, & \text{if } \hat{X} = X \\ 1, & \text{if } \hat{X} \neq X \end{cases}$$

Then

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}) \end{aligned}$$

- 马尔可夫链implies $H(E|X, \hat{X}) = 0$
- 第二步, $H(X|\hat{X}, E = 1) \leq H(x) = H(P_e)$, 熵永远小于字母表的对数值
- 此外, $H(X|E, \hat{X}) \leq P_e \log |x|$ 因为

$$\begin{aligned} H(X|E, \hat{X}) &= \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1) \\ &\leq (1 - P_e)0 + P_e \log |x| \end{aligned}$$

Corollary Let $P_e = \Pr(X \neq \hat{X})$, and let $\hat{X} : y \rightarrow x$; then $H(P_e) + P_e \log(|x| - 1) \geq H(X|Y)$, 由于已知 X 和 \hat{X} 不等, 在估计时, 熵的上界可以调小 (corollary)

直观理解:

$P_e \rightarrow 0$ implies $H(P_e) \rightarrow 0$ implies $H(X|Y) \rightarrow 0$ 错误率趋向于0时, X和Y的关系趋向确定.

Recall: binary entropy function

$H(p) = -p \log p - (1 - p) \log(1 - p)$ 实际是简写的记号, 计算的是两点分布的熵

Convexity/ Concavity of Information Measures

Log Sum Inequality for nonnegative a_1, \dots, a_n and b_1, \dots, b_n ,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

with equality if and only if $\frac{a_i}{b_i} = \text{const.}$

Pf. by moving $(\sum_{i=1}^n a_i)$ to the left, the coefficient can be regarded as a probability distribution.

Corollaries:

- Concavity of $H(P)$
- (X, Y) $p(x, y) = p(x)p(y|x)$, then $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$
 - Note given $p(x)$, **ParseError: KaTeX parse error: Undefined control sequence: \rightarrow at position 27: ...htarrow p(x,y) \rightarrow p(y)**
- Convexity of relative entropy. $D(p||q)$ is a convex function for pair (p, q) .
- 可以把 $p(x), p(y)$ 等看作高维空间上的一个点, 而不是概率分布. for $X = 1, 2, 3, \dots, n$, define $x_1 = p_1, x_2 = p_2, \dots$

0316 AEP

Law of Large Numbers

随机变量的收敛性

1. **In probability** if for every $\epsilon > 0$, $\Pr\{|X_n - X| > \epsilon\} \rightarrow 0$
2. **In mean square** if $E(X_n - X)^2 \rightarrow 0$
3. **With probability 1** (almost surely) if $\Pr\{\lim_{n \rightarrow \infty} X_n = X\}$
Note $(2) \rightarrow (1)$ $(3) \rightarrow (1)$, proof by Markov and Chebyshev.

强大数定律: For i.i.d random variables, $\bar{X}_n \rightarrow E(X_1)$ with probability 1.

弱大数定律: For i.i.d random variables, $\bar{X}_n \rightarrow E(X_1)$ in probability.

AEP (Asymptotic Equipartition Property)

渐进均分性, 大数定律在信息论中的体现.

Theorem If X_1, X_2, \dots are i.i.d. $\sim p(x)$, then

$$\begin{aligned} -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) &= -\frac{1}{n} \sum_i \log p(X_i) \\ &\rightarrow -E \log p(X) \text{ in probability} \\ &= H(X) \end{aligned}$$

Proof.

AEP应用于数据压缩算法中, 我们用极限的语言写出证明:

$$\begin{aligned} H(X) - \epsilon &\leq -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \leq H(X) + \epsilon \\ 2^{-n(H(X)+\epsilon)} &\leq p(X_1, X_2, \dots, X_n) \leq 2^{-n(H(X)-\epsilon)} \Rightarrow A_\epsilon^{(n)} \end{aligned}$$

Reason:

- Functions of independent random variables are also independent random variables.
- Since the X_i are i.i.d., so are $\log p(X_i)$
- By the weak law of large numbers

将极限的定义展开, 我们可以通过如下方式导出并定义典型集

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \leq H(X) + \epsilon$$

也即

$$2^{-n(H(X)+\epsilon)} \leq p(X_1, X_2, \dots, X_n) \leq 2^{-n(H(X)-\epsilon)} \Rightarrow A_\epsilon^{(n)}$$

导出典型集的概念.

Typical Set

Definition 样本空间中, 所有满足 $2^{-n(H(X)+\epsilon)} \leq p(X_1, X_2, \dots, X_n) \leq 2^{-n(H(X)-\epsilon)}$ 的 (x_1, x_2, \dots, x_n)

Properties

1. If $(x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots, x_n) \leq H(X) + \epsilon$

Proof. 见上一节的概念导出

2. 典型集中所有元素概率之和接近于1. $\Pr \left\{ A_\epsilon^{(n)} \right\} \geq 1 - \epsilon$ for n sufficiently large.

Proof. 由AEP, 我们知道给定任意 ϵ , 对任意 $\delta > 0$, 存在 n_0 , 对任意 $n > n_0$,

$$\Pr \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) - H(X) \right| > \epsilon \right\} < \delta$$

equivalently,

$$\Pr \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) - H(X) \right| < \epsilon \right\} > 1 - \delta$$

注意典型集的定义, 有

$$\Pr \left\{ A_\epsilon^{(n)} \right\} = \Pr \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) - H(X) \right| < \epsilon \right\}$$

Setting $\delta = \epsilon$,

$$\Pr \left\{ A_{\epsilon}^{(n)} \right\} \geq 1 - \epsilon$$

3. 典型集的大小存在上界: $\left| A_{\epsilon}^{(n)} \right| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in the set A

Proof.

$$\begin{aligned} 1 &= \sum_{x \in \mathcal{X}^n} p(x) \\ &\geq \sum_{x \in A_{\epsilon}^{(n)}} p(x) \\ &\geq \sum_{x \in A_{\epsilon}^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= 2^{-n(H(X)+\epsilon)} \left| A_{\epsilon}^{(n)} \right|_{(X)} \end{aligned}$$

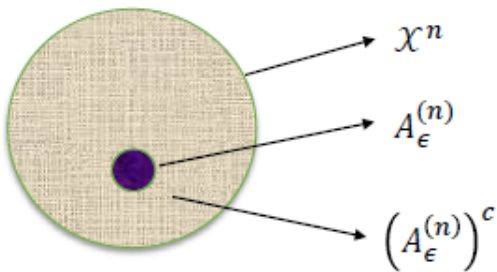
Thus, $\left| A_{\epsilon}^{(n)} \right| \leq 2^{n(H(X)+\epsilon)}$

进一步, 我们可以说明

$$\frac{\left| A_{\epsilon}^{(n)} \right|}{|\mathcal{X}^n|} \leq 2^{n(H(X)-\log |\mathcal{X}|)} \rightarrow 0$$

(在 \mathcal{X} 非均匀分布时, $H(X) < \log |\mathcal{X}|$)

Recall: 性质2实质说明了 $\Pr(x^n) \approx \Pr(A_{\epsilon}^{(n)})$, 结合这两点我们可以获得如图所示的典型集的直观理解.



4. $\left| A_{\epsilon}^{(n)} \right| \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Proof. For sufficiently large n , $\Pr \left\{ A_{\epsilon}^{(n)} \right\} > 1 - \epsilon$, so that

$$\begin{aligned}
1 - \epsilon &< \Pr \left\{ A_e^{(n)} \right\} \\
&\leq \sum_{x \in A_e^{(n)}} 2^{-n(H(X) - \epsilon)} \\
&= 2^{-n(H(X) - \epsilon)} \left| A_e^{(n)} \right|
\end{aligned}$$

$$\text{Thus } \left| A_e^{(n)} \right| \geq (1 - \epsilon) 2^{n(H(X) - \epsilon)}$$

High Probability Set

给出更宽泛的定义, 高概率集

Definition For each $n = 1, 2, \dots$, let $B_\delta^{(n)} \subseteq x^n$ be the smallest set with

$$\Pr \left\{ B_\delta^{(n)} \right\} \geq 1 - \delta$$

Theorem 高概率集元素大小的下界 Let X_1, X_2, \dots, X_n be i.i.d $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\Pr \left\{ B_\delta^{(n)} \right\} \geq 1 - \delta$, then

$$\frac{1}{n} \log \left| B_\delta^{(n)} \right| > H - \delta'$$

for n sufficiently large.

高概率集和 2^{nH} 是同阶的, 我们可以说典型集可能是最小的高概率集.

Intuition As $A_\epsilon^{(n)}$ has $2^{n(H \pm \epsilon)}$ elements, $\left| B_\delta^{(n)} \right|$ and $\left| A_\epsilon^{(n)} \right|$ are equal to the first order in the exponent

Idea: 高概率集和典型集的概率分布的交集应该也是很大的, 否则缺失会比较严重, 会与1有明显的差距. 因此我们的证明就是研究 $\Pr \left(A_\epsilon^{(n)} \cap B_\delta^{(n)} \right)$

Proof. For any two sets A, B , if $\Pr(A) \geq 1 - \epsilon_1$ $\Pr(B) \geq 1 - \epsilon_2$, then $\Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2$ 首先从概率意义上得到一个trivial的结论

$$\begin{aligned}
1 - \epsilon - \delta &\leq \Pr \left(A_\epsilon^{(n)} \cap B_\delta^{(n)} \right) = \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H - \epsilon)} \\
&= \left| A_\epsilon^{(n)} \cap B_\delta^{(n)} \right| 2^{-n(H - \epsilon)} \leq \left| B_\delta^{(n)} \right| 2^{-n(H - \epsilon)}
\end{aligned}$$

得出结论: 高概率集必定占据了典型集大部分的空间

$$\left| B_\delta^{(n)} \right| \geq \left| A_\epsilon^{(n)} \cap B_\delta^{(n)} \right| \geq 2^{n(H - \epsilon)} (1 - \epsilon - \delta)$$

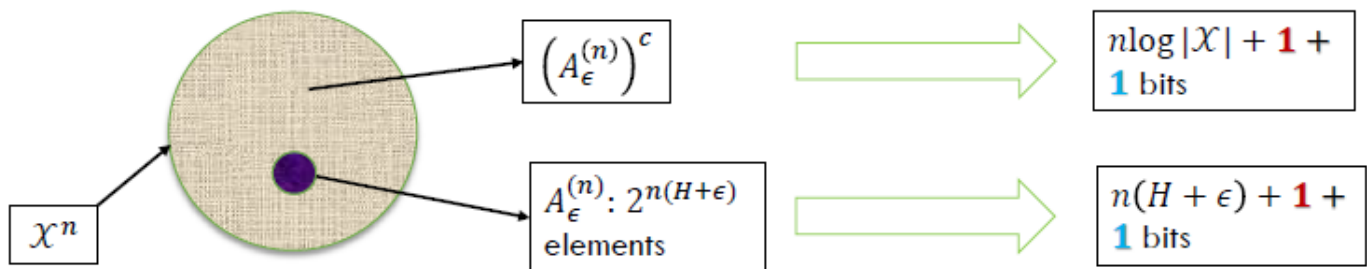
Data Compression

数据源(iid) $X^n = (X_1, \dots, X_n)$ ----> Encoder ----m bits --> Decoder --> \hat{X}^n

Problem Formulation

- Source: X_1, X_2, \dots , are i.i.d. $\sim p(X)$. 数据源, iid的假设虽然有时对实际问题过于强, 但理论上是需要的
- Source sequences: $X^n = (X_1, \dots, X_n)$ denotes the n -tuple that represents a sequence of n source symbols
- Alphabet: $x = \{1, 2, \dots, |x|\}$ — the possible values that each X_i can take on
- Encoder and decoder are a pair of functions f, g such that **ParseError: KaTeX parse error: Can't use function '\$' in math mode at position 28: ...rrow\{0,1\}^*\}\$ and \$g:\{0,1\}...**
- Probability of error $P_e = P(X^n \neq \hat{X}^n)$ 我们通过解码器获得信息, 希望解码后的错误率能够在n很大时, 无穷趋向于0
 - If $P_e = 0$, "lossless" 无损编码, otherwise "lossy" 有损编码
- The rate of a scheme 码率: $R = \frac{m}{n}$ ($R = \log |X|$ is trivial!) n 个随机变量用m个码来编, 这里的R不一定是最优的, 我们希望找到尽可能小的R.(及其对应的encoder,decoder)
ToDo: Find an encoder and decoder pair such that $P_e \rightarrow 0$, as $n \rightarrow \infty$

Procedure



实际中, 我们没有必要给每个样本相同的编码长度, 我们要对样本空间进行划分. 比如, 对典型集等高概率集区分开来

- 非典型集中, 我们至少需要 $n \log |\mathcal{X}| + 1 + 1$ 1个bit凑整, 另一个bit用以区分非典型集
- 典型集需要 $n(H + \epsilon) + 1 + 1$ 区分

Divide and conquer: $x^n \in A_\epsilon^{(n)}$ and $x^n \notin A_\epsilon^{(n)}$

- $x^n \in A_\epsilon^{(n)}$:
 - since there are $\leq 2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, the indexing requires no more than $n(H + \epsilon) + 1$ bits. [The extra bit may be necessary because $n(H + \epsilon)$ may not be an integer.]
- $x^n \notin A_\epsilon^{(n)}$:
 - Similarly, we can index each sequence not in $A_\epsilon^{(n)}$ by using not more than $n \log |\mathcal{X}| + 1$ bits.

- To deal with overlap in the $\{0, 1\}$ sequences, 但这样的编码可能会带来冲突, 比如 $\{0,0\},\{0,0\}$, 所以我们在起始位置再加一位进行区分
 - We prefix all these sequences by a 0 , giving a total length of $\leq n(H + \epsilon) + 2$ bits to represent each sequence in $A_\epsilon^{(n)}$
 - Prefixing these indices by 1 , we have a code for all the sequences in X^n .

Analysis

分析一下期望长度.

$$\begin{aligned}
 E(l(X^n)) &= \sum_{x^n} p(x^n) l(x^n) \\
 &= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) \\
 &\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) (n(H + \epsilon) + 2) + \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) (n \log |x| + 2) \\
 &= \Pr \left\{ A_\epsilon^{(n)} \right\} (n(H + \epsilon) + 2) + \Pr \left\{ \left(A_\epsilon^{(n)} \right)^c \right\} (n \log |x| + 2) \\
 &\leq n(H + \epsilon) + \epsilon n (\log |x|) + 2 \\
 &= n(H + \epsilon') \quad (\epsilon' \text{ 是另一个同样一阶的无穷小量})
 \end{aligned}$$

导出平均长度的期望值是以 $H(X)$ 为上界的

$$E \left[\frac{1}{n} l(X^n) \right] \leq H(X) + \epsilon$$

Thus, we can represent sequences X^n using $nH(X)$ bits on the average. 才能保证恢复

为了说明这一点, 我们还要说明 $H(X)$ 是最小的码率

Converse For any scheme with rate $r < H(X)$, $P_e \rightarrow 1$ 不仅不趋向于0, 而且直接趋向于1

Proof. Let $r = H(X) - \epsilon$. For any scheme with rate r , it can encode at most 2^{nr} different symbols in \mathcal{X}^n .

The correct decoding probability is $\approx 2^{nr} 2^{-nH} = 2^{-n(H-r)} \rightarrow 0$

Thus, $P_e \rightarrow 1$

0318 Entropy Rate

AEP研究了独立同分布的随机变量列, 我们希望得到更广泛的结论.

Stochastic Process

Introduction

A stochastic process $\{X_i\}$ is an indexed sequence of random variables.

一个例子: 赌徒的破产.

- 下一局的输赢概率与上一局无关
- p 概率得1元, $1-p$ 概率丢一元
- 得到 $X_{i+1} = X_i \pm 1$
- Thus X_i 's are not i.i.d.

Stationary Process

Definition A stochastic process is said to be

stationary (稳态) if the **joint distribution of any subset of the sequence of random variables is invariant** with respect to shifts in the time

index; that is, 任意随机变量的分布按照时间平移, 得到的联合概率分布是不变的.

$$\begin{aligned} & \Pr \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} \\ &= \Pr \{X_{1+l} = x_1, X_{2+l} = x_2, \dots, X_{n+l} = x_n\} \end{aligned}$$

for every n and every shift l and for all $x_1, x_2, \dots, x_n \in X$

性质:

1. 平移不变性:
 - $p(X_1) = p(X_2) = \dots = p(X_n)$
 - $p(X_1, X_3) = p(X_2, X_4) \dots$
2. 高斯过程是一个稳态过程
3. 达到稳定状态后的马尔可夫链, 同分布而不独立
4. 我们的定义中只说明了稳态的分布, 这是强稳态, 与之相对应的是弱稳态. 在使用时, 两者没有严格的推出关系

通过稳态分布的定义, 我们可以证明一些结论.

Theorem 时间单向性, Time's arrow. Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0 | X_{-1}, X_{-2}, \dots, X_{-n}) = H(X_0 | X_1, X_2, \dots, X_n)$$

即时间上的前后行动对系统没有影响.

Proof. 平移不变性, 首先有相同分布 $H(X_{-n}, \dots, X_0) = H(X_0, \dots, X_n)$, $H(X_{-n}, \dots, X_{-1}) = H(X_1, \dots, X_n)$. 相减得证.

Markov Chain

The Markov chain is said to be **time invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n ; that is, for $n = 1, 2, \dots$

$$\Pr\{X_{n+1} = b | X_n = a\} = \Pr\{X_2 = b | X_1 = a\} \quad \text{for all } a, b \in X$$

We will assume that the Markov chain **is time invariant unless otherwise stated**

A time-invariant Markov chain is characterized by its initial state and a probability transition matrix $P = [P_{ij}]$, $i, j \in \{1, 2, \dots, m\}$, where

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

Example:

- Gambler's ruin
- Random Walk

Stationary Distribution of MC

- By the definition of stationary, a Markov chain is stationary iff $p(X_{n+1}) = p(X_n)$
- If the probability mass function at time n is $p(x_n)$, then

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}} \quad \text{or} \quad x^T P = x^T$$

- If the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain is stationary
- Example: Consider a two state Markov chain with a probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

$$(\mu_1, \mu_2) \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = (\mu_1, \mu_2)$$

另解, 对小规模的网络, 由最大流-最小割定理, 考虑任意割集之间的流进与流出概率为0, For stationary distribution, the net probability flow across any cut set is zero

$$\mu_1 \alpha = \mu_2 \beta$$

$$\mu_1 + \mu_2 = 1$$

$$\mu_1 = \frac{\beta}{\alpha + \beta} \quad \text{and} \quad \mu_2 = \frac{\alpha}{\alpha + \beta}$$

Entropy Rate

对复杂系统, 我们难以用一个时刻随机变量的熵, 我们希望描述熵的演化形式. 我们取系统联合熵的极限.

Definition The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$H(x) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

when the limits exists(熵率也可能不存在)

计算方式:

$$H(X_n, \dots, X_1) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$

For $H(X_i | X_{i-1}, \dots, X_1)$, we now need to make clear of

- the existence of

$$\lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1)$$

- In a series $\{a_n\}$, if $a_n \rightarrow a$, the existence of

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n a_i$$

H'(X)

对稳态随机过程, 我们有如下性质:

Theorem For a stationary stochastic process, $H(X_n | X_{n-1}, \dots, X_1)$ is nonincreasing in n and has a limit.

Proof.

$$\begin{aligned} & H(X_{n+1} | X_n, \dots, X_1) \\ & \leq H(X_{n+1} | X_n, \dots, X_2) \\ & = H(X_n | X_{n-1}, \dots, X_1) \\ & H(X_n | X_{n-1}, \dots, X_1) \geq 0 \end{aligned}$$

根据数列极限的结论(MCT): since $\{H(X_n | X_{n-1}, \dots, X_1)\}$ is nonincreasing and $H(X_n | X_{n-1}, \dots, X_1) \geq 0$, the limit exists.

Define

$$H'(x) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$$

we have The limits $H'(x)$ exists

Cesaro Mean

Recall in Calculus,

$$\text{If } a_n \rightarrow a \text{ and } b_n = \frac{1}{n} \sum_{i=1}^n a_i, \text{ then } b_n \rightarrow a$$

Proof.

Let $\epsilon > 0$. since $a_n \rightarrow a$, there exists a number $N(\epsilon)$ such that $|a_n - a| \leq \epsilon$ for all $n \geq N(\epsilon)$. Hence

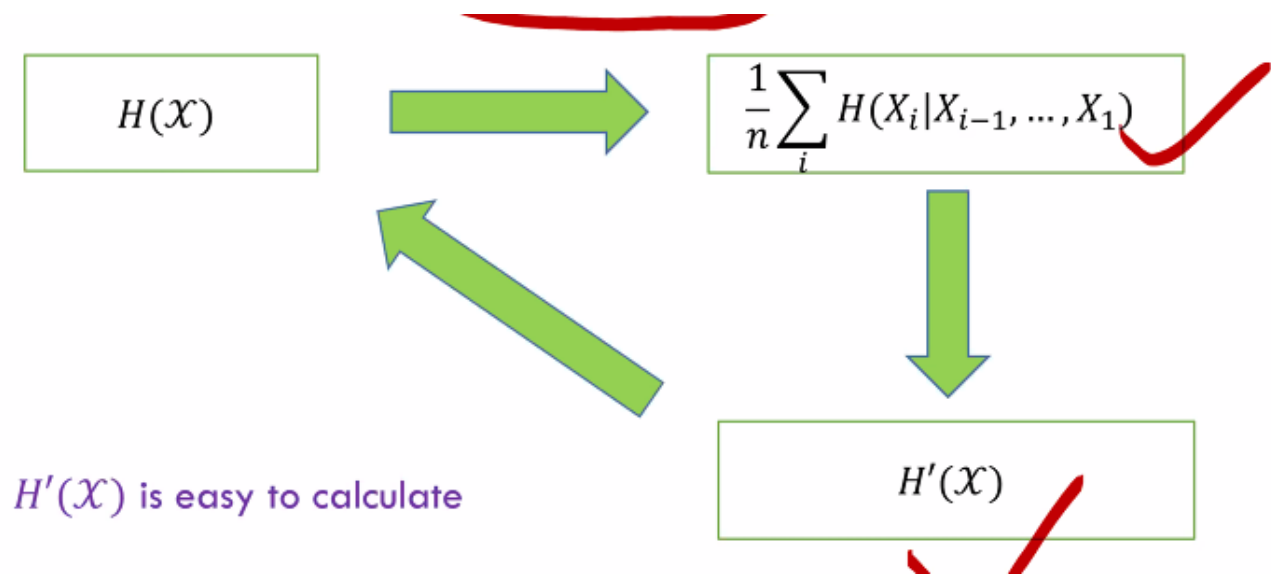
$$\begin{aligned} |b_n - a| &= \left| \frac{1}{n} \sum_{i=1}^n (a_i - a) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n |a_i - a| \\ &\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n - N(\epsilon)}{n} \epsilon \\ &\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon \end{aligned}$$

Thus, $|b_n - a| \leq \epsilon'$, for all $n \geq N(\epsilon)$.

Entropy Rate for Stationary Process

Theorem. For a **stationary stochastic process**, the limits in $H(X)$ and $H'(X)$ exist and are equal:

$$H(x) = H'(x)$$



Entropy Rate for Markov Chain

For a **stationary Markov chain**, the entropy rate is given by

$$\begin{aligned} H(x) &= H'(x) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}) \\ &= H(X_2 | X_1) \end{aligned}$$

where the conditional entropy is calculated using the given stationary distribution.

Recall that the stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij} \text{ for all } j$$

Theorem Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P . Let $X_1 \sim \mu$. Then the entropy rate is

$$H(x) = - \sum_{ij} \mu_i P_{ij} \log P_{ij}$$

Proof.

$$H(x) = H(X_2 | X_1) = \sum_i p(x_i) H(X_2 | X_1 = x_i) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij} \right)$$

Example: Random Walk

Undirected graph with weight $W_{ij} \geq 0$ and $W_{ij} = W_{ji}$

我们计算如下参量, 解题思路:

$$P_{ij} = W_{ij} / \sum_k W_{ik}$$

$$W_i = \sum_j w_{ij}$$

$$W = \sum_i \frac{W_i}{2}$$

The stationary distribution is

$$\mu_i = \frac{W_i}{2W}$$

Verify it by $\mu P = \mu$

$$H(x) = H(X_2|X_1)$$

$$= H\left(\dots, \frac{W_{ij}}{2W}, \dots\right) - H\left(\dots, \frac{W_i}{2W}, \dots\right)$$

Second Law of Thermodynamics

Intuition: 我们观测的状态会受到基本原则的控制. 我们将物理过程抽象成马尔可夫链, 基本原则则抽象成转移矩阵

- One of the basic laws of physics, the second law of thermodynamics, states that the entropy of an isolated system is nondecreasing.
- We model the isolated system as a **Markov chain with transitions obeying the physical laws governing the system.**
 - Implicit in this assumption is the notion of an overall state of the system and the fact that knowing the present state, the future of the system is independent of the past.

Some Results

- Relative entropy $D(\mu_n || \mu'_n)$ decreases with n
(Pf. by 相对熵链式法则, 因为转移矩阵是不变的, 系统会趋于稳定)
- The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process
- Shuffles increase entropy: $H(TX) \geq H(X)$

Reference: Neri Merhav (2010), "Statistical Physics and Information Theory," Foundations and Trends® in Communications and Information Theory

Extension: Functions of Markov Chains

本节的重点是熵率的性质, 此处介绍一个算法, 了解即可.

问题背景: 我们通过一些现象知道了一些看不见物质的存在. 我们用 X 表示看不到的现象, 他们通过一些物理法则, Φ 生成了现象 Y . 如果 X 是稳态的, 我们如何计算观测结果的熵率?

$$\begin{array}{ccccccc}
 & X_1 & & X_2 & & \dots & & X_n & & \dots \\
 & \downarrow & & \downarrow & & \dots & & \downarrow & & \dots \\
 Y_1 = \phi(X_1) & & Y_2 = \phi(X_2) & & \dots & & Y_n = \phi(X_n) & & \dots
 \end{array}$$

Let $X_1, X_2, \dots, X_n, \dots$ be a stationary Markov chain, and let $Y_i = \phi(X_i)$ be a process each term of which is a function of the corresponding state in the Markov chain. What is the entropy rate of $H(y)$?

- $\{Y_i\}$: A very special case of hidden Markov model (HMM) 隐马尔科夫模型, 在信号处理等情况下十分常见
- $\{Y_i\}$ is not a Markov chain in general 并不一定是马尔可夫链, 因为Y取决于单个X的取值. 我们是无法直接得到Y之间的状态转移矩阵.
- $\{X_i\}$ is stationary $\Rightarrow \{Y_i\}$ is stationary, 但稳态分布是肯定的, 熵率还是可以通过求条件熵的极限获得

$$H(y) = \lim_{n \rightarrow \infty} H(Y_n | Y_{n-1}, \dots, Y_1)$$

- Drawback: Hard to ensure the convergence by n
- Solution: We have already known that $H(Y_n | Y_{n-1}, \dots, Y_1)$ is lower bounded by $H(Y)$ 已知熵率是条件熵的下界.
 - Find a lower bound for $H(y)$ which is close to $H(Y_n | Y_{n-1}, \dots, Y_1)$
- Let's have a look at X_1
 - X_1 contains much information about Y_n as Y_1, Y_0, Y_{-1}, \dots

$$H(Y_n | Y_{n-1}, \dots, Y_1, X_1)$$

(Y_1 could be ignored) 这是没问题的, 因为 Y_1 是 X_1 的函数.

Theorem. If X_1, X_2, \dots, X_n form a stationary Markov chain, and $Y_i = \phi(X_i)$, then

$$H(Y_n | Y_{n-1}, \dots, Y_1, X_1) \leq H(y) \leq H(Y_n | Y_{n-1}, \dots, Y_1)$$

and $\lim H(Y_n | Y_{n-1}, \dots, Y_1, X_1) = H(y) = \lim H(Y_n | Y_{n-1}, \dots, Y_1)$ 我们希望运用夹逼定理.

Proof.

1. (handled) Y的条件熵是熵率的上界
2. 带 X_1 的条件熵是熵率的下界

运用马尔可夫链的性质添加负项, 挪去条件熵增大, 应用平移不变性, 发现递减, 运用MCT

$$\begin{aligned}
& H(Y_n | Y_{n-1}, \dots, Y_2, X_1) \\
&= H(Y_n | Y_{n-1}, Y_2, Y_1, X_1) \\
&= H(Y_n | Y_{n-1}, \dots, Y_1, X_1, X_0, X_{-1}, \dots, X_{-k}) \\
&= H(Y_n | Y_{n-1}, \dots, Y_1, X_1, X_0, X_{-1}, \dots, X_{-k}, Y_0, \dots, Y_{-k}) \\
&\leq H(Y_n | Y_{n-1}, \dots, Y_1, Y_0, \dots, Y_{-k}) \\
&= H(Y_{n+k+1} | Y_{n+k}, \dots, Y_1) \\
& \quad k \rightarrow \infty \\
& H(Y_n | Y_{n-1}, \dots, Y_2, X_1) \leq H(\mathcal{Y})
\end{aligned}$$

3. 带 X_1 条件下, 随着 n 的增大, 不等式两边熵无限接近, 即研究互信息趋向于0.

$$H(Y_n | Y_{n-1}, \dots, Y_1) - H(Y_n | Y_{n-1}, \dots, Y_1, X_1) = I(X_1; Y_n | Y_{n-1}, \dots, Y_1)$$

首先注意到互信息小于熵

$$I(X_1; Y_1, Y_2, \dots, Y_n) \leq H(X_1)$$

这在极限条件下也成立, 运用链式法则展开

$$\begin{aligned}
H(X_1) &\geq \lim_{n \rightarrow \infty} I(X_1; Y_1, Y_2, \dots, Y_n) \\
&= \lim_{n \rightarrow \infty} \sum_{i=1}^n I(X_1; Y_i | Y_{i-1}, \dots, Y_1) \\
&= \sum_{i=1}^{\infty} I(X_1; Y_i | Y_{i-1}, \dots, Y_1)
\end{aligned}$$

无穷级数, 每一项都是正的, 极限存在, 那么对 n 足够大级数项趋向于0.

$$\begin{aligned}
& I(X_1; Y_n | Y_{n-1}, \dots, Y_2, Y_1) \rightarrow 0 \\
& \quad \parallel \\
& (Y_n | Y_{n-1}, \dots, Y_2, Y_1) - H(Y_n | Y_{n-1}, \dots, Y_1, X_1)
\end{aligned}$$

0323 Data Compression (1)

Example of Codes

Let X be a random variable with the following distribution and codeword assignment:

$$\begin{aligned}\Pr(X = 1) &= 1/2, & \text{codeword } C(1) &= 0 \\ \Pr(X = 2) &= 1/4, & \text{codeword } C(2) &= 10 \\ \Pr(X = 3) &= 1/8, & \text{codeword } C(3) &= 110 \\ \Pr(X = 4) &= 1/8, & \text{codeword } C(4) &= 111\end{aligned}$$

- Without loss of generality, we can assume that the D -ary alphabet is $\mathcal{D} = \{0, 1, \dots, D-1\}$. 二进制中, $D=2$
- 信源编码 C for a random variable X is a mapping from X to D^* , the set of finite-length strings of symbols from a D -ary (D 元组) alphabet.
- Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$
- The expected length $L(C)$ of a source code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by

$$L(C) = \sum_{x \in X} p(x)l(x)$$

- What is $\min L(C)$
- How to construct such an optimal code
- Recall 在AEP中, 码率大于等于熵

Nonsingular Code

- 编码系统不希望两个字母有同样的编码. A code is said to be nonsingular if every element of the range of X maps into a different string in D^* ; that is,

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

- 定义编码的连接 The extension C^* of a code C is the mapping from finite length strings of X to finite-length strings of D , defined by

$$C(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n)$$

唯一可解码, 在扩展是非奇异的情况下, 编码是可解码的. A code is called uniquely decodable if its extension is nonsingular.

- In other words, any encoded string in a uniquely decodable code has only one possible source string producing it.

在对系统要求更高的情况下, 我们希望实时的解码,

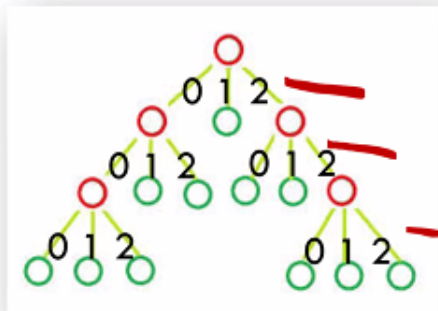
- 前缀码: 任意编码都不是另一个码的前缀
- 后缀码: 任意编码都不是另一个码的后缀

How to construct?

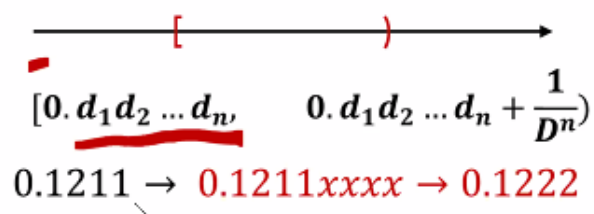
Prefix Code

由前缀码的性质, 我们可以用一些良好的结构描述编码方式.

如, 我们用三叉树表示三进制的码制.



我们也可以用区间来表示.



各个码制对应的左闭右开区间是 $[0, 1)$ 的一个互不相交的分割.

Kraft Inequality

从数学上解决了前缀码存在的一个重要条件.

(Kraft Inequality 1949) For any instantaneous code (prefix code) over an alphabet of size D , the codeword lengths l_1, l_2, \dots, l_m must satisfy the inequality

$$\sum_{i=1}^m D^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

任意两个码制对应的路径是互相不能覆盖的. 当我选定一个路径之后, 我就不可能再将路径的终点展开(叶子转换成子树).

Assume $l_1 \leq l_2 \leq \dots \leq l_m$ (The maximum depth is l_m)

- For l_i , it "occupied" a subtree in size $D^{l_m - l_i}$ 它把接下来的空间覆盖了
- The aggregate size of subtrees

$$\sum_{i=1}^m D^{l_m - l_i}$$

- 对D叉树, 在 l_m 层有必要条件

$$\sum_{i=1}^m D^{l_m - l_i} \leq D^{l_m} \Rightarrow \text{"only if"}$$

- "if": mathematical induction 利用对m进行数学归纳法证明充分性
- m-1:

$$\sum_{i=1}^{m-1} D^{l_m - l_i} < D^{l_m}$$

- 一定存在没有覆盖的m

Extended Kraft Inequality

(Extended Kraft Inequality) For any countably infinite set of codewords that form a prefix code, the codeword lengths satisfy the extended Kraft inequality,

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1$$

Conversely, given any l_1, l_2, \dots satisfying the extended Kraft inequality, we can construct a prefix code with these codeword lengths.

证明上, 树的表示不再适用. 我们用区间的方法.

必要性:

Let the D-ary alphabet be $\{0, 1, \dots, D-1\}$. Consider the i th codeword $y_1 y_2 \dots y_{l_i}$

Let $0.y_1 y_2 \dots y_{l_i}$ be the real number given by the D-ary expansion

$$0.y_1 y_2 \dots y_{l_i} = \sum_{j=1}^{l_i} y_{l_j} D^{-j}$$

This codeword corresponds to the interval

$$\left[0.y_1 y_2 \dots y_{l_i}, 0.y_1 y_2 \dots y_{l_i} + \frac{1}{D^{l_i}} \right)$$

- This is a subinterval of the unit interval $[0, 1]$
- By the prefix condition, these intervals are disjoint.

充分性:

将 l_1, l_2 递增排序, 从左往右依次切割 $[0, 1]$ 区间, 就可以构造前缀码.

Optimal Codes

Problem Formulation

Kraft inequality gives a mathematical expression on the existence of prefix code. The problem of finding the prefix code with the minimum expected length could be formulated as a standard optimization problem

$$\min L = \sum p_i l_i$$

such that $\sum D^{-l_i} \leq 1$

考虑到 l_i 是整数较为复杂, 我们首先考虑实数的情况

By Lagrange, their gradient vectors are parallel $\nabla f(X) = \lambda \nabla g$

Solution

The Lagrange multipliers

$$J = \sum p_i l_i + \lambda \left(\sum D^{-l_i} - 1 \right)$$

Differentiating with respect to l_i , we obtain

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \log_e D$$

Setting the derivatives to 0, we obtain

$$D^{-l_i} = \frac{p_i}{\lambda \log_e D}$$

又因为 $\sum D^{-l_i} = 1$

Substituting this in the constraint to find λ , we find $\lambda = 1/\log_e D$, and hence

$$p_i = D^{-l_i}$$

yielding optimal code lengths,

$$l_i^* = -\log_D p_i$$

This noninteger choice of codeword lengths yields expected codeword length

$$L^* = \sum p_i l_i^* = \sum -p_i \log p_i = H_D(X)$$

不一定是整数取得到, 但这可以是一个下界.

In general, $H_D(X)$ cannot be attained

$$L^* \geq H_D(X)$$

然而, 我们通过推导得知, 在最优情况下, 有下面的关键关系.

$$p_i = D^{-l_i}$$
$$l_i = -\log p_i$$

Bounds

进一步, 码制在最优情况下的平均情况能满足以下不等式, 得到最优编码的上界是熵+1bit

Let $l_1^*, l_2^*, \dots, l_m^*$ be optimal codeword lengths for a source distribution \mathbf{p} and D -ary alphabet, and let L^* be the associated expected length of an optimal code ($L^* = \sum p_i l_i^*$) Then

$$H_D(X) \leq L^* < H_D(X) + 1$$

Proof.

向上取整

Recall that $p_i = D^{-l_i}$ and $l_i = -\log_D p_i$

since $\log_D \frac{1}{p_i}$ may not equal to an integer, we round it up to give integer word-length assignments,

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil \Rightarrow \text{Shannon codes}$$

我们可以验证新的编码长度满足Kraft不等式(由prefix区间不相交的性质保证)

Check l_i' s satisfying Kraft inequality.

$$\log_D \frac{1}{p_i} \leq l_i < \log_D \frac{1}{p_i} + 1$$

Take expectations

$$H_D(X) \leq L < H_D(X) + 1$$

下节课我们会继续介绍能否消除一个bit

Approach the limit

将n个随机变量一同统一处理

Encode n symbols X_1, X_2, \dots, X_n on X together, where X_i 's are i.i.d $\sim p(x)$ Denote the alphabet by \mathcal{X}^n 整体字母表, the expected codeword length by L_n , the length of codeword associated with (x_1, x_2, \dots, x_n) by $l(x_1, x_2, \dots, x_n)$

$$L_n = \frac{1}{n} \sum p(x_1, x_2, \dots, x_n) l(x_1, x_2, \dots, x_n) = \frac{1}{n} El(X_1, X_2, \dots, X_n)$$

我们计算的是per symbol,所以记得除以n

Treat X_1, X_2, \dots, X_n as a whole and apply the lower bound aforementioned

$$H(X_1, X_2, \dots, X_n) \leq El(X_1, X_2, \dots, X_n) < H(X_1, X_2, \dots, X_n) + 1$$

since X_i 's are i.i.d, $H(X_1, X_2, \dots, X_n) = nH(X)$

$$H(X) \leq L_n \leq H(X) + \frac{1}{n}$$

定理: 区块编码

(Theorem.) The minimum expected codeword length per symbol statisfies

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq L^* < \frac{H(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}$$

Moreover, if X_1, X_2, \dots, X_n is a stationary stochastic process,

→ 11

当n很大时, 期望以熵率为极限, 通过这种方式, 我们可以通过逼近的方式把+1bit去掉, 但坏处是字母表太大了 $|\mathcal{X}|^n$ 码制数量是指数级的. 但至少, 系统的熵率是编码问题的极限.

Wrong Code

What happens to the expected description length if the code is designed for the wrong distribution $(q(x))$. For example, the wrong distribution may be the best estimate that we can make of the unknown true distribution.

Recall 相对熵衡量随机变量之间的距离.

(Wrong code) The expected length under $p(x)$ of the code assignment $l(x) = \log \frac{1}{q(x)}$ satisfies

$$H(p) + D(p||q) \leq E_p l(x) < H(p) + D(p||q) + 1$$

$D(p||q)$ 是我们估计产生偏差的惩罚项, 这是难以避免的, 但我们可以用数值量化.

$$\begin{aligned}
El(x) &= \sum_x p(x) \left[\log \frac{1}{q(x)} \right] \\
&< \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\
&= D(p||q) + H(p) + 1
\end{aligned}$$

Kraft Inequality For Uniquely Decodable Codes

我们进一步说明, **Kraft**不等式可以描述任意唯一可解码的编码方式, 因此我们仅研究前缀码就够了。

对任意可解码的编码方式, **Kraft**都成立。推论, 对任意满足的 l_i , 如果没有其他特殊要求, 前缀码就够了。

(McMillan) The codeword lengths of **any uniquely decodable** D-ary code must satisfy the Kraft inequality $\sum D^{-l_i} \leq 1$

Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.

- Consider $C^k = C(x_1, \dots, x_k)$, the k th extension of the code (i.e., the code formed by the concatenation of k repetitions of the given uniquely decodable code C). 考虑每一种编码方式的k次扩展, 连在一起作为新的码制。
- By the definition of unique decodability, the k th extension of the code is nonsingular. 由于可以解码, 那么k次扩展也是可解码的。
- since there are only D^n different D-ary strings of length n , unique decodability implies that the number of code sequences of length n in the k th extension of the code must be no greater than D^n . 考虑D元字符串中, 最多有 D^n 不同的字符串, 唯一可解码性意味着k次扩展码制的长度不能超过 D^n

以上三个结论有助于我们进一步的证明。

- Let the codeword lengths of the symbols $x \in X$ be denoted by $l(x)$. For the extension code, the length of the code sequence is $l(x_1, x_2, \dots, x_k) = \sum_{i=1}^k l(x_i)$ 由于扩展是直接相连
- The inequality we wish to prove is $\sum_{x \in X} D^{-l(x)} \leq 1$
- Consider the k th power of this quantity

$$\begin{aligned}
\left(\sum_{x \in X} D^{-l(x)} \right)^k &= \sum_{x_1 \in X} \sum_{x_2 \in X} \dots \sum_{x_k \in X} D^{-l(x_1)} D^{-l(x_2)} \dots D^{-l(x_k)} \\
&= \sum_{x_1, x_2, \dots, x_k \in X^k} D^{-l(x_1)} D^{-l(x_2)} \dots D^{-l(x_k)} \\
&= \sum_{x^k \in X^k} D^{-l(x^k)} = \sum_{m=1}^{kl_{\max}} a(m) D^{-m} \\
&\leq \sum_{m=1}^{kl_{\max}} D^m D^{-m} = kl_{\max}
\end{aligned}$$

第一行，展开，第二行，合并，第三行，对不同的 k ， $l(x_k)$ 可能是相同的，我们合并同类项。

- $a(m)$ 用到了生成函数的一些性质，即长度为 m 的编码方式所对应的数目。概念上类似二项式定理中的组合数。
- kl_{\max} is the maximum codeword length and $a(m)$ is the number of source sequences x^k mapping into codewords of length m
- 由于我们考虑的是 k 次扩展，最长长度是 kl_{\max} ，
- $a(m) \leq D^m$ 是我们之前证明了的结论
- 由于 k 任取，极限可以趋向于1。

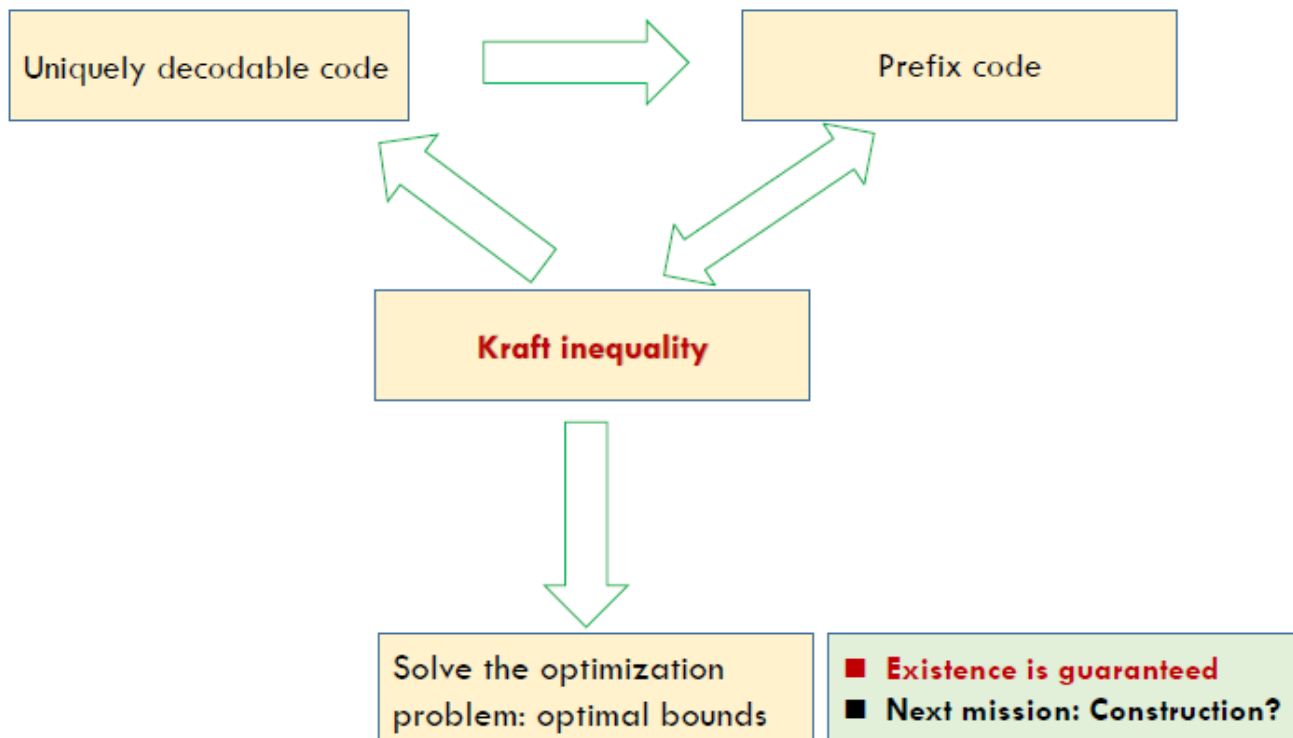
$$\sum_{x \in X} D^{-l(x)} \leq (kl_{\max})^{\frac{1}{k}} \rightarrow 1, \text{ as } k \rightarrow \infty$$

以上是必要性的证明，充分性的证明与唯一可分解码类似。

idea: k 次扩展刚好对应于 $\left(\sum_{x \in X} D^{-l(x)} \right)^k$

因此，Kraft可以约束所有可解码的方式

Summary



本节我们介绍了最优编码的存在性, 下节课我们介绍最优编码的算法.

0325 Data Compression (2)

Huffman coding

Algorithm

本质是一种贪心算法, D元组(字符串可以不仅是二进制)

D-ary Huffman codes (prefix code) for a given distribution:

Each time **combine** D symbols with the **lowest probabilities** into a single source symbol, until there is only one symbol

对三元组, 二元组编码的Huffman编码实现

Codeword	X	Probability
1	1	0.25
2	2	0.25
00	3	0.2
01	4	0.15
02	5	0.15

Codeword Length	Codeword	X	Probability
2	01	1	0.25
2	10	2	0.25
2	11	3	0.2
3	000	4	0.15
3	001	5	0.15

- Huffman Coding is optimal: $\min \sum p_i l_i$
- Huffman coding for weighted codewords w_i

$$p_i \Rightarrow w_i \rightarrow \frac{w_i}{\sum w_i}$$

等价问题: 用贪心策略取带权重期望长度的最小值. Huffman's algorithm for minimizing $\sum w_i l_i$ can be applied to any set of numbers $w_i \geq 0$

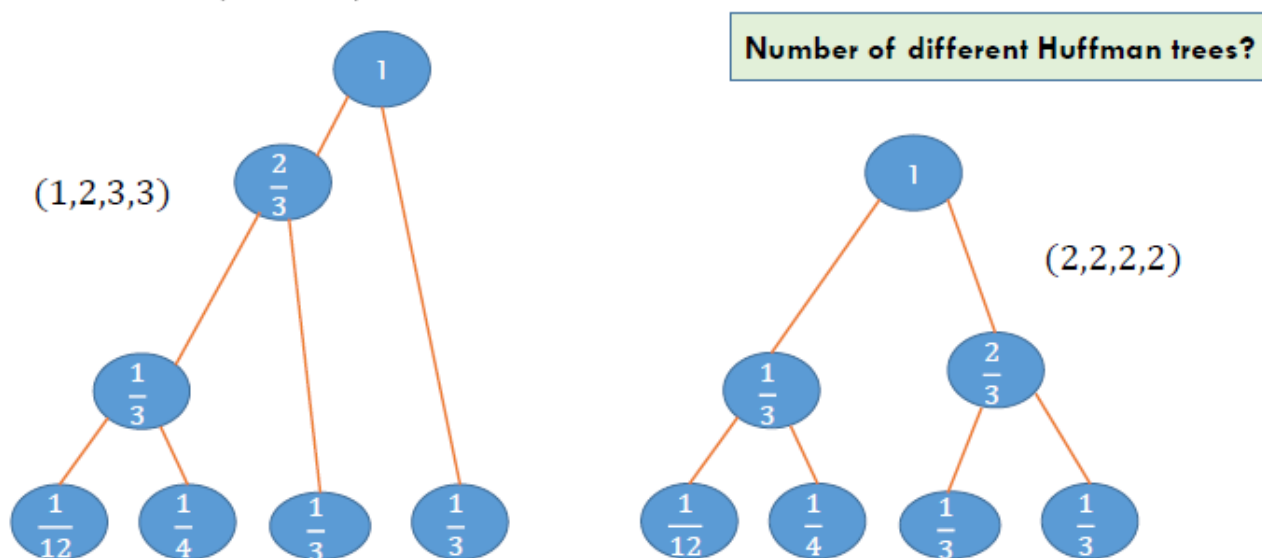
If $D \geq 3$, we may not have a sufficient number of symbols so that we can combine them D at a time. In such a case, we **add dummy symbols to the end of the set of symbols**. The dummy symbols have probability 0 and are inserted to fill the tree.

- 如何计算哑符号的数量: since at each stage of the reduction, the number of symbols is reduced by $D - 1$, we want the total number of symbols to be $1 + k(D - 1)$, where k is the number of merges.
- 本质上原理是一致的 Morse Vs. Huffman Morse code could be regarded as a certain Huffman code when p'_i s are estimated
- Adaptive Huffman coding

Extension

Huffman code is not unique: $l_i, 1 \leq i \leq n$

- Counterexample: $0 \rightarrow 1, 1 \rightarrow 0$ 一个简单的方式, 交换0,1的编码
- For $p(X) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$, both $(2,2,2,2)$ and $(1,2,3,3)$ are optimal Huffman code. 即便是从码长不同的角度看, 我们也可能得到最优情况下不同的码长分布.
- 对number of different huffman trees, 还没有完善的结论



A probability distribution $\Pr(X)$ is called **D -adic** if each of the probabilities

$$\Pr(X = x_i) = D^{-n}$$

for some n

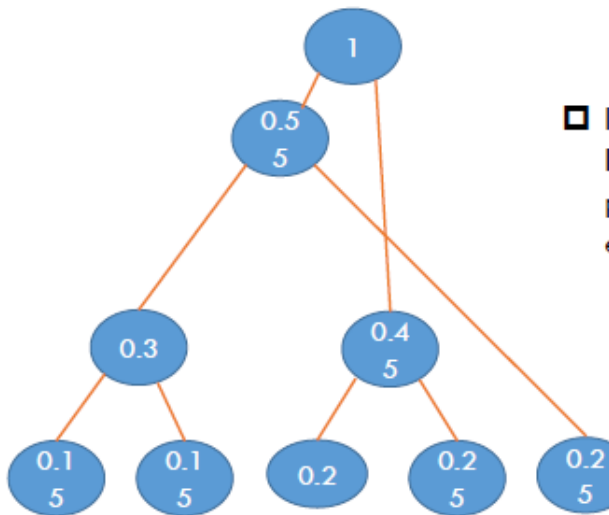
- 因为如果概率分布满足D-adic, 那么可以保证取完对数后可以得到整数值. For a D -adic distribution, the optimal solution in Lagrange is unique: $l_i = \log \frac{1}{p_i} = n_i$ 与上一节的香农码的优化问题对应.
- Huffman Vs. Shannon codes
 - Shannon codes $\left\lceil \log \frac{1}{p_i} \right\rceil$ attain optimality within 1 bit. If the prob. distribution is D adic, **Shannon codes are optimal** 香农码有些情况可以取到最优值
 - **Shannon codes may be much worse when $p_i \rightarrow 0$** : Consider two symbols, one with probability 0.9999 and the other with probability 0.0001. The optimal codeword length is 1 bit for both symbols. The lengths of Shannon codes are 1 and 14. 但Huffman整体情况下比香农码好
- Huffman codes in application
 - JPEG, PNG, ZIP, MP3
 - Cryptography
 - Internet protocol, HTTP header (RFC)

Canonical Codes

我们希望证明哈弗曼编码的最优性,我们先讨论一些有关最优编码(规范化编码)的基本性质

基本假设: Without loss of generality, we will assume that the probability masses are ordered, so that $p_1 \geq p_2 \geq \dots \geq p_m$. Recall that a code is **optimal** if $\sum p_i l_i$ is **minimal**.

Then, For any optimal coding scheme



1. 概率越大,码长越小 The lengths are ordered inversely with the probabilities (i.e., if $p_j > p_k$, then $l_j \leq l_k$).

If $p_j > p_k$, then $l_j \leq l_k$ 反证法: 交换一下就发现码长期望变小了, contradict.

If not, swap the codewords of j and k . Denote the new code by C'_m .

$$\begin{aligned}
L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i \\
&= p_j l_k + p_k l_j - p_j l_j - p_k l_k \\
&= (p_j - p_k)(l_k - l_j) < 0
\end{aligned}$$

2. 最长的两个码制长度一致 The two longest codewords have the same length

If the two longest codewords are not of the same length, one can delete the last bit of the longer one, preserving **the prefix property** and achieving lower expected codeword length. 因为去除最后一位后, 由前缀码的性质, 依然是合法的编码, 但却降低了码长

3. 最长码制一定是兄弟姐妹 Two of the longest codewords differ only in the last bit and correspond to the two least likely symbols.

If there is a maximal length codeword without a sibling(兄弟姐妹), we can delete the last bit of the codeword and still satisfy the prefix property. 如果存在这样的孤儿, 那么根据前缀码的性质, 去掉最后的bit, 依然是合法的编码, 但却降低了码长

Optimality: Strategy

We prove the optimality of Huffman coding **for a binary alphabet**

- When $m = 2$, it is trivial
- For any probability mass function for an alphabet of size m , $p = (p_1, p_2, \dots, p_m)$ with $p_1 \geq p_2 \geq \dots \geq p_m$, we define $p' = (p_1, p_2, \dots, p_{m-2}, p_{m-1} + p_m)$ over an alphabet of size $m - 1$
Now we need to prove the optimality Huffman coding on p by the Huffman code on p' Challenge:
Not so obvious 归纳证明.

Huffman coding for $m - 1$



Huffman coding for m

For any probability mass function for an alphabet of size m , $p = (p_1, p_2, \dots, p_m)$ with $p_1 \geq p_2 \geq \dots \geq p_m$, we define $p' = (p_1, p_2, \dots, p_{m-2}, p_{m-1} + p_m)$ over an alphabet of size $m - 1$.

Let $C_{m-1}^*(p')$ be an optimal code for p' . Let $C_m(p)$ be a code for p

$$C_{m-1}^*(p') \Rightarrow C_m(p)$$

	$C_{m-1}^*(\mathbf{p}')$		$C_m(\mathbf{p})$	
p_1	w'_1	l'_1	$w_1 = w'_1$	$l_1 = l'_1$
p_2	w'_2	l'_2	$w_2 = w'_2$	$l_2 = l'_2$
\vdots	\vdots	\vdots	\vdots	\vdots
p_{m-2}	w'_{m-2}	l'_{m-2}	$w_{m-2} = w'_{m-2}$	$l_{m-2} = l'_{m-2}$
$p_{m-1} + p_m$	w'_{m-1}	l'_{m-1}	$w_{m-1} = w'_{m-1} 0$	$l_{m-1} = l'_{m-1} + 1$
			$w_m = w'_{m-1} 1$	$l_m = l'_{m-1} + 1$

假设左边是最优的编码方案, 我们希望构造出右边的构造方案, 一个简单的操作是把后者的概率分布拆开. 新的码制都是原本码制基础上分别加0和1, 通过这样的操作, 我们可以计算出它们的平均码长符合如下的性质.

Expand an optimal code for \mathbf{p}' to construct a code for \mathbf{p}

$$L(\mathbf{p}) = L^*(\mathbf{p}') + p_{m-1} + p_m$$

(L and L^*) L^* 是已知的最优编码, L 是通过 Huffman 方法构造出的编码

$C_m(p)$ is a Huffman code. Maybe not optimal

Huffman coding for m



Huffman coding for $m - 1$

From the canonical code for \mathbf{p} , we construct a code for \mathbf{p}' by merging the codewords for the two lowest-probability symbols $m - 1$ and m with probabilities p_{m-1} and p_m , which are siblings by the properties of the canonical code. 由此前的规范化编码性质, 我们可以说明最长两者是兄弟节点, 因此这样的操作是合法的. The new code for \mathbf{p}' has average length:

$$\begin{aligned}
 L(\mathbf{p}') &= \sum_{i=1}^{m-2} p_i l_i + p_{m-1} (l_{m-1} - 1) + p_m (l_m - 1) \\
 &= \sum_{i=1}^m p_i l_i - p_{m-1} - p_m \\
 &= L^*(\mathbf{p}) - p_{m-1} - p_m
 \end{aligned}$$

- Expand an optimal code for \mathbf{p}' to construct a code for \mathbf{p}

$$L(\mathbf{p}) = L^*(\mathbf{p}') + p_{m-1} + p_m$$

- Condense an optimal canonical code for \mathbf{p} to construct a code for the reduction \mathbf{p}'

$$L(\mathbf{p}') = L^*(\mathbf{p}) - p_{m-1} - p_m$$

- 两式相加, Together,

$$L(p) + L(p') = L^*(p) + L^*(p')$$

since $L(p) \geq L^*(p), L(p') \geq L^*(p')$ 假设中的最优性

$$L(p) = L^*(p) \quad \text{and} \quad L(p') = L^*(p')$$

- 由此, 我们通过双向构造完成了归纳证明. 即如果 p' 上哈夫曼编码是最优的, 那么 p 上面的最优性也是保证的.
Let the optimal code on p' be a Huffman code, then the expanded code on p is also a Huffman code and it is optimal for p .
- 哈夫曼编码最优性的表述: Huffman coding is optimal; that is, if C^* is a Huffman code and C' is any other uniquely decodable code, $L(C^*) \leq L(C')$

Shannon-Fano-Elias coding

Formulation

Motivation: 虽然在平均码长上不如哈夫曼编码, 但提出了一种全新的构造方式, 且在后续的应用中获得了更高的改进. the codeword lengths $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil \Rightarrow$ Kraft's inequality

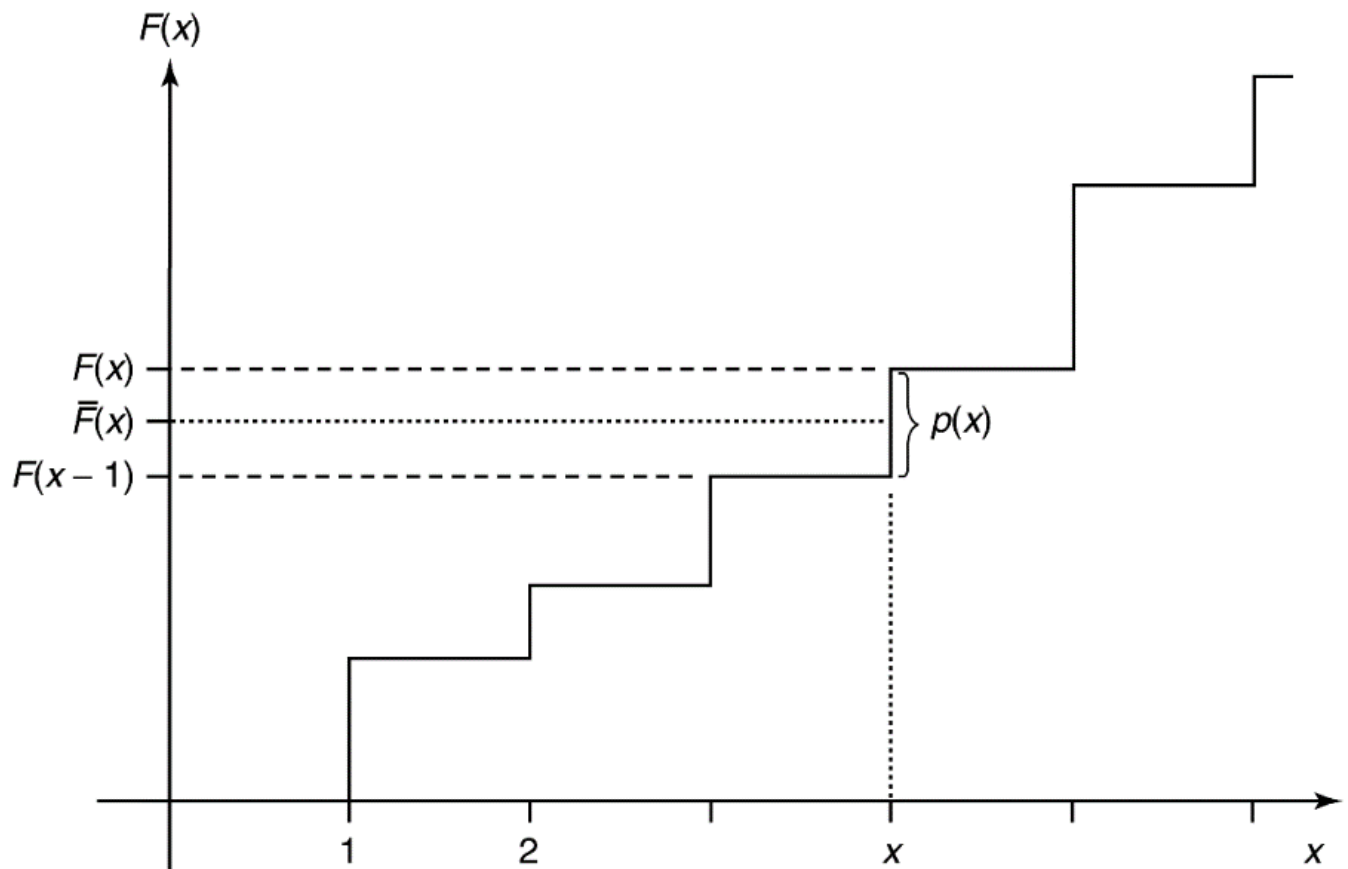
Without loss of generality, we can take $x = \{1, 2, \dots, m\}$. Assume that $p(x) > 0$ for all x . The **cumulative(累积) distribution function** $F(x)$ is defined as $F(x) = \sum_{a \leq x} p(a)$

Consider the modified cumulative distribution function

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2}p(x) = F(x) - \frac{1}{2}p(x)$$

折线段上, 我们取中间的节点. 这样设置的好处:

- The step size is $p(x)$. $\bar{F}(x)$ is the midpoint.
- $\bar{F}(x)$ can determine x . **Thus is a code for x**



- $\bar{F}(x)$ is a real number. Truncate $\bar{F}(x)$ to $l(x)$ bits and use the first $l(x)$ bit of $\bar{F}(x)$ as a code for x . Denote by $\lfloor \bar{F}(x) \rfloor_{l(x)}$
- We have: $\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{l(x)} \leq \frac{1}{2^{l(x)}}$
If $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$

$$\frac{1}{2^{l(x)}} \leq \frac{p(x)}{2} = \bar{F}(x) - \bar{F}(x-1)$$

$\lfloor \bar{F}(x) \rfloor_{l(x)}$ lies within the step corresponding to x . Thus, $l(x)$ bits suffice to describe x . (Prefix-free code)

$$L = \sum p(x)l(x) < H(X) + 2$$

ROADMAP: CDF \rightarrow improved CDF

$$p(x) \Rightarrow F(x) = \sum_{a \leq x} p(a) \Rightarrow \bar{F}(x) = F(x) - \frac{1}{2}p(x) \Rightarrow l(x) + 1 \text{ bits}$$

Example

x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ in Binary	$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$	Codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.5	0.75	0.5	0.10	2	10
3	0.125	0.875	0.8125	0.1101	4	1101
4	0.125	1.0	0.9375	0.1111	4	1111

The average codeword length is 2.75 bits and the entropy is 1.75 bits. The Huffman code for this case achieves the entropy bound.

- Direct application of Shannon-Fano-Elias coding would also need arithmetic **whose precision grows with the block size**, which is not practical when we deal with long blocks.
- Shannon-Fano-Elias \Rightarrow 改进 Arithmetic coding

Optimality

(Optimality) Let $l(x)$ be the codeword lengths associated with the Shannon code, and let $l'(x)$ be the codeword lengths associated with any other uniquely decodable code. Then

$$\Pr(l(x) \geq l'(x) + c) \leq \frac{1}{2^{c-1}}$$

Hence, no other code can do much better than the Shannon code most of the time 注意,是具体码制的随机变量,而不是平均码长, 不与Huffman矛盾.

For example, the probability that $l'(X)$ is 5 or more bits shorter than $l(X)$ is less than $\frac{1}{16}$

$$\begin{aligned}
 \Pr(l(X) \geq l'(X) + c) &= \Pr\left(\left\lceil \log \frac{1}{p(X)} \right\rceil \geq l'(X) + c\right) \\
 &\leq \Pr\left(\log \frac{1}{p(X)} \geq l'(X) + c - 1\right) \\
 &= \Pr\left(p(X) \leq 2^{-l'(X) - c + 1}\right) \\
 &= \sum_{x: p(x) \leq 2^{-l'(x) - c + 1}} p(x) \\
 &\leq \sum_{x: p(x) \leq 2^{-l'(x) - c + 1}} 2^{-l'(x) - (c-1)} \\
 &\leq \sum_x 2^{-l'(x)} 2^{-(c-1)} \\
 &\leq 2^{-(c-1)}
 \end{aligned}$$

since $\sum 2^{-l'(x)} \leq 1$ by the Kraft inequality.

Hence, no other code can do much better than the Shannon code most of the time. We now strengthen this result. In a game-theoretic setting, one would like to ensure that $l(x) < l'(x)$ more often than $l(x) > l'(x)$. The fact that $l(x) \leq l'(x) + 1$ with probability $\geq \frac{1}{2}$ does not ensure this. We now show that even under this stricter criterion, Shannon coding is optimal. Recall that the probability mass function $p(x)$ is dyadic if $\log \frac{1}{p(x)}$ is an integer for all x .

0330 Data Compression (3)

Random Variable Generation

Introduction

We are given a sequence of fair coin tosses Z_1, Z_2, \dots , and we wish to generate X on $X = \{1, 2, \dots, m\}$ with probability mass function $\mathbf{p} = (p_1, \dots, p_m)$.

直观上, 只要我们抛的够多, 就可以生成 X 的分布.

但我们希望不要抛太多.

Let the random variable T denote the number of coin flips used in the algorithm.

Generate a random variable according to the outcome of fair coin flips:

HHHH, TTTTT, HTHTHT, THTHTH

比如, 我们要构造这样的概率分布, if $X = \{0, 1, 2\}$, $p(X) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, 利用抛硬币的结果, 我们可以设置这样的生成机制.

H : $X = 0$

TH : $X = 1$

TT : $X = 2$

算法优化的目标: How many fair coin flips to generate X ?

Recall: The entropy of X $H(X) = 1.5$

如果一个问题跟信息有关系, 联想到熵

The expected number of coin flips $E(T) = 1.5$

Formulation

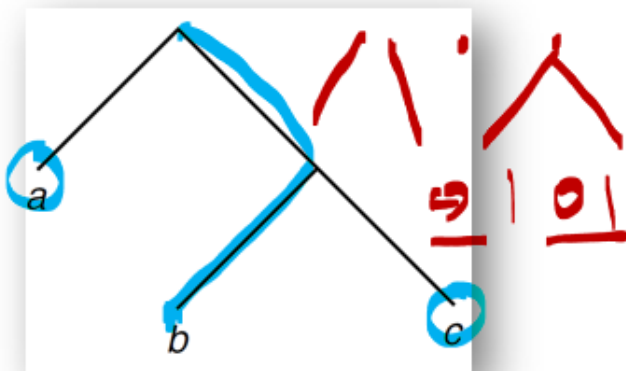
Representation of a generation algorithm

We can describe the algorithm mapping strings of bits Z_1, Z_2, \dots , to possible outcomes X by a **binary tree**.

The **leaves** of the tree are marked by output symbols X , and the path to the leaves is given by the sequence of bits produced by the fair coin

The tree representing the algorithm must satisfy certain properties:

- The tree should be **complete** (i.e., every node is either a leaf or has two descendants in the tree). The tree may be infinite, as we will see in some examples.
- The probability of a leaf **at depth k is 2^{-k}** . **Many leaves may be labeled with the same output symbol**
不同的叶子节点可能对应同一个随机变量 - the total probability of all these leaves should equal the desired probability of the output symbol.
- The expected number of fair bits ET required to generate X is equal to the expected depth of this tree.



Tree for generation of the distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$

Intuition: Each coin tossing generates 1 bit.

$$E(T) \geq H(X)$$

Properties

Let y denote the set of leaves of a complete tree. Consider a distribution on the leaves such that the probability of a leaf at depth k on the tree is 2^{-k} . Let Y be a random variable with this distribution. (Lemma). For any complete tree, consider a probability distribution on the leaves such that the probability of a leaf at depth k is 2^{-k} . Then **the expected depth of the tree is equal to the entropy of this distribution** ($H(Y) = ET$).

Proof. trivial.

$$E(T) = \sum_{y \in y} k(y) 2^{-k(y)}$$

The entropy of the distribution of Y is

$$\begin{aligned}
 H(Y) &= - \sum_{y \in \mathcal{Y}} \frac{1}{2^{k(y)}} \log \frac{1}{2^{k(y)}} \\
 &= \sum_{y \in \mathcal{Y}} k(y) 2^{-k(y)}
 \end{aligned}$$

where $k(y)$ denotes the depth of leaf y . 在这里, 暂时不考虑 Y 有相同的情况. Thus,

在 Y 的样本空间和叶子节点一一对应时, $H(Y) = ET$.

(Theorem). 考虑任意生成算法 For any algorithm generating X , the expected number of fair bits used is greater than the entropy $H(X)$, that is, $E(T) \geq H(X)$

- Any algorithm generating X from fair bits can be represented by a complete binary tree. 生成算法其实是给每一个叶子节点编号. Label all the leaves of this tree by distinct symbols $y \in Y = \{1, 2, \dots\}$. 特别的, 考虑 If the tree is **infinite**, the alphabet Y is also **infinite**.
- Now consider the random variable Y defined on the leaves of the tree, such that for any leaf y at depth k , the probability that $Y = y$ is 2^{-k} . The expected depth of this tree is equal to the entropy of Y :

$$ET = H(Y)$$

- Now the random variable X is a **function of Y** (one or more leaves map onto an output symbol), and hence we have $H(X) \leq H(Y) \leq E(T)$

给出了基本下界, 证明是在上一页基本 lemma 上, 将算法的生成描述成了一个函数关系.

Algorithm

(Theorem). 下界是否可以取到 Let the random variable X have a **dyadic distribution**. (每一个概率分布都可以写作 D^{-n}) The optimal algorithm to generate X from fair coin flips requires an expected number of coin tosses precisely equal to the entropy:

$$ET = H(X)$$

- For the constructive part, we use the Huffman code tree for X as the tree to generate the random variable. Each $X = x$ will correspond to a leaf.
- For a dyadic distribution, the Huffman code is the same as the Shannon code and achieves the entropy bound. (Recall Lagrange Formula, 最优解可以取到)

$$-l_i = \log D^{-n_i} = n_i$$

- 又, 每个事件只对应一个节点, For any $x \in X$, the depth of the leaf in the code tree corresponding to x is the length of the corresponding codeword, which is $\log \frac{1}{p(x)}$. Hence, when this code tree is used to generate X , the leaf will have a probability

$$2^{-\log_p(x)} = p(x)$$

- The expected number of coin flips is the expected depth of the tree, which is equal to the entropy (because the distribution is dyadic). Hence, for a dyadic distribution, the optimal generating algorithm achieves

$$-ET = H(X)$$

一般情况？

- If the distribution is not dyadic? In this case we cannot use the same idea, since **the code tree for the Huffman code will generate a dyadic distribution on the leaves, not the distribution** with which we started 样本空间不再与叶子节点一一对应.
- since all the leaves of the tree have probabilities of the form 2^{-k} , it follows that **we should split any probability p_i that is not of this form into atoms of this form**. We can then allot these atoms(原子形式) to leaves on the tree

$$p(x) = \frac{7}{8} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$$

****Finding the binary expansions of the probabilities p_i 's.**** 二进制展开式 Let the binary expansion of the probability p_i be

$$p_i = \sum_{j \geq 1} p_i^{(j)}$$

对应叶子节点 where $p_i^{(j)} = 2^{-j}$ or 0. Then the atoms of the expansion are the

$$\left\{ p_i^{(j)} : i = 1, 2, \dots, m, j \geq 1 \right\}$$

Since $\sum_i p_i = 1$, **the sum of the probabilities of these atoms is 1**. We will allot an atom of probability 2^{-j} to a leaf at depth j on the tree.

The **depths (j) of the atoms satisfy the Kraft inequality**, we can always construct such a tree with all the atoms at the right depths.

Example

Let X have distribution

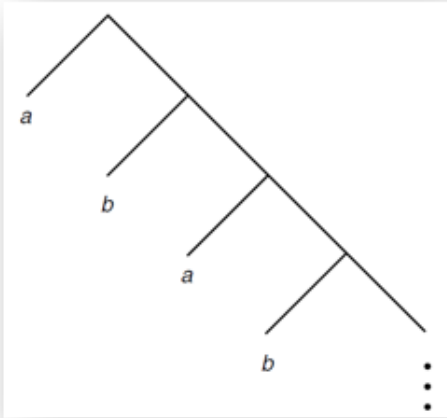
$$X = \begin{cases} a & \text{with prob. } \frac{2}{3} \\ b & \text{with prob. } \frac{1}{3} \end{cases}$$

We find the binary expansions of these probabilities:

$$\begin{aligned} \frac{2}{3} &= 0.10101010 \dots 2 \\ \frac{1}{3} &= 0.01010101 \dots 2 \end{aligned}$$

Hence, the atom for the expansion are:

$$\begin{aligned} \frac{2}{3} &\rightarrow \left(\frac{1}{2}, \frac{1}{8}, \frac{1}{32}, \dots\right) \\ \frac{1}{3} &\rightarrow \left(\frac{1}{4}, \frac{1}{16}, \frac{1}{64}, \dots\right) \end{aligned}$$



Tree to generate a $(\frac{2}{3}, \frac{1}{3})$ distribution

This procedure yields a tree that generates the random variable X . We have argued that this procedure is optimal (gives a tree of minimum expected depth) 因为是最优的, 所以也是完全的.

我们也可以从理论上证明, 期望深度的最小值具有上界.

(Theorem) The expected number of fair bits required by the optimal algorithm to generate a random variable X lies between $H(X)$ and $H(X) + 2$:

$$H(X) \leq ET < H(X) + 2$$

Universal Source Coding

理论应用实际的限制: Challenge: For many practical situations, however, the probability distribution underlying the source may be unknown 信源的分布是未知的

- One possible approach is to wait until we have seen all the data, **estimate the distribution** from the data, use this distribution to construct the best code, and then **go back** to the beginning and compress the data using this code.
 - This two-pass procedure is used in some applications where there is a fairly small amount of data to be compressed.
- In yet other cases, there is no probability distribution underlying the data-all we are given is an individual sequence of outcomes. How well can we compress the sequence?
 - If we do not put any restrictions on the class of algorithms, we get a meaningless answer- there always exists a function that compresses a particular sequence to one bit while leaving every other sequence uncompressed. This function is clearly "overfitted" to the data. 我们需要构造通用的算法. 为此, 压缩算法要有一些限制(counter.e.g.: 1 bit)

Assume we have a random variable X drawn according to a distribution from the family $\{p_\theta\}$, **where the parameter $\theta \in \{1, 2, 3, \dots, m\}$ is unknown** We wish to find an efficient code for this source

Minmax Redundancy

根据上面的问题假设, 我们进行如下分析

- **If we know θ** , we can construct a code with codeword length $l(x) = \log \frac{1}{p_\theta(x)}$ 为了分析的方便, 我们省略向上取整的一步. (+1 bit)

$$\min_{l(x)} E_p[l(X)] = E_p \left[\log \frac{1}{p_\theta(X)} \right] = H(p_\theta)$$

- What happens if **we do not know the true distribution p_θ** , yet wish to code as efficiently as possible? In this case, using a code with codeword lengths $l(x)$ and implied probability $q(x) = 2^{-l(x)}$, we define the redundancy of the code as the difference between the expected length of the code and the lower limit for the expected length: 定义冗余

$$\begin{aligned} R(p_\theta, q) &= E_{p_\theta}[l(x)] - E_{p_\theta} \left[\log \frac{1}{p_\theta(x)} \right] = \sum_x p_\theta(x) \left(l(x) - \log \frac{1}{p_\theta(x)} \right) \\ &= \sum_x p_\theta(x) \left(\log \frac{1}{q(x)} - \log \frac{1}{p_\theta(x)} \right) = D(p_\theta \| q) \end{aligned}$$

- 这样, 我们定义了在不**知道真实分布**的情况下, 某种编码的冗余程度是它和真实分布之间的相对熵
- 目标: 最大冗余最小化, We wish to find a code that does well irrespective of the true distribution p_θ , and thus we define the **minimax redundancy as**

$$R^* = \min_a \max_{ne} R = \min_a \max_{ne} D(p_\theta \| q)$$

Redundancy and Capacity

对任意冗余我们可以构造一个信道. 任意最小最大冗余的计算可以归结为信道容量的计算.

(Theorem) The capacity of a channel $p(x|\theta)$ with rows p_1, p_2, \dots, p_m is given by

$$C = R^* = \min_q \max_\theta D(p_\theta \| q)$$

将信源的所有可能分布写作一个状态转移矩阵, 看作一个信道

How to compute R^* : Take $\{p_\theta : 1 \leq \theta \leq m\}$ as a transition a matrix

$$\theta \rightarrow \begin{bmatrix} \dots p_1(x) \dots \\ \dots p_2(x) \dots \\ \vdots \\ \dots p_\theta(x) \dots \\ \dots p_m(x) \dots \end{bmatrix} \rightarrow X$$

This is a channel $\{\theta, p_\theta(x), x\}$. The capacity of this channel is given by

$$C = \max_{\pi(\theta)} I(\theta; X) = \max_{\pi(\theta)} \sum_{\theta} \pi(\theta) p_\theta(x) \log \frac{p}{q}$$

where $q_\pi(x) = \sum_{\theta} \pi(\theta) p_\theta(x)$

信道和相对熵是等价的

Arithmetic Coding

Recall: Shannon-Fano-Elias Coding: $F(a) = \Pr(x \leq a)$

$$l(x) = \left\lceil \frac{1}{p(x)} \right\rceil + 1$$

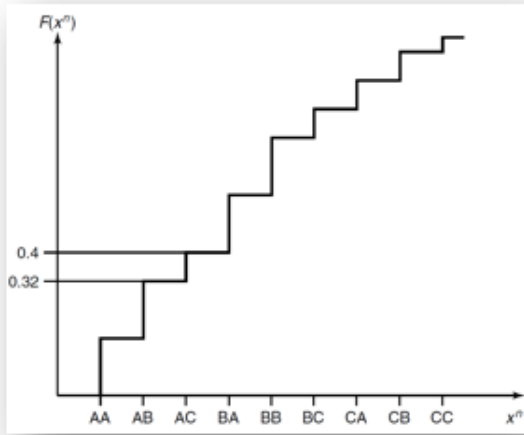
$$H(X) \leq E(l(x)) < H(X) + 2$$

Motivation: using intervals to represent symbols

Consider a random variable X with a ternary alphabet $\{A, B, C\}$, with probabilities 0.4, 0.4, and 0.2 respectively. $F(x) = (0.4, 0.8, 1.0)$.

Let the sequence to be encoded by **ACAA**

- $A \rightarrow [0, 0.4)$
- $AC \rightarrow [0.32, 0.4)$ (scale with ratio $(0.4, 0.8, 1.0)$) 在上一区间的基础上进一步缩放
- $ACA \rightarrow [0.32, 0.352)$
- $ACAA \rightarrow [0.32, 0.3328)$



Combination of x_1x_2 ,
 $x_1x_2 \in \{A, B, C\}$

Lempel-Ziv Coding: Introduction

基于字典的适应性的编码, 动态调整字典大小

Use dictionaries for compression dates back to the invention of the telegraph.

- ".25: Merry Christmas"
- "26: May Heaven's choicest blessings be showered on the newly married couple."

The idea of **adaptive dictionary-based** schemes was not explored until Ziv and Lempel wrote their papers in 1977 and 1978. The two papers describe two distinct versions of the algorithm. We refer to these versions as LZ77 or sliding window Lempel-Ziv and LZ78 or tree-structured Lempel-Ziv.

Gzip, pkzip, compress in unix, GIF

LZ编码有两种方式: 滑动窗口/

Sliding Window

The key idea of the Lempel-Ziv algorithm is to **parse the string into phrases** and to replace phrases by pointers to where the same string has occurred in the past. 重复出现的短语用记号替代它

Sliding Window Lempel-Ziv Algorithm

We assume that we have a string x_1, x_2, \dots to be compressed from a finite alphabet. A **parsing** S of a string $x_1x_2 \dots x_n$ is a division of the string into phrases, separated by commas. Let W be the **length of the window**.

Assume that we have compressed the string until time $i - 1$.

- Then to find the next phrase, find the largest k such that for some $j, i - W \leq j \leq i - 1$, the string of length k starting at x_j is equal to the string (of length k) starting at x_i (i.e., $x_{j+l} = x_{i+l}$ for all

$0 \leq l < k$). The next phrase is then of length k (i.e., $x_i \dots x_{i+k-1}$) and is represented by the pair (P, L) , where P is the location of the beginning of the match and L is the length of the match. If a match is not found in the window, the next character is sent uncompressed.

01010101010101011010101010101101, W = 7
01010101010101011010101010101101
01010101010101011010101010101101
Find the maximum repeated substrng inside W

For example, if $W = 4$ and the string is ABBABBABBBAABABA and the initial window is empty, the string will be parsed as follows: A, B, B, ABBABB, BA, A, BA, BA, which is represented by the sequence of "pointers": $(0, A), (0, B), (1, 1, 1), (1, 3, 6), (1, 4, 2), (1, 1, 1), (1, 3, 2), (1, 2, 2)$, where the flag bit is 0 if there is no match and 1 if there is a match, and the location of the match is measured backward from the end of the window. [In the example, we have represented every match within the window using the (P, L) pair; however, it might be more efficient to represent short matches as uncompressed characters. See Problem 13.8 for details.]

We can view this algorithm as using a dictionary that consists of all substrings of the string in the window and of all single characters. The algorithm finds the longest match within the dictionary and sends a pointer to that match. We later show that a simple variation on this version of LZ77 is asymptotically optimal. Most practical implementations of LZ77, such as gzip and pkzip, are also based on this version of LZ77.

Tree-Structure

思想一致, 解析标准改变了:
each phrase is the shortest phrase not seen earlier.

我们用搜索树为字典建模.

ABBABBABBBAABABAA

	ABBABBABBBAABABAA
A	BBABBABBBAABABAA
A, B	BABBABBBAABABAA
A, B, BA	BBABBBAABABAA
A, B, BA, BB	ABBBAAABABAA
A, B, BA, BB, AB	BBAABABAA
A, B, BA, BB, AB, BBA	ABABAA
A, B, BA, BB, AB, BBA, ABA	BAA
A, B, BA, BB, AB, BBA, ABA, BAA	

"Trie" in data structure.

Since this is the shortest such string, all its prefixes must have occurred earlier. (Thus, we can build up a tree of these phrases.) In particular, the string consisting of all but the last bim of this string must have occurred earlier. We code this phrase by giving the location of the prefix and the value of the last symbol. Thus, the string above would be represented as $(0, A), (0, B), (2, A), (2, B), (1, B), (4, A), (5, A), (3, A)$

0401 Channel Capacity (1)

Noise cannot be eliminated from our life. We should learn how to cope with it.

Noise in Information Transmission

When you send your friend a message via Email/QQ/wechat, you might experience the following failures due to current network environment. 将信息传递中的噪音干扰问题用数学建模

For each task, the message is M with alphabet \mathcal{M} 将信息用字母表表示

How to model the end-to-end pipeline between the sender and the receiver

- The input is X with alphabet \mathcal{X} , the output is Y with alphabet \mathcal{Y} . \mathcal{X} and \mathcal{Y} may be disjoint
- The change from $X \rightarrow Y$ can be modeled as a transition matrix between X and Y

$$p(Y|X)$$

The channel is just like a phone. Each time, you could use it to make a call (M)

The message may be too large to send in just one use of the channel. Thus

$$M \rightarrow X_1, \dots, X_n$$

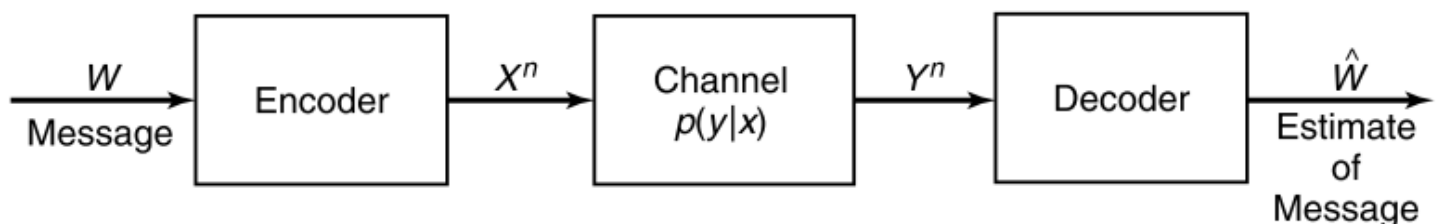
That is, the channel is used n times and we use a random process $\{X_i\}$ to denote it.

Does $p(Y|X)$ remain the same for each X_i^2 . Or we need to define $p_i(Y|X)$ for X_i

无论背景噪声如何变化, 我们都定义成状态转移矩阵

Discrete Memoryless Channel

离散无记忆信道(DMC)



Discrete memoryless channel

A discrete channel is a system consisting of an input alphabet \mathcal{X} and output alphabet \mathcal{Y} and a probability transition matrix $p(y|x)$ that expresses the probability of observing the output symbol y given that we send the symbol x

根据我们的实际需求, 我们要求 $\Pr(W \neq \hat{W}) \rightarrow 0$

The channel is said to be **memoryless** 即第二次与第一次无影响 if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs. (Each time, it is a new channel)

ParseError: KaTeX parse error: Expected '}', got 'EOF' at end of input: ...x), \mathcal{Y}: When you try to send x , with probability $p(y|x)$, the receiver will get y DMC被定义为一个三元组

Channel Capacity

We define the "information" channel capacity of a discrete memoryless channel as

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

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

通常, $p(x)$ 的实际意义是编码方式, 因此我们实际上是在对给定的信道进行优化, 得到最好的编码目标

- $C \geq 0$ since $I(X; Y) \geq 0$
- $C \leq \log |X|$ since $C = \max I(X; Y) \leq \max H(X) = \log |X|$
- $C \leq \log |Y|$ for the same reason
- $I(X; Y)$ is a continuous function of $p(x)$
- $I(X; Y)$ is a concave function of $p(x)$ (recall 信息熵)
 - since $I(X; Y)$ is a concave function over a closed convex set, a local maximum is a global maximum
 - 因此, 在这个意义上(闭区间凹函数): $\sup I(X; Y) = \max I(X; Y)$

" $C = I(X; Y)$ " the most important formula in information age

Properties Of Channel Capacity

General strategy to calculate C :

- $I(X; Y) = H(Y) - H(Y|X)$
 - Estimate $H(Y|X) = \sum_x H(Y|X = x)p(x)$ by the given transition probability matrix
 - Estimate $H(Y)$
- In very few situations, $I(X; Y) = H(X) - H(X|Y)$
 - Estimate $H(X|Y)$ by the given conditions in the problem
 - Estimate $H(X)$

In general, we do not have a closed form expression (显式表达式) for channel capacity except for some special $p(y|x)$

Examples

Noiseless Binary Channel

Suppose that we have a channel whose the binary input is reproduced exactly at the output
In this case, any transmitted bit is received without error

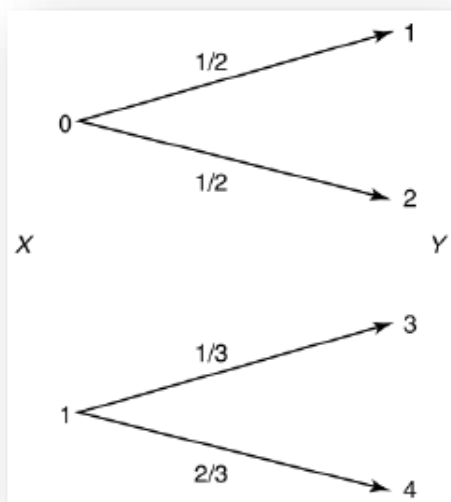


$$C = \max I(X; Y) = \max I(X; X) = \max H(X) \leq 1$$

which is achieved by using $p(x) = (\frac{1}{2}, \frac{1}{2})$

Noisy Channel with Nonoverlapping Outputs

This channel has two possible outputs corresponding to each of the two inputs. 噪音没有使信号叠加在一起
The channel appears to be noisy, but really is not.



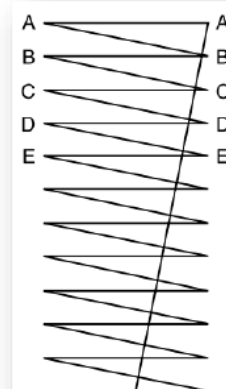
Y can determine X : X is a function of Y

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

Noisy Typewriter



The channel input is either received **unchanged** at the output with probability $\frac{1}{2}$ or is transformed into the next letter with probability $\frac{1}{2}$.



The transition matrix: For each $x \in \{A, B, \dots, Z\}$

$$p(x|x) = \frac{1}{2}, \quad p(x+1|x) = \frac{1}{2}$$

The channel looks symmetric (由对称性)

$$H(Y|X = x) = 1$$

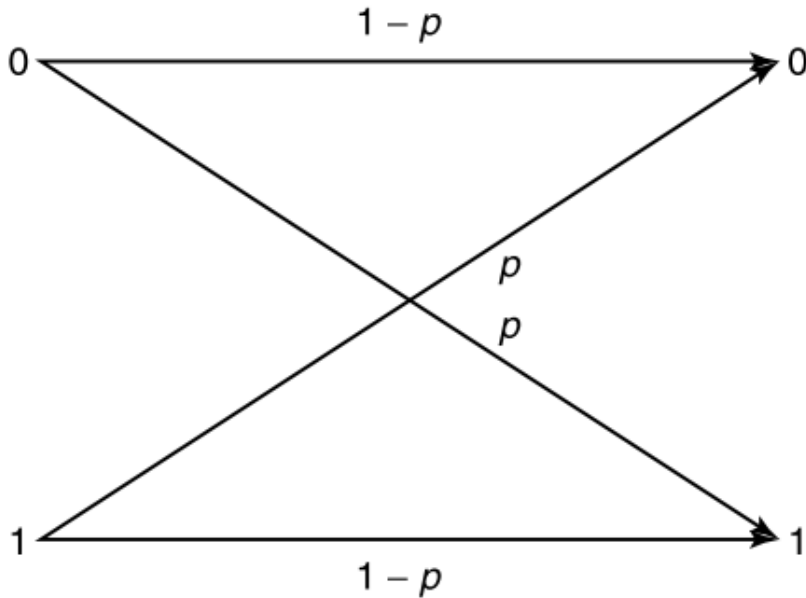
$$H(Y|X) = \sum p(x)H(Y|X = x) = 1$$

The capacity

$$\begin{aligned}
 C &= \max I(X; Y) \\
 &= \max (H(Y) - H(Y|X)) = \max H(Y) - 1 = \log 26 - 1 = \log 13 \\
 p(x) &= \frac{1}{26}
 \end{aligned}$$

Example: Binary Symmetric Channel

错误率为 p



由信道的特殊性, 我们可以用单独的随机变量表示噪声, 吧噪声的影响建模为数学取模过程.

$$\begin{aligned}
 X, Y, Z &\in \{0, 1\} \\
 \Pr(Z = 0) &= 1 - p \\
 Y &= X + Z \pmod{2} \\
 H(Y|X = x) &= H(p)
 \end{aligned}$$

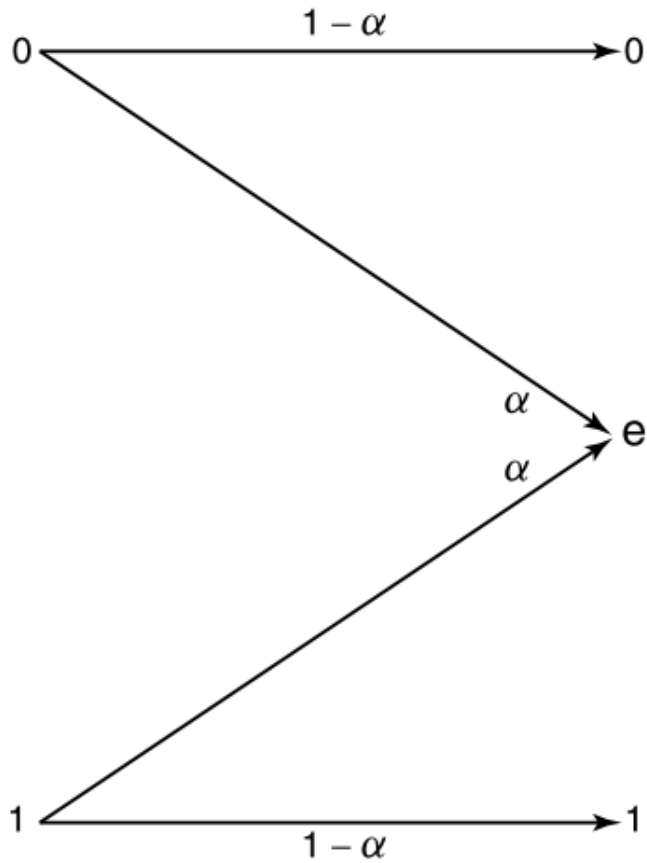
$$\begin{aligned}
 C &= \max I(X; Y) \\
 &= \max H(Y) - H(Y|X) \\
 &= \max H(Y) - \sum p(x) H(Y|X = x) \\
 &= \max H(Y) - \sum p(x) H(p) \\
 &= \max H(Y) - H(p) \\
 &\leq 1 - H(p) \\
 C &= 1 - H(p)
 \end{aligned}$$

BSC is the simplest model of a channel with errors, yet it captures most of the complexity of the general problem 记住 $C = 1 - H(p)$ 的结论

Example: Binary Erasure Channel

The analog of the binary symmetric channel in which some bits are lost (rather than corrupted) is the binary erasure channel. In this channel, a fraction α of the bits are erased.

The receiver knows which bits have been erased. The binary erasure channel has two inputs and three outputs



$$H(Y|X = x) = H(\alpha)$$

$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} H(Y) - H(\alpha) \end{aligned}$$

By letting $\Pr(X = 1) = \pi$

$$\begin{aligned} H(Y) &= H((1 - \pi)(1 - \alpha), \alpha, \pi(1 - \alpha)) \\ &= H(\alpha) + (1 - \alpha)H(\pi) \end{aligned}$$

$$\begin{aligned} C &= \max_{p(x)} H(Y) - H(\alpha) = \max_{\pi} ((1 - \alpha)H(\pi) + \\ &H(\alpha) - H(\alpha)) = \max_{\pi} (1 - \alpha)H(\pi) = 1 - \alpha \end{aligned}$$

注意, 这里不能将 $H(Y)$ 看作均匀分布, 否则取不到.

Symmetric Channel

考虑一般性质

A channel is said to be **symmetric** if the rows of the channel transition matrix $p(y|x)$ are permutations of each other and the columns are permutations of each other. 状态矩阵的每一行和每一列都是其他行/列的排列. (如左例) A channel is said to be **weakly symmetric** if every row of the transition matrix $p(\cdot|x)$ is a permutation (如右例)

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, \quad p(y|x) = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}$$

Letting \mathbf{r} be a row of the transition matrix, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\mathbf{r}) && \text{行对称性} \\ &\leq \log |\mathcal{Y}| - H(\mathbf{r}) && \text{列对称性} \end{aligned}$$

When $p(x) = \frac{1}{|x|}$

$$C = \log |y| - H(\mathbf{r})$$

Computation Of Channel Capacity

信道容量公式的算法(ref

cover

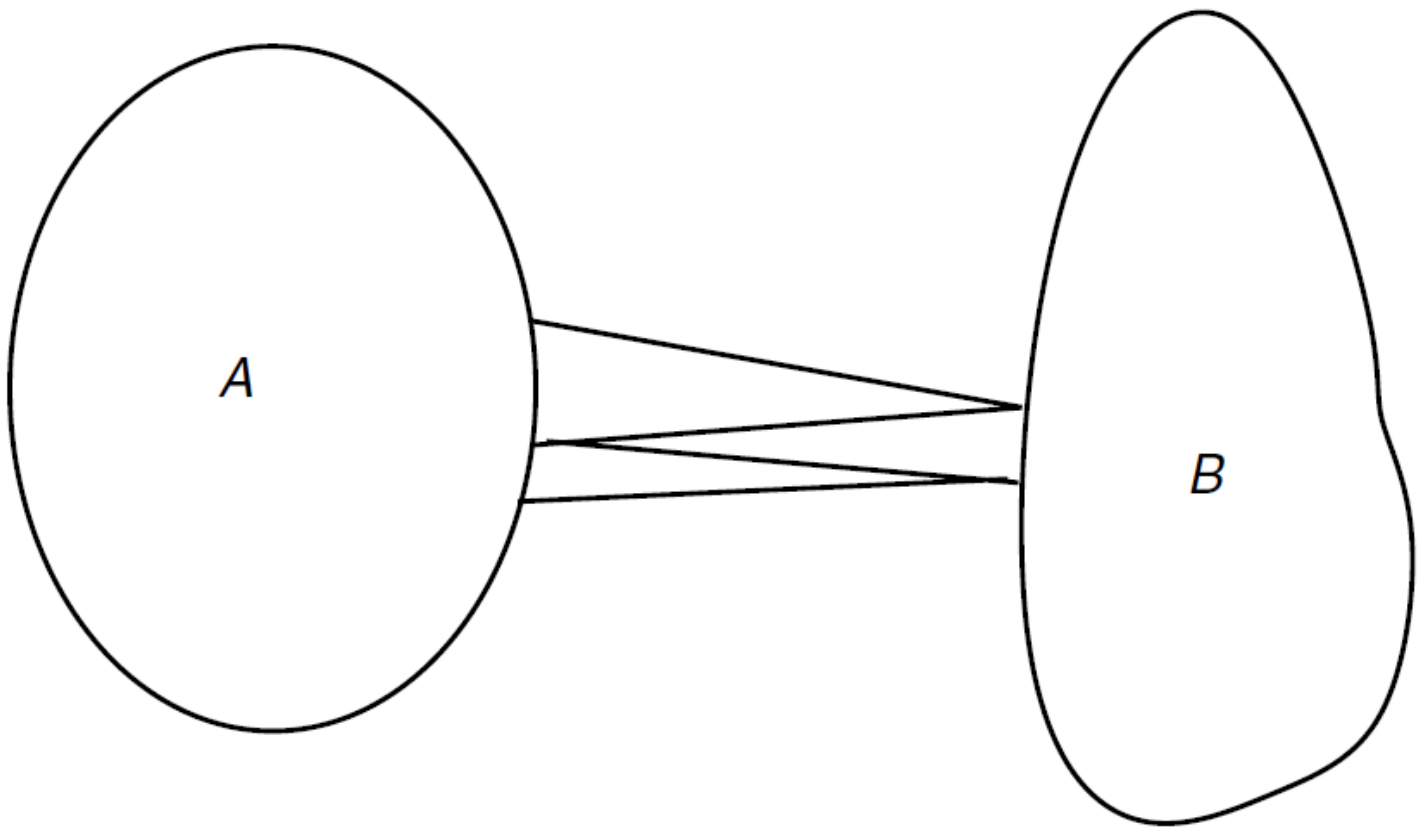
Ch10.8)

优化目标, 两个凸集间最近的点对.

Given two convex sets A and B in \mathcal{R}^n , we would like to find the minimum distance between them:

$$d_{\min} = \min_{a \in A, b \in B} d(a,b)$$

where $d(a,b)$ is the Euclidean distance between a and b



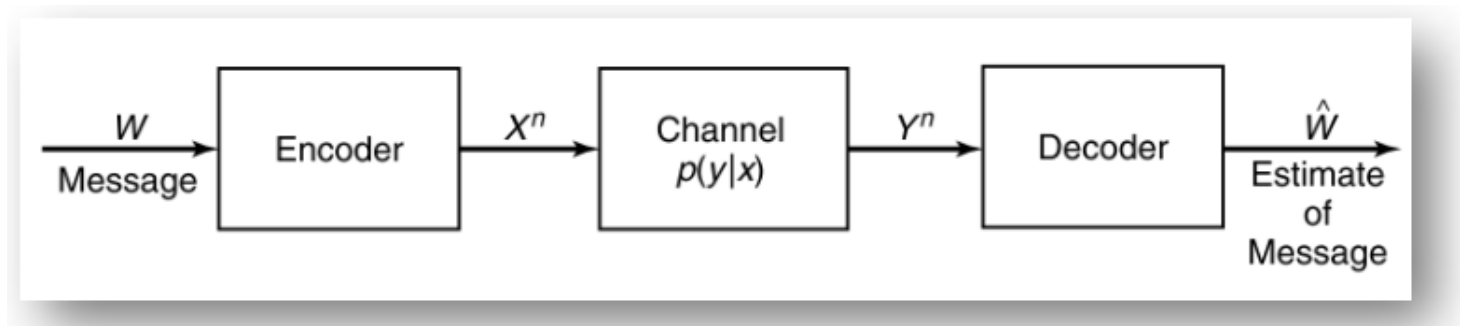
An intuitively obvious algorithm to do this would be to take any point $x \in A$, and find the $y \in B$ that is closest to it. Then fix this y and find the closest point in A . Repeating this process, it is clear that the distance decreases at each stage.

是否收敛?

In particular, if the sets are sets of **probability distributions** and the **distance measure is the relative entropy**, the algorithm does converge to the minimum relative entropy between the two sets of distributions.

0408 Channel Capacity (2)

Recall: Channel Model for Telegraph



- **Codebook** shared by two sides: e.g. a:110, b:111 ...
- 将发电报的过程抽象如下, 由于系统存在随机性, 我们可能解码错误或失败

- $\mathbf{W} \rightarrow \mathbf{X}^n, \mathbf{Y}^n \rightarrow \widehat{\mathbf{W}}$ **could be designed** by us
- $\mathbf{X}^n \rightarrow \mathbf{Y}^n : p(y|x)$ is **out of our control**. (Physical law)
- Aim: a good design with $n(= 40)$ as small as possible

$$\max \frac{H(\mathbf{W})}{n}$$

- 我们希望信息发送的速率(信道容量)越大越好
- 我们不能改变的是信道的状态转移矩阵, 但我们能改变的是编码方式.
- 进一步, 我们发现具有马尔可夫链性质. $\mathbf{W} \rightarrow \mathbf{X}^n, \mathbf{Y}^n \rightarrow \widehat{\mathbf{W}}$

Memory and Feedback

Definition

以磁带为例, 我们翻录磁带的时候, 由于设备原因, 有时翻录能听到原声. 我们把磁带看作一个信道, 反复读写可能会使当前传输的信息受上一次信息保留的影响. 我们在模型中忽略这一因素. (memoryless channel)

- Notation: $y^{k-1} := y_{k-1}, y_{k-2}, \dots, y_1$
- 定义信道的n阶扩展. **The nth extension of the discrete memoryless channel (DMC) is the channel $(x^n, p(y^n|x^n), y^n)$, where**

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$$

When x_k is given, y_k is determined by $p(y|x)$ and is independent of all the generated before time k : $x_1, \dots, x_{k-1}, y_1, \dots, y_{k-1}$

- If the channel is used **without feedback**, i.e., if the input symbols do not depend on the past output symbols, namely, $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$. 定义无反馈, 表明随机变量的生成也是独立于此前的y的.
- When we refer to the discrete memoryless channel, we mean the **discrete memoryless channel without feedback** unless we state explicitly otherwise, 当我们谈论DMC时, 我们是默认无反馈的.

Analysis

Memoryless: $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$

No feedback: $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$

Memoryless + No Feedback \implies

$$\begin{aligned}
p(y^n|x^n) &= p(y^{n-1}|x^n) p(y_n|y^{n-1}, x^n) && \text{条件概率定义式} \\
&= p(y^{n-1}|x^{n-1}, x_n) p(y_n|y^{n-1}, x^n) && x_n \text{ 展开} \\
&= p(y^{n-1}|x^{n-1}) p(y_n|y^{n-1}, x^{n-1}, x_n) && \text{无反馈性质的变形} \\
&= p(y^{n-1}|x^{n-1}) p(y_n|x_n) && \text{无记忆性展开} \\
&= \prod_{i=1}^n p(y_i|x_i) && \text{数学归纳法}
\end{aligned}$$

从信息度量的角度看,

$$\begin{aligned}
p(y^n|x^n) &= \prod_{i=1}^n p(y_i|x_i) \\
H(Y^n|X^n) &= \sum_{i=1}^n H(Y_i|X_i)
\end{aligned}$$

Furthermore, 我们将左侧用链式法则展开, 可以得到更有意思的结论.

Interpretation

reformulate: 信息传输过程

- A message W , drawn from the index set $\{1, 2, \dots, M\}$, results in the signal $X^n(W)$, which is received by the receiver as a random sequence $Y^n \sim p(y^n|x^n)$
- The receiver **guesses** the index W from Y^n : $\widehat{W} = g(Y^n)$. **An error** if $\widehat{W} \neq W$

- A discrete channel, denoted by $(x, p(y|x), y)$
- **Memoryless**: The n th extension of the discrete **memoryless** channel (DMC) is the channel $(x^n, p(y^n|x^n), y^n)$, where

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

- NO Feedback: If the channel is used without feedback [i.e, if the input symbols do not depend on the past output symbols, namely, $p(x_k|x^{k-1}, y_{k-1}) = p(x_k|x^{k-1})$]

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

- Markov chain

$$w \rightarrow X^n \rightarrow Y^n \rightarrow \widehat{w}$$

recall 上节课中, $C = \max I(X; Y)$

接下来我们要证明, 为什么信道容量是最大互信息.

Channel Model

下面我们定义信道模型中的编码

Code

- A code consists of **the message set** \mathcal{M} , an **encoder** and a **decoder**
- Encoder: The channel is used n times to send a symbol $w \in \mathcal{M}$
 - An encoder is a function f such that $f(w) : \mathcal{M} \rightarrow \mathcal{X}^n$ (**one-to-one**) 要求编码/解码函数一一对应
 - f yields a distribution on $x^n = \xrightarrow{\text{DMC}}$ a distribution on x 一般情况下, 我们不能认为定义在 \mathcal{X}^n 上的分布于 \mathcal{X} 分布式等价的, 但在DMC中我们可以通过数学证明说明这一点.
 - The encoding rule $f(w) = x^n \in \mathcal{X}^n$ generates a **codebook** 码本
 - The codebook is **shared** between the sender and the receiver
 - When f is given, a random variable X^n was also defined. 当我们定义了编码器, 我们也就定义了一个随机变量 X^n
- Decoder received $y^n \sim p(y^n|x^n) = \prod p(y_n|x_n)$
 - The decoder need to **guess** the possible x^n by y^n in some genius manner
 - By the codebook $f^{-1}(x^n) = w$. \hat{w} could be recovered by decoder. Error if $\hat{w} \neq w$

综合以上分析,

An (M, n) code for the channel $(X, p(y|x), y)$ consists of the following:

1. **An index set** $\{1, 2, \dots, M\}$: 所有可能信息的集合, 进行索引(如hello-1, world-2, ...)
2. **An encoding function** $X^n : \{1, 2, \dots, M\} \rightarrow \mathcal{X}^n$, yielding **codewords** $x^n(1), x^n(2), \dots, x^n(M)$. The set of codewords is called the **codebook**
3. **A decoding function**

$$g : \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

which is a deterministic rule that assigns a guess to each possible received vector.

$$f : \mathcal{M} \rightarrow \mathcal{X}^n \text{ and } g : \mathcal{Y}^n \rightarrow \mathcal{M}$$
$$W \rightarrow X^n \rightarrow Y^n \rightarrow \widehat{W}$$

w 的维数(M)一般比 x 多, 所以需要利用信道 n 次. 由于我们考虑平均意义, 所以我们认为codewords都等长. 否则编码与解码会很麻烦.

recall 码率 $r = \frac{\log |M|}{n}$.

M 可以看做一个随机过程.

Probability of Error

- Definition (Conditional probability of error) 条件错误概率 Let

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

be the conditional probability of error given that index i was sent, where $I(\cdot)$ is the indicator function. $x^n(i)$ 表示信息*i*对应的码制. 根据定义式,按照所有*y*的可能性展开, 我们可以得到计算式.

$$I(x \neq y) = 0, \quad I(x = y) = 1$$

- Maximal probability of error: 最大错误概率

$$\lambda^{(n)} = \max_{i \in \{1, 2, \dots, M\}} \lambda_i$$

- The (arithmetic) average probability of error

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

我们有

1. 平均错误概率不大于最大错误概率, $P_e^{(n)} \leq \lambda^{(n)}$
2. If M is uniformly distributed,

$$\Pr(W \neq g(Y^n)) = \sum_{i=1}^m \Pr(X^n = x^n(i)) \Pr(g(Y^n) \neq i | X^n = x^n(i)) = P_e^{(n)}$$

Rate and Capacity

- The **rate** R of (M, n) code is

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

因为信息*M*一定能用长度为 $\log M$ 的码进行编码, 可以通过*n*次传输.

- A rate R is said to be **achievable** if there exists a sequence of $(2^{nR}, n)$ codes such that **the maximal probability of error $\lambda(n)$ tends to 0** as $n \rightarrow \infty$

码率可取, if 我们能找到一个编码, 使**最大错误概率**,在码长很长时可以趋向于0.

- The capacity of a channel is the supremum of all achievable rates.

信道容量: 所有可取码率的上确界.

(Channel coding theorem) For a **discrete memoryless channel**, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with

maximum probability of error $\lambda^{(n)} \rightarrow 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \rightarrow 0$ must have $R \leq C$

即,我们要证明 $C = \max_{p(x)} I(X; Y)$

就要证明两点

1. Achievability

For any $r < C$, there exists an $(2^{nr}, n)$ code

2. Converse

For any $r > c, \lambda_e > 0$

Joint Typicality

定义联合典型集.

Roughly speaking, we decode a channel output Y^n as the i th index if the codeword $X^n(i)$ is "jointly typical" with the received signal Y^n

现在我们有两组随机变量, The set $A_\epsilon^{(n)}$ of jointly typical sequences 联合典型集序列 $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ -close to the true entropies:

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in X^n \times Y^n : \begin{aligned} & \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \\ & \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon \\ & \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \end{aligned}\}$$

我们要求, 典型集元素要求满足单个元素在典型集中, 合起来看也要在典型集不等式中.

注意, 一般情况下, $X^n \in A_\epsilon^{(n)}, Y^n \in A_\epsilon^{(n)}$ cannot imply $(X^n, Y^n) \in A_\epsilon^{(n)}$

有了典型集, 顺理成章我们要证明联合AEP.

- $\Pr \left((X^n, Y^n) \in A_\epsilon^{(n)} \right) \rightarrow 1$ as $n \rightarrow \infty$
- $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$ 证明类似单个随机变量的AEP
- $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)}$ 证明类似单个随机变量的AEP
- If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$(1 - \epsilon)2^{-n(I(X,Y)+3\epsilon)} \leq \Pr \left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)} \right) \leq 2^{-n(I(X,Y)-3\epsilon)}$$

(即对典型集中特殊的 X_n, Y_n , 上下界用互信息衡量, 而不是熵)

我们主要证明第三个性质.

定义展开, 根据典型集大小放缩. 右侧得证

$$\begin{aligned}\Pr\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_{\epsilon}^{(n)}\right) &= \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p\left(x^n\right) p\left(y^n\right) \\ &\leq 2^{n(H(X, Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} = 2^{-n(I(X, Y)+3 \epsilon)}\end{aligned}$$

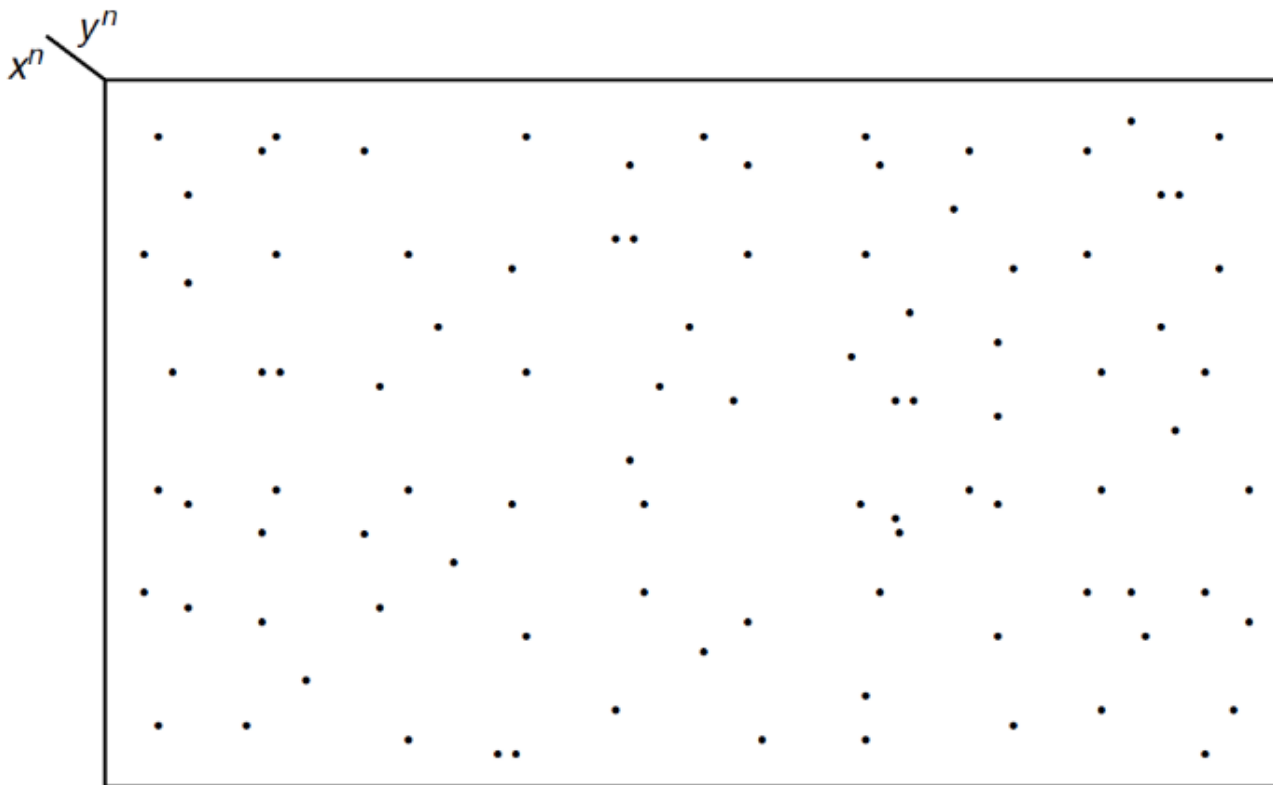
左侧, 先证性质3

$$1-\epsilon \leq \Pr\left(A_{\epsilon}^{(n)}\right)=\sum_{\left(x^n, y^n\right) \in A_{\epsilon}^{(n)}} p\left(x^n, y^n\right) \leq\left|A_{\epsilon}^{(n)}\right| 2^{-n(H(X, Y)-\epsilon)}$$

用类似的方法,

$$\begin{aligned}\Pr\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_{\epsilon}^{(n)}\right) &= \sum_{\left(x^n, y^n\right) \in A_{\epsilon}^{(n)}} p\left(x^n\right) p\left(y^n\right) \\ &\geq(1-\epsilon) 2^{n(H(X, Y)-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(Y)+\epsilon)}=(1-\epsilon) 2^{-n(I(X ; Y)+3 \epsilon)}\end{aligned}$$

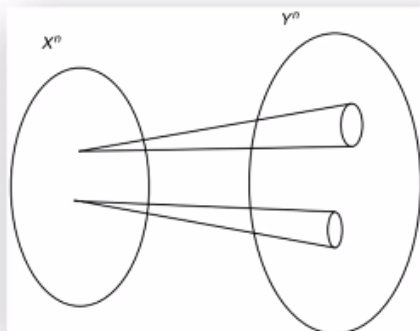
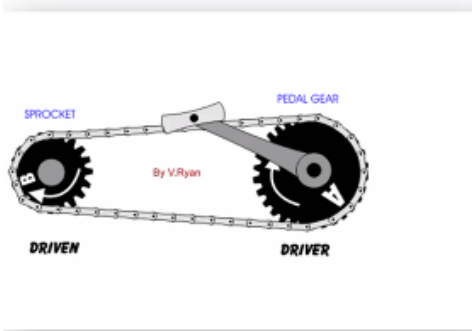
为什么信道编码中会出现互信息?



因为在所有 X^n, Y^n 中, 能够符合典型集性质3的必须用互信息衡量.

$$\frac{2^{nH(X, Y)}}{2^{nH(X)} 2^{nH(Y)}} = 2^{-nI(X ; Y)}$$

Intuition for Channel Capacity

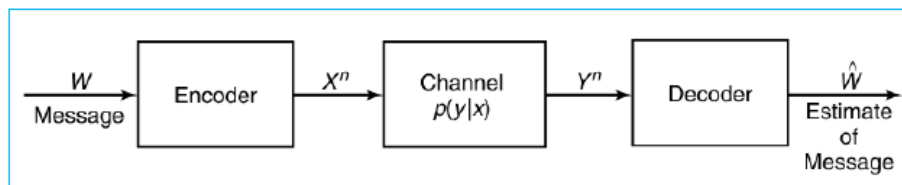


- 手电筒
- 编码在 X^n 上定义了一个随机变量 X^n

- X 与 Y 的关系类似 X^n 驱动了 Y^n , 由于噪声的作用, X^n 的点变成了 Y^n 的区域.
- 那我们最好希望, no two X sequences produce the same Y output sequence.
 - 那比较好的做法就是把 Y^n 的投影分成越来越多的不相交的小集合.
- 从典型集的角度, 当 X^n 给定的情况下, 也就对应的会有 $2^{nH(Y|X)}$ 的 Y^n 个序列可能会构成典型性.
- 由于 Y^n 的典型性, 我们最多能找到互不相交的集合数目就是, $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$
- 因此, 优化目标是最大化 $I(X;Y)$

0413 Channel Capacity (3)

Coverse Proof Special Case: Zero-Error Codes



$$W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W}$$

$X^n \rightarrow Y^n$ is memoryless

我们希望证明, 对于任何合格的编码方案, 码率要小于等于信道容量。

先考虑一个特殊情况: Y^n 可以完美地恢复出 W 。即 $H(W|Y^n) = 0$ 。

The outline of the proof of the converse is most clearly motivated by going through the argument when absolutely no errors are allowed.

$$\begin{aligned}
 nR &= H(W) = H(W|Y^n)_{=0} + I(W; Y^n) \\
 &= I(W; Y^n) \\
 &\leq I(X^n; Y^n) \quad (W \rightarrow X^n \rightarrow Y^n) \\
 &\leq \sum_i I(X_i; Y_i) \\
 &\leq nC \\
 R &\leq C
 \end{aligned}$$

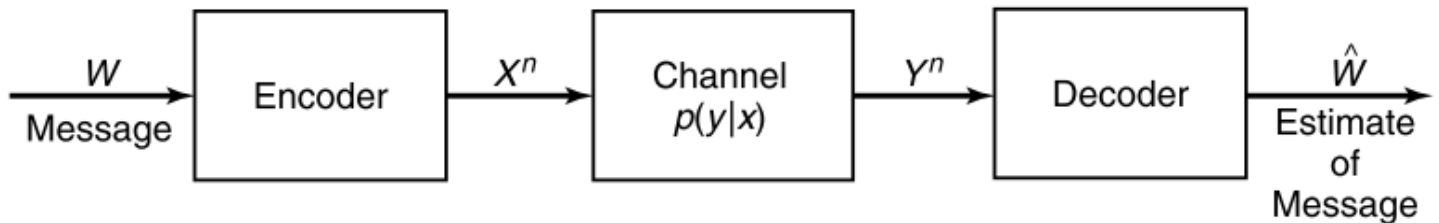
In general, $H(W|Y^n) > 0$: Fano's inequality

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum H(Y_i|X_i) \\ &\leq \sum H(Y_i) - \sum H(Y_i|X_i) = \sum I(X_i; Y_i) \end{aligned}$$

注意，在这个问题（DMC）中， Y_i 与 X_i 相互独立，但不代表 Y_i 之间相互独立。

上面的证明中，我们加强的条件简化了不等式的证明，in general $H(W|Y^n) > 0$ ，我们就要用到Fano's Inequality。

Coverse Proof: Channel Coding Theorem



我们假设错误概率率趋小，即根据fano不等式， $H(W|W) \leq 1 + P_e^{(n)} nR$

$$\begin{aligned} nR &= H(W) \\ &= H(W|\widehat{W}) + I(W; \widehat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W; \widehat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(X^n; Y^n) \quad \text{马尔可夫链性质，数据处理不等式} \\ &\leq 1 + P_e^{(n)} nR + nC \end{aligned}$$

因此我们有

$$R \leq P_e^{(n)} R + \frac{1}{n} + c \rightarrow C$$

反证法，如果 $R > C$ ，则 $P_e^{(n)}$ 不会趋向于0.

$$P_e^{(n)} \geq 1 - \frac{c}{R} - \frac{1}{nR} > 0 \text{ as } R > C$$

接下来我们证明，对任意小于 C 的编码，我们都能找到一种编码（可达性）

Achievability

Code Construction

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

编码的构造：实际上就是随机生成码本矩阵 C 。

行：消息，列：编码。

Fix $p(x)$. Generate $a(2^{nR}, n)$ code at random according to $p(x)$

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

The probability the we generate a particular code C is 对任意编码，生成该编码（码本）的概率是 $\Pr(C) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^n p(x_i(w))$

The code C will be shared both by the sender and receiver, both know $p(y|x)$.

A message W is chosen according to a uniform distribution暂时忽略 nR 不是整数的问题。 $\Pr(W = w) = 2^{-nR}, w = 1, 2, \dots, 2^{nR}$

The w th codeword $X^n(w)$ is sent over the channel The receiver receives a sequence Y^n according to the distribution

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

Joint Decoding

如何解码，我们要用联合AEP解码。

The receiver guess which message was sent. In jointly typical decoding, the receiver declares that the index \widehat{W} was sent if the following conditions are satisfied:

- $(X^n(\widehat{W}), Y^n)$ is jointly typical 存在
- There is no other index $W' \neq W$, such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$ 唯一

If no such \widehat{W} exists or if there is more than one such, an error is declared. (We may assume that the receiver **outputs a dummy index such as 0** in this case.)

Let \mathcal{E} be the event $\{\widehat{W} \neq W\}$

We need to show that

$$\Pr(\mathcal{E}) \rightarrow 0$$

$$\Pr(\mathcal{E}) \rightarrow 0$$

Main idea: If we could prove that for all the codebook (all the possible C), the average $\Pr(\varepsilon) \leq \epsilon_i$ **then the error probability of the best code** (one of C' 's $\leq \epsilon$)

We let W be drawn according to a uniform distribution over $\{1, 2, \dots, 2^{nR}\}$ and use jointly typical decoding $\hat{W}(y^n)$

Let $\mathcal{E} = \{\hat{W}(y^n) \neq W\}$ denote the error event

We will calculate the average probability of error, averaged over all codewords in the codebook, and averaged over all codebooks 每一个码本、每一个码制上的平均错误率。

$$\begin{aligned}\Pr(\varepsilon) &= \sum_c \Pr(C) P_e^{(n)}(C) \quad \text{定义展开} \\ &= \sum_c \Pr(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C) \quad \text{码制均匀分布} \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_c \Pr(C) \lambda_w(C) \quad \text{求和交换位置}\end{aligned}$$

我们分析第二个求和表达式。

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

是在传递信息1的情况下的平均错误概率，也即

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

由对称性，我们以证明信息为1时的结论为例。

$$E_i = \left\{ \left((X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \right), i \in \{1, 2, \dots, 2^{nR}\} \right\}$$

定义集合 E_i ，那么解码错误的概率可以形式化地表达为

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &= P(E_1^c \cup E_2 \cup E_3 \cup \dots \cup E_{2^n R} | W=1) \\ &\leq P(E_1^c | W=1) + \sum_{i=2}^{2^n R} P(E_i | W=1)\end{aligned}$$

要么不在典型集中，要么在其他典型集中。我们直接对集合进行放缩。并 \rightarrow 不相交并。

我们具体分析两个部分有多大。

- 由联合AEP, By Joint AEP, $P(E_1^c | W=1) \rightarrow 0$, and hence $P(E_1^c | W=1) \leq \epsilon$, for n sufficiently large
- For $i \geq 2, (E_i | W=1)$:
 - since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, 源码的生成相互独立
 - so are Y^n and $X^n(i)$.
 - Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &\leq \epsilon + \sum_{i=2}^{2^n R} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR} - 1) 2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}\end{aligned}$$

If n is sufficiently large and $R < I(X;Y) - 3\epsilon$

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &\leq 2\epsilon \\ \Pr(\mathcal{E}) &\leq 2\epsilon\end{aligned}$$

这样，我们就证明了，给定 $p(x)$ 情况下，随机编码的错误概率 $\leq 2\epsilon$ 。这样也就证明了一定存在这样的编码。

注意到 $p(x)$ 的选择是不受限的，因此Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieving capacity. Then

$$\begin{aligned}R &\leq I(X^*;Y) = C \\ \lambda^{(n)} &\leq 4\epsilon\end{aligned}$$

$$\Pr(\mathcal{E}) \rightarrow 0 \Rightarrow \lambda^{(n)} \rightarrow 0$$

从平均概率为0到最大错误概率为0.

在前面的证明中，我们说明了There exists a best codebook C^* such that

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

根据上一节的定义，By the definition of $(n, 2^{nR})$ code, we need to further show that

$$\lambda^{(n)} \rightarrow 0$$

Without loss of generality, assume $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{2^{nR}}$

By $\Pr(\varepsilon|C^*) \leq 2\epsilon$, we have 我们知道前半一半一定足够小

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{2^{nR-1}} \leq 4\epsilon$$

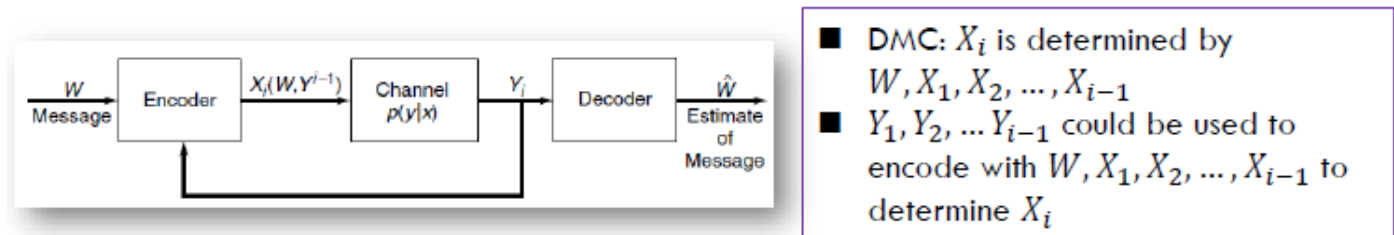
反证法: $\left(\text{Or } \lambda_{2^{nR-1}} > 4\epsilon, \frac{1}{2^{nR}} \sum_{i=1+2^{nR-1}}^{2^{nR}} \lambda_i(C^*) > \frac{1}{2}4\epsilon = 2\epsilon, \text{ contradiction!} \right)$

只需对码本进行微调，就可以达到最大错误概率的降低。把最坏的一半码制直接扔掉。（考虑平均再通过扔掉一半证明最坏，这是一个很general的证明做法，在很多地方都很常见）Further refine the codebook C^*

- Throw away the worst half of the codewords in the best codebook C^*
- The best half of the codewords have a maximal probability of error less than 4ϵ
- If we reindex these codewords, we **have 2^{nR-1} codewords**. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n . 在极限情况下不会对码率产生影响。

(Introduction) Feedback Capacity

Recall, DMC中，收到的信息是什么情况我们是不知情的。信号的发送端与接收端在物理上是隔离开的。



- We assume that **all the received symbols are sent back immediately and noiselessly** to the transmitter, which can then use them to decide **which symbol to send next** 即，在有反馈的情况下，我们有更高的灵活性生成码制。这里我们假设输出的信号马上会发到输入端，输入端可以根据该反馈生成下一个信号。
- We define a $(2^{nR}, n)$ **feedback code** as a sequence of mappings $x_i(W, Y^{i-1})$, where each x_i is a **function only of** the message $W \in 2^{nR}$ and the previous received values, Y_1, Y_2, \dots, Y_{i-1} , and a sequence of decoding functions $g: \mathcal{Y}^n \rightarrow \{1, 2, \dots, 2^{nR}\}$. Thus, 定义解码器没有得到正确结果的概率

$$P_e^{(n)} = \Pr(g(Y^n) \neq W)$$

when W is uniformly distributed over $\{1, 2, \dots, 2^{nR}\}$

我们发现: Feedback capacity

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

Feedback cannot increase capacity. (因为该问题中, 无记忆性仍然是保留的)

TODO:看一下证明

(Introduction) Source-Channel Separation

信源信道分离定理

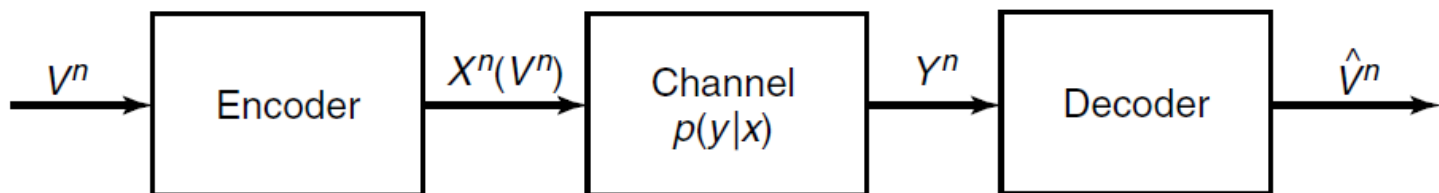
Recall, 我们学习了数据的压缩与数据的传输

- In data compression: $R > H$
- In data transmission: $R < C$

Is the condition $H < C$ sufficient and necessary?

是不是 $H < C$ 就保证了我们可以以很低的错误概率?

YES



Formal Problem

- We want to send the sequence of symbols $V^n = V_1, V_2, \dots, V_n$ over the channel so that the receiver can reconstruct the sequence
- To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel
- The receiver looks at his received sequence Y^n and makes an estimate \hat{V}^n of the sequence V^n that was sent. The receiver makes an error if $V^n \neq \hat{V}^n$. We define the probability of error as

$$\Pr(V^n \neq \hat{V}^n) = \sum_{y^n} \sum_{v^n} p(v^n) p(y^n | x^n(v^n)) I(g(y^n) \neq v^n)$$

Where I is the indicator function and $g(y^n)$ is the decoding function

TODO: Theorem

Theorem (Source-channel coding theorem). If V_1, V_2, \dots, V_n is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source-channel code with probability of error $\Pr(\hat{V}^n \neq V^n) \rightarrow 0$. Conversely, for any stationary stochastic process, if $H(v) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.

Error Correction Code

The object of coding is to introduce redundancy so that even if some of the information is lost or corrupted, it will still be possible to recover the message at the receiver. 在前面的学习中，我们知道，我们可以用冗余抵抗噪声的干扰，下面列出一些冗余手段。

- Repetition code: For example, to send a 1, we send 11111, and to send a 0, we send 00000. The decoding scheme is to take the majority vote. 接收端数0多还是1多，少数服从多数。
- Parity check code: Starting with a block of $n - 1$ information bits, we choose the n th bit so that the parity of the entire block is 0. 奇偶校验位。
- The code does not detect an even number of errors and does not give any information about how to correct the errors that occur.

Hamming Code

RECALL: BSC信道

由大数定律，因为我们会以 p 的概率翻转，如果信息量够大，那么会有大约 np 个bits被修改，我们可以证明， $d(x, y) \leq np$ （曼哈顿距离）。另一种意义上，也就是说，All the points y are within the sphere with center x and radius np

- Decode the codeword by x , then the noisy version y of x stays inside the sphere with center x and radius r
- Sphere packing: the art of error correction code
- https://en.wikipedia.org/wiki/Sphere_packing

0415 Differential Entropy (1)

Differential Entropy

Definition

概率论中的一些概念

- Let X be a random variable with **cumulative distribution function** $F(x) = \Pr(X \leq x)$
- If $F(x)$ is continuous, the random variable is said to be **continuous**.
- Let $f(x) = F'(x)$ when the derivative is defined. If $\int_{-\infty}^{\infty} f(x) = 1$, $f(x)$ is called the **probability density function** for X .
- The set where $f(x) > 0$ is called the **support set** of X .

微分熵的定义

The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

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

where S is the support set of the random variable. The differential entropy is sometimes written as $h(f)$ rather than $h(X)$

平移不改变微分熵.

$h(X + c) = h(X)$ (Translation does not change the differential entropy)

对比离散熵

$$\begin{aligned} p(x) &\Rightarrow f(x) \\ \sum &= \int \\ H(X) &\Rightarrow h(X) \end{aligned}$$

$H(X)$ is always non-negative. $h(X)$ may be negative

Example

Consider a random variable distributed uniformly from 0 to a , then $h(X) = \log a$.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $h(X) = \frac{1}{2} \log 2\pi e \sigma^2$

When X is uniformly distributed in $[0, a]$

$$\begin{aligned} f(x) &= 1/a \\ h(X) &= - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a \end{aligned}$$

发现, 微分熵已经可以小于0了.

When X is Gaussian $\mathcal{N}(\mu, \sigma^2)$, then

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

$$\begin{aligned} h(f(x)) &= - \int f(x) \log f(x) dx \\ &= - \int f(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} + f(x) \left(-\frac{(x-\mu)^2}{2\sigma^2} \right) dx \end{aligned}$$

$$\int f(x) dx = 1 \text{ and } \text{Var}(X) = \int (x-\mu)^2 f(x) dx = \sigma^2$$

用代入方差的方法, 简化运算.

$$h(f(x)) = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} = \frac{1}{2} \log 2\pi e\sigma^2$$

我们用 e 为底, 方便代入概率密度函数.

h(X): infinite information

连续变量不加以任何限制, 它的信息是无穷大的. 微分熵无法再度量系统中的信息量了.

- Differential entropy does not serve as a measure of the average amount of information contained in a continuous random variable.
- In fact, a continuous random variable generally contains an infinite amount of information

我们计算连续型随机变量的离散熵:

Let X be uniformly distributed on $[0, 1)$. Then we can write

$$X = 0.X_1X_2, \dots$$

The dyadic expansion of X , where X'_i 's is a sequence of i.i.d bits. Then

$$\begin{aligned}
 H(X) &= H(X_1, X_2, \dots) \\
 &= \sum_{i=1}^{\infty} H(X_i) \\
 &= \sum_{i=1}^{\infty} 1 \\
 &= \infty
 \end{aligned}$$

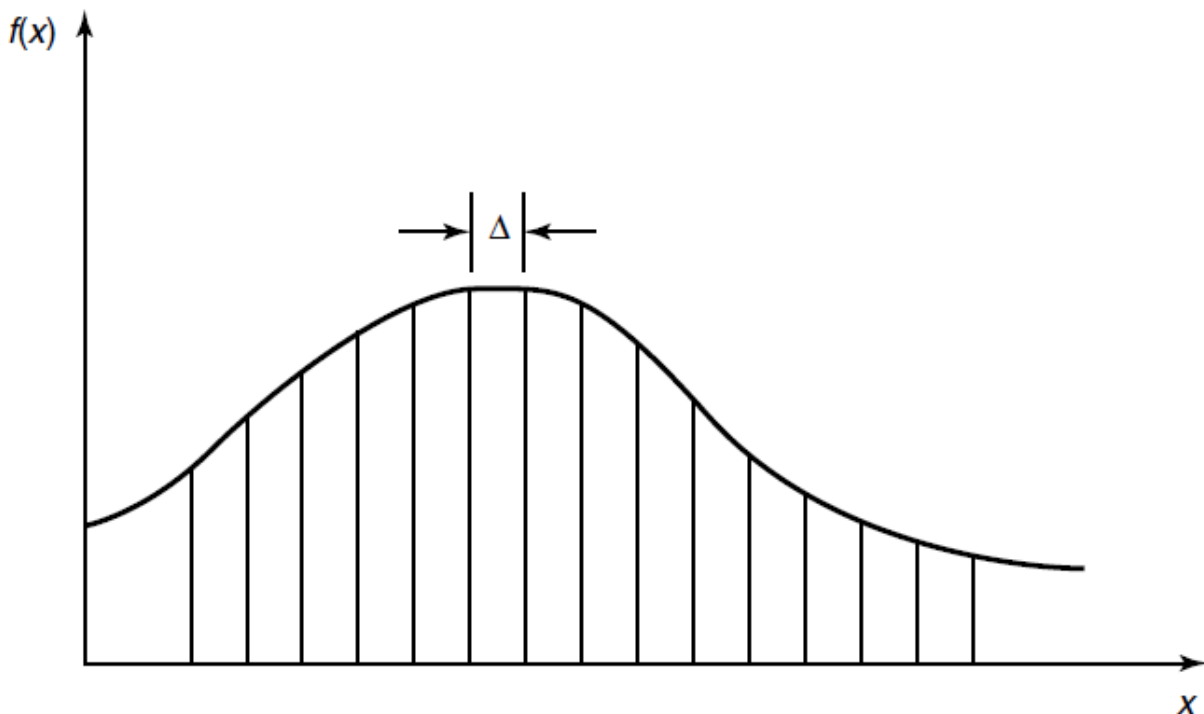
h(aX): Stretching Random Variable

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

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

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

Differential and Discrete Entropy



Suppose that we divide the range of X into bins of length Δ .

By the mean value theorem, there exists a value x_i within each bin such that

$$f(x_i) \Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) dx$$

Consider the quantized random variable X^Δ , which is defined by

ParseError: KaTeX parse error: Can't use function '\$' in math mode at position 17: ...^{\Delta}=x_{\{i\}}\$ if \$i \Delta \dots

基于连续随机变量定义一个切割意义上的离散随机变量. Then the probability that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) dx = f(x_i) \Delta$$

$$H(X^\Delta) = - \sum \Delta f(x_i) \log f(x_i) - \log \Delta$$

结论:

这里说的是 $-\infty + \infty \rightarrow h(f)$.

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \text{ as } \Delta \rightarrow 0$$

AEP For Continuous Random Variable

- AEP for continuous random variables: Let X_1, X_2, \dots, X_n be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow E(-\log f(X)) = h(f)$$

in probability

- For $\epsilon \geq 0$ and any n , we define the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \dots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, x_2, \dots, x_n) - h(X) \right| \leq \epsilon \right\}$$

where $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$

连续情况,考虑元素的体积

The volume of a set $A \subset \mathcal{R}^n$ is defined as

$$\text{Vol}(A) = \int_A dx_1 dx_2 \dots dx_n$$

在这里, $2^{nh(X)}$ is the volume

The typical set $A_\epsilon^{(n)}$ has the following properties:

1. $\Pr \left(A_\epsilon^{(n)} \right) > 1 - \epsilon$ for n sufficiently large.
2. $\text{Vol} \left(A_\epsilon^{(n)} \right) \leq 2^{n(h(X)+\epsilon)}$ for all n
3. $\text{Vol} \left(A_\epsilon^{(n)} \right) \geq (1 - \epsilon)2^{n(h(X)-\epsilon)}$ for n sufficiently large.

Joint and Conditional Differential Entropy

- The differential entropy of a set X_1, X_2, \dots, X_n of random variables with density $f(x_1, x_2, \dots, x_n)$ is defined as

$$h(X_1, X_2, \dots, X_n) = - \int f(x^n) \log f(x^n) dx^n$$

- If X, Y have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$\begin{aligned} h(X|Y) &= - \int f(x, y) \log f(x|y) dx dy \\ h(X|Y) &= h(X, Y) - h(Y) \end{aligned}$$

考虑到,很多情况下,积分是没法求/不存在的,我们在本课程中仅假设积分是存在的. 但是这在研究中是需要特别谨慎的地方.

可以证明条件减少熵, chain rule, 联合熵的结论依然成立

Pf by expectation (虽然定义的方式不同,但期望的写法依然成立)

- $h(X|Y) \leq h(X)$
with equality iff X and Y are independent.
- (Chain rule for differential entropy)

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

- $h(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n h(X_i)$
with equality iff X_1, X_2, \dots, X_n are independent.

Entropy of Multivariate Normal Distribution

Covariance Matrix

- The **covariance** between two random variables X and Y is defined as

$$\text{cov}(X; Y) = E(X - EX)(Y - EY) = E(XY) - (EX)(EY)$$

- For a random vector $X = [X_1, X_2, \dots, X_n]^T$, the **covariance matrix** 相关矩阵 is defined as

$$K_x = E(X - EX)(X - EX)^T = [\text{cov}(X_i; X_j)]$$

and the **correlation matrix** is defined as $\widetilde{K}_X = EXX^T = [EX_iX_j]$

- $K_X = EXX^T - (EX)(EX^T) = \widetilde{K}_X - (EX)(EX^T)$

- A covariance matrix is both symmetric and positive semidefinite. 协方差矩阵半正定
 - The eigenvalues of a positive semidefinite matrix are non-negative. 特征值非负
- 线性变换作用于协方差矩阵和关联矩阵. Let $Y = AX$, where X and Y are column vectors of n random variables and A is an $n \times n$ matrix. Then

$$K_Y = AK_XA^T$$

and

$$\widetilde{K}_Y = A\widetilde{K}_XA^T$$

A set of correlated random variables can be regarded as an orthogonal transformation of a set of uncorrelated random variables. (Ref : Ch. 10.1 Yeung, Information theory and network coding)

Multivariate Normal Distribution

- In probability theory and statistics, the multivariate normal distribution, multivariate Gaussian distribution, or joint normal distribution is a generalization of the onedimensional (univariate) normal distribution to higher dimensions.
- 原始定义: 向量的任意线性组合都是高斯分布 ~ 向量服从多元高斯分布
- More generally, let $\mathcal{N}(\mu, K)$ denote the multivariate Gaussian distribution with mean μ and covariance matrix K , i.e., the joint pdf of the distribution is given by

$$f(x) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)}$$

- One definition is that a random vector is said to be **k-variate normally** distributed if every linear combination of its k components has a univariate normal distribution.

多元高斯分布良好的性质

- In general, random variables may be uncorrelated but statistically dependent.

- 两个事件不相关不代表两个事件独立. 但对多元高斯分布 一定成立. But if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent .
- 两两独立推出联合独立. This implies that any two or more of its components that are pairwise independent are independent.

Entropy

(Entropy of a multivariate normal distribution) Let X_1, X_2, \dots, X_n have a multivariate normal distribution with mean μ and covariance matrix K

$$h(X_1, X_2, \dots, X_n) = h(\mathcal{N}(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K|$$

where $|K|$ denotes the determinant of K .

记住,对于多元高斯分布,微分熵是可以计算出来的, 且和协方差矩阵对数值相关即可.

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$$

证明不作具体要求, 与一元情况类似.

技巧:期望和求和进行交换

$$\begin{aligned} h(f) &= - \int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x} - \mu)^T K^{-1}(\mathbf{x} - \mu) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}} \right] d\mathbf{x} \\ &= \frac{1}{2} E \left[\sum_{i,j} (X_i - \mu_i)(X_j - \mu_j) (K^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_{i,j} E[(X_j - \mu_j)(X_i - \mu_i)] (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_j \sum_i K_{ji} (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_j (K K^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_j I_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \ln(2\pi e)^n |K| \\ &= \frac{1}{2} \log(2\pi e)^n |K| \end{aligned}$$

Relative Entropy and Mutual Information

- The **relative entropy (or Kullback-Leibler distance)** $D(f\|g)$ between two densities f and g is defined by

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

- The **mutual information** $I(X; Y)$ between two random variables with joint density $f(x, y)$ is defined as

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy$$

有关相对熵和互信息的性质:

- $I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y)$,
 $I(X; Y) = D(f(x, y)\|f(x)f(y))$
信息图依然可用,但对非负部分需要考虑.
- $D(f\|g) \geq 0$ with equality iff $f = g$ almost everywhere (a.e.). 与此前证明类似
- $I(X; Y) \geq 0$ with equality iff X and Y are independent. 从相对熵推出

Mutual Information: Master Definition

互信息和熵还是有很大区别的. 我们考虑进阶的定义方式:

The mutual information between two random variables is the limit of the mutual information between their quantized versions

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta|Y^\Delta) \\ &\approx h(X) - \log \Delta - (h(x|y) - \log \Delta) \\ &= I(X; Y) \end{aligned}$$

我们发现, 离散型随机变量与连续型随机变量的互信息(离散化)是近似相等的.

Definition. The mutual information between two random variables X and Y is given by

$$I(X; Y) = \sup_{\mathcal{P}, \mathcal{Q}} I([X]_{\mathcal{P}}, [Y]_{\mathcal{Q}})$$

where the supremum is over all finite partitions \mathcal{P} and \mathcal{Q}

这里, X 和 Y 既可以是离散的也可以是连续的(4种情况), 但定义是通用的.

- Let \mathcal{X} be the range of a random variable X . A partition \mathcal{P} of \mathcal{X} is a finite collection of disjoint sets P_i such that $\bigcup_i P_i = \mathcal{X}$. The quantization of X by \mathcal{P} (denoted $[X]_{\mathcal{P}}$) is the discrete random variable

defined by

$$\Pr([X]_P = i) = \Pr(X \in P_i) = \int_{P_i} dF(x)$$

- For two random variables X and Y with partitions \mathcal{P} and \mathcal{Q} , we can calculate the mutual information between the quantized versions of X and Y

This is the master definition of mutual information that always applies, even to joint distributions with atoms, densities, and singular parts.

0420 Differential Entropy (2)

Correlated Gaussian

我们计算两个随机变量的互信息，这里以联合高斯分布为例。

(Mutual information between correlated Gaussian random variables with correlation ρ) Let $(X, Y) \sim \mathcal{N}(0, K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

$I(X; Y)$?

$$\begin{aligned} h(X) &= h(Y) = \frac{1}{2} \log 2\pi e \sigma^2 \\ h(X, Y) &= \frac{1}{2} \log (2\pi e)^2 |K| = \frac{1}{2} \log (2\pi e)^2 \sigma^4 (1 - \rho^2) \\ I(X; Y) &= h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log (1 - \rho^2) \end{aligned}$$

- $\rho = 0$, X and Y are independent and I is 0
- $\rho = \pm 1$, X and Y are perfectly correlated and I is ∞

Maximum Entropy with Constraints

$E(X^2)$, $Var(X)$ 给定的情况下，高斯分布最大化微分熵。

- Let the random variable $X \in \mathbb{R}$ have mean μ and variance σ^2 . Then

$$h(X) \leq \frac{1}{2} \log 2\pi e \sigma^2$$

with equality iff $X \sim \mathcal{N}(\mu, \sigma^2)$

- Let the random variable $X \in \mathbb{R}$ satisfy $EX^2 \leq \sigma^2$. Then

$$h(x) \leq \frac{1}{2} \log 2\pi e \sigma^2$$

with equality iff $X \sim \mathcal{N}(0, \sigma^2)$

证明域平均分布最大化离散熵的证明如下，我们用相对熵推出。

1. Let $X_G \sim \mathcal{N}(\mu, \sigma^2)$. Consider

$$D(X \| X_G) \geq 0$$

Then

$$\int f \log \frac{f}{g} \geq 0$$

把对数函数展开，由于 g 是高斯分布，可以进一步展开。

$$h(X) = h(f) \leq - \int f \log g = - \int f \log \frac{1}{\sqrt{2\pi\sigma^2}} + f \left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

由于右侧都是常数，可代入化简。

$$h(X) \leq \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} = \frac{1}{2} \log 2\pi e \sigma^2$$

2. $\text{Var}(X) = E(X^2) - E(X)^2 \leq \sigma^2 \Rightarrow \text{Case 1}$

使用这两个结论时一定要注意是否存在确定的均值、方差或二阶矩是否存在上界。

Maximum Entropy

最大熵原理在不同的限制下可以得到不同的结论。（详见 Cover Ch.12）

Consider the following problem: Maximize the entropy $h(f)$ over all probability densities f satisfying (条件)

1. $f(x) \geq 0$, with equality outside the support 非负性
2. $\int_S f(x) dx = 1$ 规范性
3. $\int_S f(x) r_i(x) dx = \alpha_i$ for $1 \leq i \leq m$. ($r_i(x)$ is a function of x). Thus, f is a density on support set S meeting certain moment constraints $\alpha_1, \alpha_2, \dots, \alpha_m$ 即某些关于 x 的函数的均值是一定的。

Theorem 12.1.1 (Maximum entropy distribution) Let

$$f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)}$$

$x \in S$, where $\lambda_0, \dots, \lambda_m$ are chosen so that f^* satisfies $(++)$. Then f^* uniquely maximizes $h(f)$ over all probability densities f satisfying constraints $(++)$

最大熵的分布是 f^* 取到的，但其中一些系数需要通过 λ_i 作为待定系数，还需更多条件可确定待定系数。

Proof.

$$\begin{aligned} h(g) &= - \int_S g \ln g \\ &= - \int_S g \ln \frac{g}{f^*} f^* \\ &= -D(g||f^*) - \int_S g \ln f^* \\ &\stackrel{(a)}{\leq} - \int_S g \ln f^* \\ &\stackrel{(b)}{=} - \int_S g \left(\lambda_0 + \sum \lambda_i r_i \right) \\ &\stackrel{(c)}{=} - \int_S f^* \left(\lambda_0 + \sum \lambda_i r_i \right) \\ &= - \int_S f^* \ln f^* \\ &= h(f^*) \end{aligned}$$

where (a) follows from the nonnegativity of relative entropy, (b) follows from the definition of f^* , and (c) follows from the fact that both f^* and g satisfy the constraints. Note that equality holds in (a) if and only if $g(x) = f^*(x)$ for all x , except for a set of measure 0, thus proving uniqueness.

The same approach holds for discrete entropies and for multivariate distributions.

Examples.

- Let $S = [a, b]$, with no other constraints. Then the maximum entropy distribution is the uniform distribution over this range.
- $S = [0, \infty)$ and $EX = \mu$. Then the entropy-maximizing distribution is

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0$$

- $S = (-\infty, \infty)$, $EX = \alpha_1$, and $EX^2 = \alpha_2$. The maximum entropy distribution is $\mathcal{N}(\alpha_1, \alpha_2 - \alpha_1^2)$

Hadamard's Inequality

K is a nonnegative definite symmetric $n \times n$ matrix. Let $|K|$ denote the determinant of K

Theorem (Hadamard) $|K| \leq \prod K_{ii}$, with equality iff $K_{ij} = 0, i \neq j$

Proof.

Let $X \sim \mathcal{N}(0, K)$. Then

$$\frac{1}{2} \log(2\pi e)^n |K| = h(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n h(X_i) = \sum_{i=1}^n \frac{1}{2} \log 2\pi e |K_{ii}|$$

with equality iff X_1, X_2, \dots, X_n are independent (i.e., $K_{ij} = 0, i \neq j$)

idea: 矩阵转化为多元高斯分布，联合分布熵小于边缘分布熵的和。

remark: 熵是一个基础的物理量，可以用来证明很多不等式(Cover Ch 17.9, 17.10)，比如一系列有关正定矩阵的性质

- $\log |K|$ is concave
- $\log(|K_n| / |K_{n-p}|)$ is concave in K_n
- $|K_n| / |K_{n-1}|$ is concave in K_n

Balanced Information Inequality

平衡信息不等式：离散熵与微分熵的同和不同

Differences between inequalities of the discrete entropy and differential entropy

- Both $H(X, Y) \leq H(X) + H(Y)$ and $h(X, Y) \leq h(X) + h(Y)$ are valid
- $H(X, Y) \geq H(X)$ but neither $h(X, Y) \geq h(X)$ nor $h(X, Y) \leq h(X)$ is valid

Take $H(X, Y, Z) \leq \frac{1}{4}H(X) + \frac{1}{2}H(Y, Z) + \frac{3}{4}H(Z, X)$ for example.

Count the weights of random variables X, Y, Z in both sides $X : 1, 1; Y : 1, \frac{1}{2}; Z : 1, \frac{5}{4}$ 定义 X, Y, Z 的净权重。

The net weights of X, Y, Z are $0, \frac{1}{2}, -\frac{1}{4}$

比如，下面的不等式是平衡的：

Balanced: If the net weights of X, Y, Z are all zero.

$$h(X, Y) \leq h(X) + h(Y) \text{ and } h(X, Y, Z) \leq \frac{1}{2}h(X, Y) + \frac{1}{2}h(Y, Z) + \frac{1}{2}h(Z, X)$$

对更为一般的情况，

Let $[n] := \{1, 2, \dots, n\}$. For any $\alpha \subseteq [n]$, denote $(X_i : i \in \alpha)$ by X_α . For example, $\alpha = \{1, 3, 4\}$, we denote X_1, X_3, X_4 by $X_{(1,3,4)}$ for simplicity.

- We could write any information inequality in the form $\sum_{\alpha} w_{\alpha} H(X_{\alpha}) \geq 0$ or $\sum_{\alpha} w_{\alpha} h(X_{\alpha}) \geq 0$
- An information inequality is called balanced if for any $i \in [n]$, the net weight of X_i is zero.

- The linear continuous inequality $\sum_{\alpha} w_{\alpha} h(X_{\alpha}) \geq 0$ is valid if and only if its corresponding discrete counterpart $\sum_{\mathbf{g}} w_{\mathbf{g}} H(X_{\mathbf{g}}) \geq 0$ is valid and balanced.

由此，我们可以建立微分熵不等式和离散熵不等式的关系。这个不等式是正确的当且仅当它对应的离散熵不等式是正确的且平衡的。

Ref: Balanced Information Inequalities, T. H. Chan, IEEE Transactions on Information Theory, Vol. 49, No. 12, December 2003

Han's Inequality

Let (X_1, X_2, \dots, X_n) have a density, and for every $S \subseteq \{1, 2, \dots, n\}$, denoted by $X(S)$ the subset $\{X_i : i \in S\}$. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} \frac{h(X(S))}{k}$$

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} \frac{h(X(S)|X(S^c))}{k}$$

从n个随机变量中取k个，求联合熵、条件熵。

When $n = 3$,

$$h_1^{(3)} = \frac{H(X_1) + H(X_2) + H(X_3)}{3} \geq h_2^{(3)} = \frac{H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)}{3}$$

$$\geq h_3^{(3)} = H(X_1, X_2, X_3)$$

$$g_1^{(3)} = \frac{H(X_1|X_2, X_3) + H(X_2|X_1, X_3) + H(X_3|X_1, X_2)}{3}$$

$$\leq g_2^{(3)} = \frac{H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_3, X_1|X_2)}{3}$$

$$\leq g_3^{(3)} = H(X_1, X_2, X_3)$$

Han's inequality:

$$h_1^{(n)} \geq h_2^{(n)} \dots \geq h_n^{(n)} = H(X_1, X_2, \dots, X_n) = g_n^{(n)} \geq \dots \geq g_2^{(n)} \geq g_1^{(n)}$$

Information Heat

Heat Equation

- Heat equation (Fourier): Let x be the position and t be the time, 热传导方程。（它与高斯信道是等价的）

$$\frac{\partial}{\partial t} f(x, t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f(x, t)$$

- Let X be any random variable with a density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance, $Z \sim \mathcal{N}(0, 1)$. Let

$$Y_t = X + \sqrt{t}Z$$

The **probability density function** $f(y; t)$ ($f(y; t)$ is a function in y , not t) of Y_t **satisfies heat equation**

$$f(y; t) = \int f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dx$$

高斯信道的输出信号与热传导方程具有一一对应的关系。

Entropy and Fisher Information

对连续型随机变量定义一个新的信息量， Fisher Information

Fisher information: Let X be any random variable with density $f(x)$. Its Fisher information is given by

$$I(x) = \int_{-\infty}^{+\infty} f(x) \left[\frac{\frac{\partial}{\partial x} f(x)}{f(x)} \right]^2 dx$$

- Let X be any random variable with a density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Let $Y_t = X + \sqrt{t}Z$

$$\frac{\partial}{\partial t} h(Y_t) = \frac{1}{2} I(Y_t)$$

表明信息量与统计量之间也存在关联。

- Let $f(y, t)$ (or f) be the p.d.f of Y_t

$$\begin{aligned} \frac{\partial}{\partial t} h(Y_t) &= \frac{1}{2} I(Y_t) = \frac{1}{2} \int \frac{f_y^2}{f} dy \geq 0 \\ \frac{\partial^2}{\partial t^2} h(Y_t) &= -\frac{1}{2} \int f \left(\frac{f_{yy}}{f} - \frac{f_y^2}{f^2} \right)^2 dy \leq 0 \end{aligned}$$

$h(Y_t)$ 关于 t 是一个递增的凹函数。

- When X is Gaussian $\mathcal{N}(0, 1)$

$$h(Y_t) = \frac{1}{2} \log 2\pi e(1 + t)$$

对高斯分布的输入， n 阶导数可求，且符号为

All the derivatives alternate in signs: $+, -, +, -, \dots$

Higher Order Derivatives of $h(Y_t)$

(Cheng 2015) Let X be any random variable with a density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Let $Y_t = X + \sqrt{t}Z$ and $f(y, t)$ (or f) be the p.d.f of Y_t . Then

$$\frac{\partial^3}{\partial t^3} h(Y_t) \geq 0 \text{ and } \frac{\partial^4}{\partial t^4} h(Y_t) \leq 0$$

Conjecture: When n is even, $\frac{\partial^n}{\partial t^n} h(Y_t) \leq 0$, otherwise $\frac{\partial^n}{\partial t^n} h(Y_t) \geq 0$

Ref: F. Cheng and Y. Geng, Higher Order Derivatives in Costa's Entropy Power Inequality

EPI and FII

(Shannon 1948, Entropy power inequality (EPI)) If X and Y are independent random n -vectors with densities, then

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}$$

若对随机变量：

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}$$

- 也可以互推FII不等式，Fisher information inequality (FII)

$$\frac{1}{I(X+Y)} \geq \frac{1}{I(X)} + \frac{1}{I(Y)}$$

- Most profound result in Shannon's 1948 paper
- EPI can imply some very fundamental results
 - Uncertainty principle in quantum physics
 - Young's inequality
 - Nash's inequality
 - Cramer-Rao bound

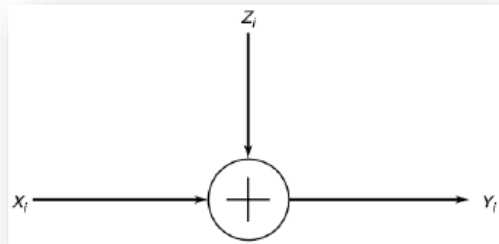
References:

- T. Cover, Information theoretic inequalities, 1990
- O. Rioul, "Information Theoretic Proofs of Entropy Power Inequalities," 2011

0422 Gaussian Channel

Gaussian Channel

连续信道中最常见的是高斯信道，背景噪声服从高斯分布。



Gaussian channel

Continuous alphabet channel

- The channel could be use at each time i
- The input X_i , noise Z_i , output Y_i are continuous

- the most important continuous alphabet channel is the Gaussian channel. For example, wireless telephone channels and satellite links
- The noise Z_i is drawn i.i.d. from a **Gaussian distribution** with variance N
- The noise Z_i is assumed to be **independent** of the signal X_i
- This is a time-discrete channel with output Y_i at time i , where Y_i is the sum of the input X_i and the noise Z_i ,

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N)$$

对连续随机变量，两个随机变量的和的概率密度函数是它们的卷积。

- Without further conditions, the capacity of this channel may be ∞ .
 - The values of X may be very sparse
 - 一个例子：Assume the variance of noise N is neglected compared to the distances of the values of X . Then $Y = X + Z \approx X$. Thus $I(X; Y) \approx H(X)$, which may be ∞ 。连续型随机变量的离散熵有可能是无穷大的，信道容量此时也失去了意义。

为进一步研究信道容量，我们需要从实际出发，为信道增加一些功能、限制。

Energy Constraint

高斯信道中常见的一个限制是能量限制。能量通常与方差相关。在一些研究中，有关“能量”可以用其他更复杂或更精细的模型进行定义，从而得到类似的推导和结论。

- The most common limitation on the input is **an energy or power constraint**
- We assume an average power constraint. For any codeword (x_1, x_2, \dots, x_n) transmitted over the channel, we require that （假设码制中的符号均匀分布）

$$\frac{1}{n} \sum_{i=1}^n x_i^2 \leq P$$

- within the sphere \sqrt{nP}
- P per channel use 每次信道使用功率消耗最大为P
- 如果码制不均匀分布，只需改成 $EX^2 \leq P$
- This communication channel models many practical channels, including radio and satellite links.

The information capacity of the Gaussian channel with power constraint P is

$$C = \max_{f(x): EX^2 \leq P} I(X; Y)$$

下面求解该优化问题。

$$\begin{aligned} I(X; Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X + Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \\ h(Z) &= \frac{1}{2} \log 2\pi e N \end{aligned}$$

$$\begin{aligned} EY^2 &= E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2 \leq P + N \\ h(Y) &\leq \frac{1}{2} \log 2\pi e(P + N) \end{aligned}$$

$$\begin{aligned} I(X; Y) &= h(Y) - h(Z) \leq \frac{1}{2} \log 2\pi e(P + N) - \frac{1}{2} \log 2\pi e N \\ &= \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \end{aligned}$$

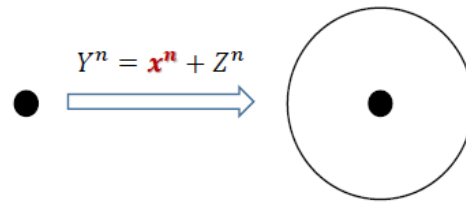
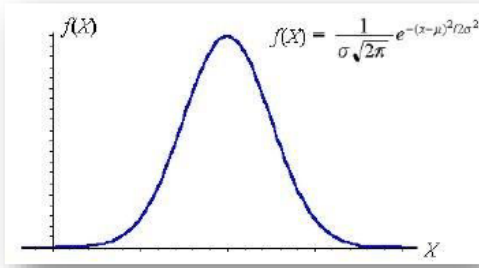
综上：

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} \right)$$

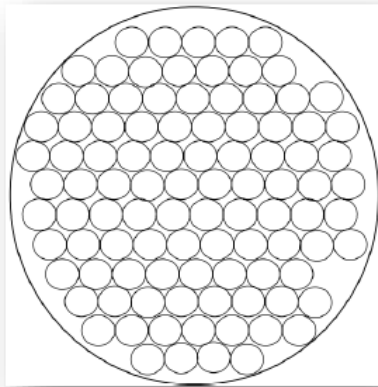
The maximum is attained when $X \sim \mathcal{N}(\mathbf{0}, P)$, $\frac{P}{N}$ 也被称为信道比。

Intuition

每个输入信号都会在接收端产生一定的区域。考虑一个给定的码制 \mathbf{x}_n



- The received vector is **normally distributed** with mean equal to the true codeword and variance equal to the noise variance.
- With high probability, the received vector is contained **in a sphere of radius $\sqrt{n(N + \epsilon)}$ around the true codeword**.
- 任何码制在高斯噪声的影响下，接收端都产生一个球体
- 我们在解码时，只要获得的 Y_n 在球体内，我们就认为 Y_n 对应 X_n ，为了降低错误率，我们希望任意两个球都是不相交的。
- If we assign everything within this sphere to the given codeword, when this codeword is sent there will be an error only if the received vector falls outside the sphere, which has low probability.
 - Each codeword is represented by a sphere
 - Low decoding error requires no intersection between any spheres
 那么填充满的小球空间率越大，码率就越高



Sphere packing for the Gaussian channel

The maximum number of nonintersecting decoding spheres is no more than

$$\frac{C_n(n(P+N))^{\frac{n}{2}}}{C_n(nN)^{\frac{n}{2}}} = \left(1 + \frac{P}{N}\right)^{\frac{n}{2}}$$

$$R \leq \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

$$C = \sup R = \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

- The received vectors ($Y = X + Z$) have energy no greater than $n(P + N)$, so they lie in a sphere of radius $\sqrt{n(P + N)}$
- The volume of an n-dimensional sphere is of the form $C_n r^n$, where r is the radius of the sphere.

$$2\pi r, \pi r^2 \text{ and } \frac{4}{3}\pi r^3$$

- The volumes are approximated by

$$C_n(nN)^{\frac{n}{2}} \text{ and } C_n(n(P + N))^{\frac{n}{2}}$$

我们不需要把球体的公式算的特别细，忽略前面的系数，计算高阶的情况。

我们完成了上界的推算。当然，高维空间小球的堆放问题，其实际的构造是一个非常复杂的问题。

Theorems

证明思路与DMC一致，定义编码，定义错误概率，定义可达区域，证明converse和achievability

Definition

对能量限制P的信道定义编码函数。

Definition. An (M, n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, \dots, M\}$
2. An encoding function $x : \{1, 2, \dots, M\} \rightarrow X^n$, yielding codewords $x^n(1), x^n(2), \dots, x^n(M)$, satisfying the power constraint P ; that is, for every codeword

$$\sum_{i=1}^n x_i^2(w) \leq nP, \quad w = 1, 2, \dots, M$$

3. A decoding function

$$g : y^n \rightarrow \{1, 2, \dots, M\}$$

The arithmetic average of the probability of error is defined by

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum \lambda_i$$

A rate R is said to be achievable for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates.

Code Construction

Generation of the codebook

We generate the codewords (x_1, x_2, \dots, x_n) with each element i.i.d. according to a normal distribution with variance $P - \epsilon$. since for large n

$$\frac{1}{n} \sum x_i^2 \rightarrow P - \epsilon$$

The probability that a codeword does not satisfy the power constraint will be small. 根据方差的定义，超过能量

限制的概率是非常小的。

Let $X_i(w)$, $i = 1, 2, \dots, n, w = 1, 2, \dots, 2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P - \epsilon)$, forming codewords $X^n(1), X^n(2), \dots, X^n(2^{nR}) \in \mathcal{R}^n$ 完成了码本的生成

Encoding

- The codebook is revealed to both the sender and the receiver.
- To send the message index w , sends the w th codeword $X^n(w)$ in the codebook.

Decoding:

根据联合典型性解码, The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector.

- If there is one and only one such codeword $X^n(w)$, the receiver declares $\widehat{W} = w$ to be the transmitted codeword.
- Otherwise, the receiver declares an error. The receiver also declares an error **if the chosen codeword does not satisfy the power constraint.** (除了不满足典型性, 错误还可能是超过了能量限制)

Probability of Error

WLOG, 假设我们发送了码制1.

Without loss of generality, assume that codeword 1 was sent. Thus,

$$Y^n = X^n(1) + Z^n$$

现在我们要分析两种限制。Define the following events:

$$E_0 = \left\{ \frac{1}{n} \sum_{j=1}^n X_j^2(1) > P \right\}$$

and

$$E_i = \left\{ (X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \right\}$$

$$\Pr(\mathcal{E}|W=1) = P(E_0 \cup E_1^c \cup E_2 \cup E_3 \dots \cup E_{2^{nR}}) \leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i)$$

E_0 表示违背了能量约束。

但根据我们前面的分析, $P(E_0) \rightarrow 0$, $P(E_1^c) \leq \epsilon$, 又根据联合典型性,

$$\sum_{i=2}^{2^{nR}} P(E_i) = (2^{nR} - 1) 2^{-n(I(X;Y)-3\epsilon)} \leq 2^{-n(I(X;Y)-R-3\epsilon)}$$

所以我们有 $P_e^{(n)} \leq 3\epsilon$ 。利用DMC中类似的减一半方法，我们可以证明最大错误概率也是趋向于0的。

Converse

Let W be distributed uniformly over $\{1, 2, \dots, 2^{nR}\}$

$$W \rightarrow X^n(W) \rightarrow Y^n \rightarrow \widehat{W}$$

By Fano's inequality

$$H(W|\widehat{W}) \leq 1 + nR P_e^{(n)} = n\epsilon_n$$

where $\epsilon_n \rightarrow 0$ as $P_e^{(n)} \rightarrow 0$

$$\begin{aligned} nR &= H(W) = I(W; \widehat{W}) + H(W|\widehat{W}) \\ &\leq I(W; \widehat{W}) + n\epsilon_n \\ &\leq I(X^n; Y^n) + n\epsilon_n \\ &= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \\ &= h(Y^n) - h(Z^n) + n\epsilon_n \quad \text{去除信号 } Y_n = X_n + Z_n \\ &\leq \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n \end{aligned}$$

Let P_i be the average power of the i th column of the codebook

$$P_i = \frac{1}{2^{nR}} \sum_w x_i^2(w) \text{ and } \frac{1}{n} \sum_i P_i \leq P$$

since X_i and Z_i are independent, then 建立微分熵 $h(Y_i)$ 的上界

$$EY_i^2 = P_i + N, h(Y_i) \leq \frac{1}{2} \log 2\pi e (P_i + N)$$

$$\begin{aligned} nR &\leq \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n \\ &\leq \sum \left(\frac{1}{2} \log 2\pi e (P_i + N) - \frac{1}{2} \log 2\pi e N \right) + n\epsilon_n \\ &= \sum \frac{1}{2} \log 2\pi e \left(1 + \frac{P_i}{N} \right) + n\epsilon_n \end{aligned}$$

利用凹函数的性质, $f(x) = \frac{1}{2} \log(1+x)$ is concave

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \log \left(1 + \frac{P_i}{N} \right) \\ & \leq \frac{1}{2} \log \left(1 + \frac{1}{n} \sum_{i=1}^n \frac{P_i}{N} \right) \leq \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \end{aligned}$$

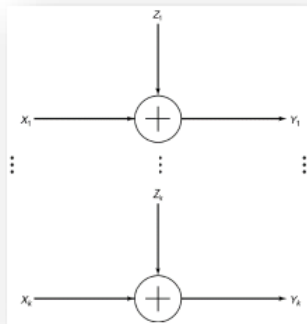
所以

$$R \leq \frac{1}{2} \log \left(1 + \frac{P}{N} \right) + \epsilon_n$$

这里我们借用凹函数的性质处理了能量的问题，完成了converse的证明。

Parallel Gaussian Channel

Problem



$$C = \max_{f(x_1, x_2, \dots, x_k): E \sum X_i^2 \leq P} I(X_1, X_2, \dots, X_k; Y_1, Y_2, \dots, Y_k)$$

高斯信道的扩展：n个信道，可以同时使用。仍然存在能量的限制。

Assume that we have a set of Gaussian channels in parallel. The output of each channel is the sum of the input and Gaussian noise. For channel j

$$Y_j = X_j + Z_j, \quad j = 1, 2, \dots, k$$

The noise is assumed to be independent from channel to channel. We assume that there is a common power constraint on the total power used, that is

$$E \sum_{j=1}^k X_j^2 \leq P$$

We wish to **distribute the power among the various channels** so as to maximize the total capacity. 将问题更细化一步，我们需要对单个信道的能量做一定的分配。

$$P_i = EX_i^2, \text{ and } \sum P_i \leq P$$

Solution

$$\begin{aligned}
 & I(X_1, X_2, \dots, X_k; Y_1, Y_2, \dots, Y_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Y_1, Y_2, \dots, Y_k | X_1, X_2, \dots, X_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Z_1, Z_2, \dots, Z_k | X_1, X_2, \dots, X_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Z_1, Z_2, \dots, Z_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - \sum_i h(Z_i) \\
 &\leq \sum h(Y_i) - h(Z_i) \\
 &\leq \sum_i \frac{1}{2} \log \left(1 + \frac{P_i}{N_i} \right) \quad \text{高斯噪声最大化微分熵}
 \end{aligned}$$

where $P_i = EX_i^2$, and $\sum P_i = P$. 等号是可以取到的, Equality is achieved by

$$(X_1, X_2, \dots, X_k) \sim \mathcal{N} \left(0, \begin{bmatrix} P_1 & 0 & \dots & 0 \\ 0 & P_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_k \end{bmatrix} \right)$$

优化目标:

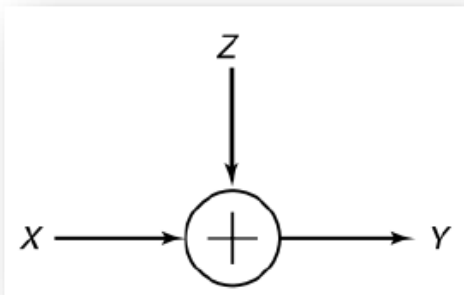
$$\max_{\sum P_i = P} \sum_i \log \left(1 + \frac{P_i}{N_i} \right)$$

凸优化问题, 可解。我们发现, 在并行高斯信道中, 我们需要专门分配能量。

Extension: 分配能量算法: water-filling

Worst Additive Noise

推广到噪声的一般分布, 信道容量定理依然成立。



$$\begin{aligned}
 Y &= X + Z \\
 C &= \max_{X: EX^2 \leq P} I(X; X + Z)
 \end{aligned}$$

Problem

- Under the energy constraint P , the channel capacity of additive channel $Y = X + Z$ is

$$\begin{aligned} C(Z) &= \max_{X: EX^2 \leq P} I(X; Y) \\ &= \max_{X: EX^2 \leq P} h(X + Z) - h(Z) \end{aligned}$$

后一项需要根据具体情况具体分析。

- 加一个限制? What is the minimum of $C(Z)$, if we could choose $Z : EZ^2 \leq N$
 - That is, to play a max-min game between X and Z 在 Z 给定的情况下设定 X

$$\begin{aligned} \max_{Z: EZ^2 \leq N} C(Z) &:= \min_{Z: EZ^2 \leq N} \max_{X: EX^2 \leq P} I(X; X + Z) \\ &= \min_{Z: EZ^2 \leq N} \left(\max_{X: EX^2 \leq P} I(X; X + Z) \right) \end{aligned}$$

- 对多重优化问题，把内部看成一个函数，分两步走。We need to find a Z^* . When $C(Z^*)$ is attained by X^*

$$I(X^*; X^* + Z^*) \leq \max_{X: EX^2 \leq P} I(X; X + Z)$$

- The $\min_{Z: EZ^2 \leq N} C(Z)$ is attained iff $Z = Z_G \sim \mathcal{N}(0, \sigma^2)$ (Shannon, 1948)

Entropy power inequality

在给定信道能量的情况下，高斯噪声是最坏的加性噪声。EPI，熵幂不等式

Entropy power inequality (EPI, Shannon 1948): If X and Y are independent random n vectors with densities, then

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}$$

证明略

利用EPI，我们证明前一节的定理

- Recall $I(X; X + Z) = h(X + Z) - h(Z)$
- By EPI, $h(X + Z) \geq \frac{1}{2} \log(e^{2h(X)} + e^{2h(Z)})$
- 我们有 $I(X; X + Z) \geq \frac{1}{2} \log(e^{2h(X)} + e^{2h(Z)}) - h(Z)$
- $f(t, s) = \frac{1}{2} \log(e^{2t} + e^{2s}) - s$, where

$$\begin{aligned} t &= h(X) \leq \frac{1}{2} \log 2\pi e P \\ s &= h(Z) \leq \frac{1}{2} \log 2\pi e N \end{aligned}$$

- In $f(t, s)$ is increasing and convex in t , and is decreasing and convex in S
- Fix s , $f(t, s)$ is maximized if $t = \frac{1}{2} \log 2\pi e P$
- Fix t , $f(t, s)$ is minimized if $s = \frac{1}{2} \log 2\pi e N$
- $X^* \sim \mathcal{N}(0, P)$, $Z \sim \mathcal{N}(0, N^*)$

In Gaussian channel

$$\begin{aligned}
 I(X; X + Z^*) &\leq I(X^*; X^* + Z^*) = C(Z^*) \\
 &= I(X^*; X^* + Z) \\
 &= h(X^* + Z) - h(Z) \\
 &\geq \frac{1}{2} \log \left(e^{2h(X^*)} + e^{2h(Z)} \right) - h(Z) \\
 &\geq \min f(t, s) \\
 &= I(X^*; X^* + Z^*)
 \end{aligned}$$

综合起来，我们获得了一个不等式链，高斯分布位于中间。同时作为下界和上界而存在。这也揭示了我们在很多问题中，以高斯信道为例的意义。

$$\begin{aligned}
 I(\mathbf{X}; \mathbf{X} + \mathbf{Z}^*) &\leq I(\mathbf{X}^*; \mathbf{X}^* + \mathbf{Z}^*) \leq I(\mathbf{X}^*; \mathbf{X}^* + \mathbf{Z}) \\
 \min_Z \max_X I(X; X + Z) &= \max_X \min_Z I(X; X + Z) \\
 &= \frac{1}{2} \log \left(1 + \frac{P}{N} \right)
 \end{aligned}$$

完结撒花~