

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Lecture 1: Introduction

- Instructor: Fan Cheng, Rm 3-513, SEIEE
(<http://www.cs.sjtu.edu.cn/~chengfan/>)
Office hour: By appointment
TA: TBA
- Textbook: David J.C. MacKay, “Information Theory, Inference, and Learning Algorithms,” Cambridge Press, 2005
(<http://www.inference.org.uk/itprnn/book.html>)
- 16 Weeks := 14 lectures + 1 in-class midterm + 1 Q&A
- Grade policy := 50% final + 30% midterm + 10% attendance + 10% homework

Birth of Information Theory

“A Mathematical Theory of Communication,” Bell System Technical Journal, 27 (3): 379-423, July, 1948.



Claude. E. Shannon
(1916-2001)

[https://en.wikipedia.org/
wiki/Claude_Shannon](https://en.wikipedia.org/wiki/Claude_Shannon)

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

C. E. Shannon, 1948

IEEE Information Theory Society

<http://www.itsoc.org>

IEEE Transactions on Information Theory

[http://ieeexplore.ieee.org/xpl/
RecentIssue.jsp?punumber=18](http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=18)

Shannon: father of information theory

Mathematician

Ph.D. in Mathematics from MIT. Worked at AT&T Bell Labs and RLE in MIT

Electrical engineer

Mater's Thesis: electrical applications of Boolean algebra could construct any logical, numerical relationship

Cryptographer

"A Mathematical Theory of Cryptography," 1949.

Friend of Turing

For two months early in 1943, Shannon came into contact with the leading British mathematician Alan Turing. Shannon and Turing met at teatime in the cafeteria. Turing showed Shannon his 1936 paper that defined what is now known as the "Universal Turing machine"



Magnetic mouse



Juggling



Unicycling

Topics in IT

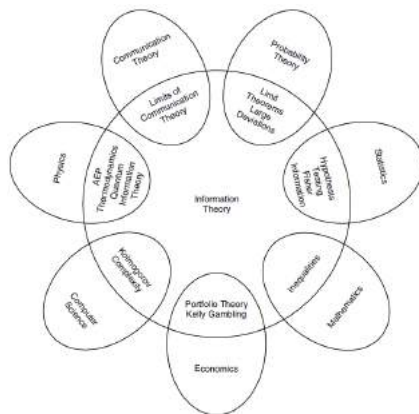
Big Data Analytics
Coding for Communication and Storage
Coding Theory
Combinatorics and Information Theory
Communication Theory
Complexity and Computation Theory
Compressed Sensing and Sparsity
Cryptography and Security

Detection and Estimation
Distributed Storage
Emerging Applications of Information Theory
Information Theory and Statistics
Information Theory in Biology
Information Theory in Computer Science
Statistical/Machine Learning
Network Coding and Applications

Network Data Analysis
Network Information Theory
Optical Communication
Quantum Information and Coding Theory
Shannon Theory
Signal Processing
Source Coding and Data Compression
Wireless Communication and Networks

<https://www.isit2018.org/authors/call-for-papers/>

Information theory to other fields



1

- Information Theory and Reliable Communication, 1st Edition, Robert G. Gallager
- Elements of Information Theory, 2nd Edition (Wiley Series in Telecommunications and Signal Processing), Thomas M. Cover, Joy A. Thomas
- Information Theory: Coding Theorems for Discrete Memoryless Systems, 2nd Edition, Imre Csiszar, Janos Korner
- Information Theory, Inference and Learning Algorithms, David J. C. MacKay
- A First Course in Information Theory (Information Technology: Transmission, Processing and Storage), 1st Edition, Raymond W. Yeung

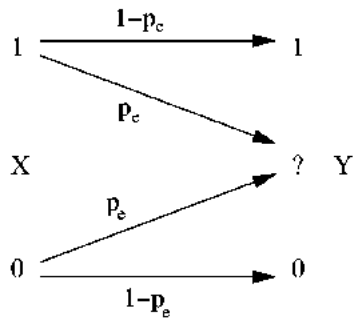
Course plan

- Elements: entropy, mutual information, information divergence, etc.
- Data compression
- Noisy-channel coding
- Probability and inference
- Neural networks
- Low-density parity-check codes

Prerequisite: Probability theory, mathematical analysis, matrix theory

In class: pen and paper

Information theory: An example



Binary Erasure Channel

For random variable X defined on alphabet \mathcal{X} , its mean and variance is defined as

$$\mathcal{E}(X) := \sum_{x \in \mathcal{X}} xp(x)$$

$$\text{Var}(x) := \mathcal{E}(X^2) - (\mathcal{E}(X))^2$$

- $\mathcal{E}(X_1 + X_2) = \mathcal{E}(X_1) + \mathcal{E}(X_2)$
- If X_1 and X_2 are independent, then
 $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$

Some probability distributions: Bernoulli, Binomial, Poisson, Gauss, etc.

A function f on (a, b) is called convex iff for any x_1, x_2 in (a, b)

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f(x_1) + f(x_2)}{2}$$

- If $f(x)$ is twice differentialable, then $f(x)$ is convex iff $f''(x) \geq 0$
- If $f(x)$ is twice differentialable, then $f(x)$ is minimized iff $f'(x) = 0$
- (Jesen's inequality) $f(x)$ is convex in (a, b) ,

$$f(\mathcal{E}(X)) \leq \mathcal{E}f(X)$$

Take $f := e^x, \sin(x), \cos(x), x^2, x^3$ for example

Binomial distribution

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads, r ? What are the mean and variance of r ?

$$p(r|f, N) = \binom{N}{r} f^r (1-f)^{N-r}$$

$$\mathcal{E}(r) = Nf$$

$$\text{Var}(r) = Nf(1-f)$$

Approximating $x!$ and $\binom{N}{r}$

Stirling's approximation

$$x! \simeq x^x e^{-x} \sqrt{2\pi x} \iff \ln x! = x \ln x - x + \frac{1}{2} \ln 2\pi x$$

- Poisson distribution: $P(r|\lambda) = e^{-\lambda} \frac{\lambda^r}{r!}$
- When λ is large and $r \rightarrow \lambda$, $P(r|\lambda) \rightarrow \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(r-\lambda)^2}{2\lambda}}$
- Plug $r = \lambda$

$$\ln \binom{N}{r} \simeq (N-r) \ln \frac{N}{N-r} + r \ln \frac{N}{r}$$

Binary Entropy Function

$$H_2(x) = -x \log x - (1-x) \log(1-x)$$

$$\binom{N}{r} \simeq 2^{NH_2(r/N)}$$

Exercise

- Plot $H_2(x)$ in python
- $H_2(x)$ is symmetric at $x = \frac{1}{2}$
- $H_2(x)$ is maximized at $x = \frac{1}{2}$
- Let $H_2^{-1}(x)$ be the inverse of $H_2(x)$ and $H_2^{-1}(x) \in [0, 1/2]$. For $p \in [0, 1]$, define $p * x := (1 - p)x + p(1 - x)$. Prove that $H_2(p * H_2^{-1}(x))$ is convex in x

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Lecture 2: Entropy and Mutual Information

- Entropy
- Mutual Information

Random Variable and Entropy

Support Set

For a random variable X , denote its alphabet by \mathcal{X} . The probability distribution of X is $p(x)$. The support set of X is defined as

$$\text{supp}(X) := \{x : p(x) > 0, x \in \mathcal{X}\}$$

- $\text{supp}(X) \subseteq \mathcal{X}$
- $x \rightarrow 0, x \log x \rightarrow 0$

Entropy

For a random variable X with probability density function $p(x)$, its entropy is defined as

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

- $H(X) \geq 0$
- $H(X) \leq \log |\mathcal{X}|$

Joint Entropy and Conditional Entropy

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of random variables X and Y is defined by

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

Conditional Entropy

For random variables of X and Y , the conditional entropy of Y given X is defined by

$$H(Y|X) := - \sum_{x,y} p(x, y) \log p(y|x) = -\mathcal{E} \log p(Y|X)$$

$$H(X, Y), H(X), H(Y)$$

Proposition

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

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

Proof from definitions.

Mutual Information

For random variables X and Y , the mutual information between X and Y is defined by

$$I(X; Y) := \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \mathcal{E} \log \frac{p(X, Y)}{p(X)p(Y)}$$

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

Proposition

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

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

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

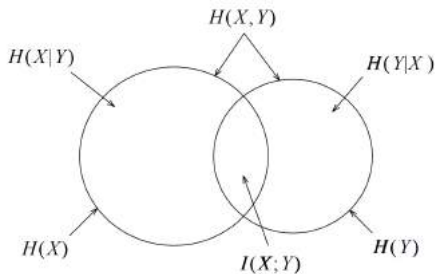


Figure: Relationship between entropies and mutual information for two random variables

Conditional Mutual Information

For random variables X , Y and Z , the mutual information between X and Y conditioning on Z is defined by

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

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

Generic information diagram

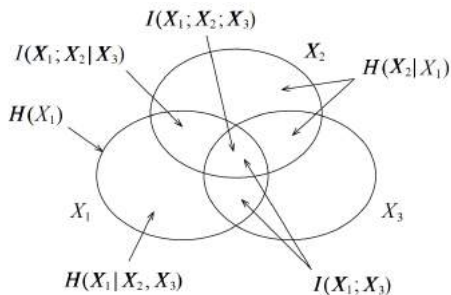


Figure: Information diagram for three random variables

Let X_1 and X_2 be independent binary random variables with

$$P(X_i = 0) = P(X_i = 1) = 0.5,$$

$i = 1, 2$. Let

$$X_3 = (X_1 + X_2) \bmod 2.$$

Calculate $I(X_1; X_2; X_3)$ ($I(X_1; X_2; X_3)$ is not an information measure)

Chain Rule: Entropy

Chain rule for entropy

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

Proof by induction.

Chain rule for conditional entropy

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

$$p(x_1, x_2, \dots, x_n) = \prod p(x_i | x_1, \dots, x_{i-1})$$

Chain Rule: Mutual Information

Chain rule for mutual information

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

Proof by induction.

Chain rule for conditional mutual information

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

$D(p||q)$

The informational divergence between two probability distributions p and q on a common alphabet \mathcal{X} is defined as

$$D(p||q) := \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathcal{E}_p \log \frac{p(X)}{q(X)},$$

where \mathcal{E}_p denotes expectation with respect to p .

- In convention, $p(x) \log \frac{p(x)}{q(x)} = \infty$ if $q(x) = 0$.
- $D(p||q)$ is not symmetric.
- $D(p||q)$ is not a metric. It does not satisfy the triangular inequality.
- $D(p||q) \geq 0$ (Proof via $\ln a \geq 1 - \frac{1}{a}$)

Two Inequalities on $D(p||q)$

Log-sum Inequality

For positive numbers a_1, a_2, \dots and nonnegative numbers b_1, b_2, \dots such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}.$$

Moreover, equality holds if and only if $\frac{a_i}{b_i} = \text{constant}$ for all i .

Let p and q be two probability distributions on a common alphabet \mathcal{X} . The variational distance between p and q is defined by

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

Pinsker's inequality

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

Chain Rule for Relative Entropy

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Proof.

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

Convexity of relative entropy

$D(p||q)$ is convex in the pair (p, q) ; that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 || q_1) + (1 - \lambda)D(p_2 || q_2)$$

for all $0 \leq \lambda \leq 1$.

Checked by log-sum inequality.

Concavity of entropy

$H(p)$ is a concave function of p .

$H(p) = \log |\mathcal{X}| - D(p||u)$, where u is the uniform distribution on $|\mathcal{X}|$ outcomes.

DPI

If X, Y, Z form a Markov chain $X \rightarrow Y \rightarrow Z$ (i.e. $p(x, y, z) = p(x)p(y|x)p(z|y)$), then

$$I(X; Y) \geq I(X; Z)$$

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

Corollary

- For any function g , $I(X; Y) \geq I(X; g(Y))$
- If $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$.

$$I(X; Y|Z) \geq 0$$

$$I(X; Y|Z) = \sum p(z) D(P_{XY|z} || P_{X|z} P_{Y|z})$$

In the information diagram, except $I(X; Y; Z)$, every region is non-negative

$H(X) = 0$ if and only if X is deterministic

$I(X; Y) = 0$ if and only if X and Y are independent

$H(Y|X) = 0$ if and only if Y is a function of X

Equivalent condition

More Information Inequalities

(Conditioning reduces entropy)(Information cant hurt)

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

with equality if and only if X and Y are independent.

(Independence bound on entropy)

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

$$H(X_1, X_2, \dots, X_n) \leq \sum_i H(X_i)$$

with equality if and only if the X_i are independent

Chain rule + conditioning

Question

Conditioning reduce mutual information?

$$I(X; Y|Z) \leq I(X; Y)$$

Axiomatic definition of entropy

If a sequence of symmetric function $H_m(p_1, \dots, p_m)$ satisfies the following properties:

- Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = 1$.
- Continuity: $H_2(p, 1 - p)$ is continuous function of p .
- Grouping: $H_m(p_1, p_2, \dots, p_m) = H_{m-1}(p_1 + p_2, p_3, \dots, p_m) + (p_1 + p_2)H_2(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$, H_m must be of the form of entropy function.

Rényi Entropy

For a discrete random variable X with probability density function $p(x)$, its Rényi entropy with index α is defined as

$$H_{\alpha}(X) = \frac{1}{1-\alpha} \log \sum_x p^{\alpha}(x)$$

When $\alpha \rightarrow 1$, $H_{\alpha}(X) \rightarrow H(X)$

Differential Entropy

For a continuous random variable $X \sim f(x)$, its differential entropy is defined as

$$h(x) := - \int_x f(x) \log f(x) dx$$

$h(x)$ may be negative

Uniform Distribution

If X is uniformly distributed from 0 to a , then

$$h(X) = \log a$$

Gaussian Distribution

If $X \sim \mathcal{N}(0, \sigma^2)$, then

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

More on $h(X)$

$$h(X + c) = h(X).$$

Translation does not change the differential entropy.

$$h(aX) = h(X) + \log |a|$$

Checked by definition

For vector-valued random variable X ,

$$h(AX) = h(X) + \log |\det(A)|$$

Proof is not required

- Ch. 2 (Yeung), Ch. 2 (Cover)
- Facets of entropy:
<http://www.inc.cuhk.edu.hk/EII2013/entropy.pdf>

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Recap: fundamental information quantities

- Definition of KL-divergence, entropy, mutual information
- Information diagram: relationship of entropy, mutual information, etc.
- Some fundamental properties: non-negative, convexity/concavity, inequalities
- How to solve problems via information quantities

Lecture 3: Error Correcting

- Repetition code
- Hamming code
- Decomposability of entropy

Noisy world



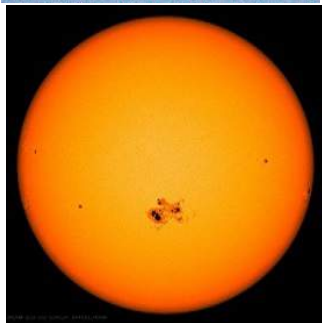
Noisy world



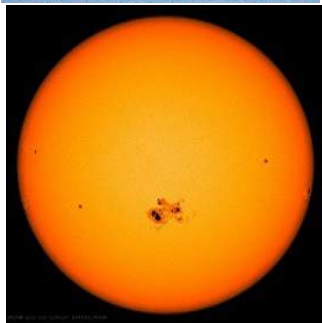
Noisy world



Noisy world

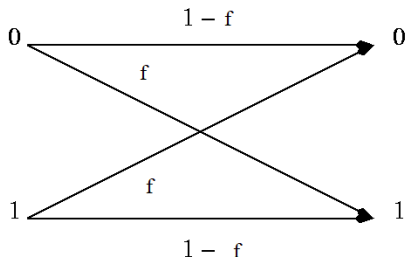


Noisy world



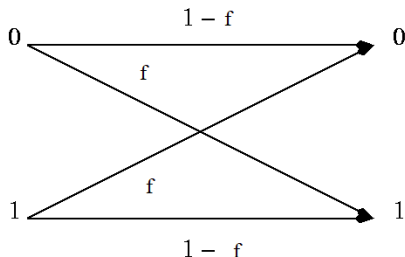
We need to correct them!

Mathematical model: binary symmetric channel



Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f .

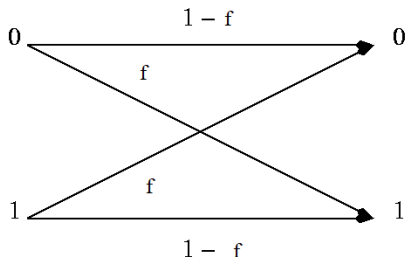
Mathematical model: binary symmetric channel



Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f .

Equivalently, BSC can be written as: $Y = X + Z \pmod{2}$, where $X, Z \in \{0,1\}$

Mathematical model: binary symmetric channel

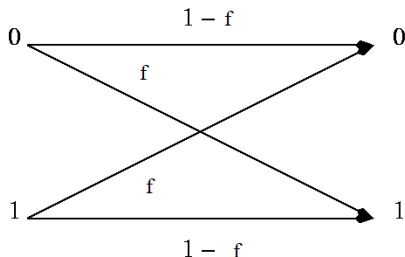


Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f .

Equivalently, BSC can be written as: $Y = X + Z \pmod{2}$, where $X, Z \in \{0,1\}$

For example: $01010101 \rightarrow 01110110$

Mathematical model: binary symmetric channel



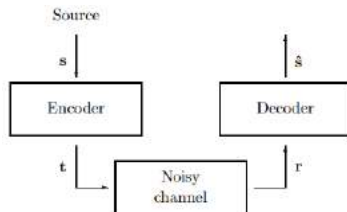
Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f .

Equivalently, BSC can be written as: $Y = X + Z \pmod{2}$, where $X, Z \in \{0,1\}$

For example: $01010101 \rightarrow 01110110$

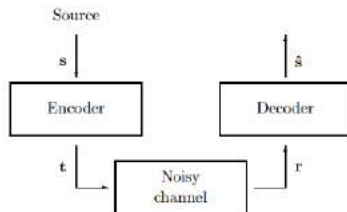
Question: how to transmit a message with very low error probability (e.g. 10^{-15})?

A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

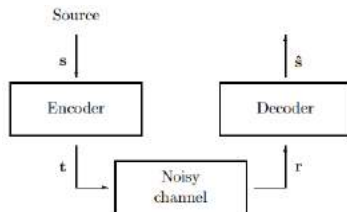
A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

Information theory: What is the best error-correcting performance we could achieve?

A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

Information theory: What is the best error-correcting performance we could achieve?

Coding theory: The creation of practical encoding and decoding systems.

Repetition codes

R_k : Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

Repetition codes

R_k : Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

Repetition codes

R_k : Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

Encoding is obvious. How to decoding? Majority vote? Is it optimal?

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given \mathbf{r} ; i.e., $\max_s P(s|\mathbf{r})$.

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given \mathbf{r} ; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \rightarrow t \rightarrow r \rightarrow \hat{s}$

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given \mathbf{r} ; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \rightarrow t \rightarrow r \rightarrow \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given \mathbf{r} ; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \rightarrow t \rightarrow r \rightarrow \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Proof.

By Bayes' theorem, $P(s|\mathbf{r}) = \frac{P(r,s)}{P(\mathbf{r})} = \frac{P(s)P(\mathbf{r}|s)}{P(\mathbf{r})}$



Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given \mathbf{r} ; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \rightarrow t \rightarrow r \rightarrow \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Proof.

By Bayes' theorem, $P(s|\mathbf{r}) = \frac{P(r,s)}{P(\mathbf{r})} = \frac{P(s)P(\mathbf{r}|s)}{P(\mathbf{r})}$



- [2.] Assume $f < 0.5$, the winning hypothesis is the one with the most 'votes'.

Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^N P(r_n|t_n(s))$$



Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^N P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$



Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^N P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$

The likelihood ratio (Since $s \in \{0,1\}$, likelihood ratio test can tell us the optimal solution) for the two hypotheses is

$$\frac{P(\mathbf{r}|s=1)}{P(\mathbf{r}|s=0)} = \prod_{n=1}^N \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$$



Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^N P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$

The likelihood ratio (Since $s \in \{0,1\}$, likelihood ratio test can tell us the optimal solution) for the two hypotheses is

$$\frac{P(\mathbf{r}|s=1)}{P(\mathbf{r}|s=0)} = \prod_{n=1}^N \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$$

Each factor $\frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$ equals $\frac{1-f}{f}$ if $r_n = 1$ and $\frac{f}{1-f}$ if $r_n = 0$. Thus majority vote is optimal if $f < 0.5$



Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
\hat{s}	0	0	1	0	0	1	0

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
\hat{s}	0	0	1	0	0	1	0

Not all the errors can be detected.

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
\hat{s}	0	0	1	0	0	1	0

Not all the errors can be detected.

- Error probability can be reduced by R_k
- Rate of information transfer has fallen by a factor of k ($k \rightarrow \infty, :($)

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
\hat{s}	0	0	1	0	0	1	0

Not all the errors can be detected.

- Error probability can be reduced by R_k
- Rate of information transfer has fallen by a factor of k ($k \rightarrow \infty$, :()

Proper redundancy is needed to reliable communication. Tradeoff exists between rate of information and error probability.

Hamming code

A **block code** is a rule for converting a sequence of source bits s , of length K , into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Hamming code

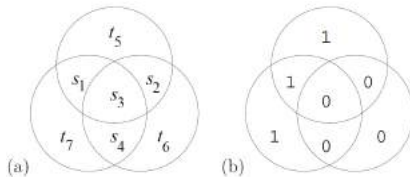
A **block code** is a rule for converting a sequence of source bits s , of length K , into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N

Hamming code

A **block code** is a rule for converting a sequence of source bits s , of length K , into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N

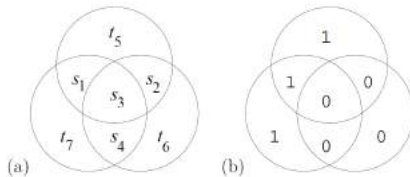


Encoder of (7,4) Hamming code:

Hamming code

A **block code** is a rule for converting a sequence of source bits s , of length K , into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N



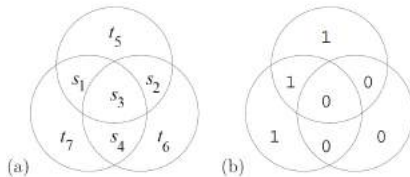
Encoder of (7,4) Hamming code:

- $t_1 t_2 t_3 t_4$ are set equal to $s_1 s_2 s_3 s_4$

Hamming code

A **block code** is a rule for converting a sequence of source bits s , of length K , into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N



Encoder of (7,4) Hamming code:

- $t_1 t_2 t_3 t_4$ are set equal to $s_1 s_2 s_3 s_4$
- The parity-check bits $t_5 t_6 t_7$ are set so that the parity within each circle is even

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{t\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{t\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),
where G is the generator matrix of the code.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{t\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),
where G is the generator matrix of the code.

- Higher information rate

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{t\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),

where G is the generator matrix of the code.

- Higher information rate
- More complicated encoder

Matrix form of Hamming (7,4) code

$$t = sG,$$

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Decoding the (7,4) Hamming code

Facts:

Decoding the (7,4) Hamming code

Facts:

- $s \rightarrow t \rightarrow r$, any of the bits may have been flipped, including the parity bits

Decoding the (7,4) Hamming code

Facts:

- $s \rightarrow t \rightarrow r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are **equiprobable**, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.

Decoding the (7,4) Hamming code

Facts:

- $s \rightarrow t \rightarrow r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are **equiprobable**, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.
- We could solve the decoding problem by measuring how far r is from each of the sixteen codewords, then picking the closest.

Decoding the (7,4) Hamming code

Facts:

- $s \rightarrow t \rightarrow r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are **equiprobable**, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.
- We could solve the decoding problem by measuring how far r is from each of the sixteen codewords, then picking the closest.

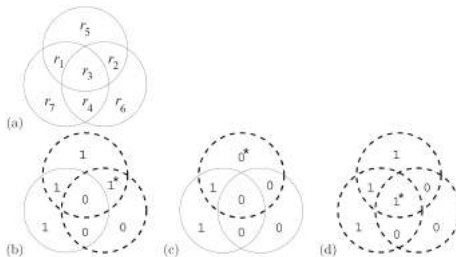
Not efficient!

Syndrome decoding for the Hamming code

The pattern of violations of the parity checks is called the syndrome, and can be written as a binary vector (In Fig. b, the syndrome is $z = (1, 1, 0)$). (Syndrome: Happy (parity 0) and unhappy (parity 1))

Syndrome decoding

Find a unique bit that lies inside all the 'unhappy' circles and outside all the 'happy' circles. Flip this bit for correction.

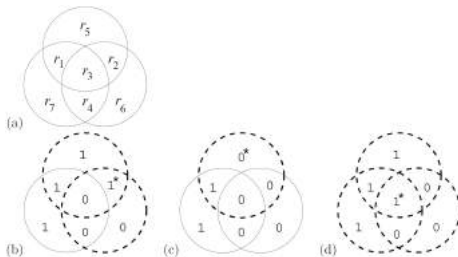


Syndrome decoding for the Hamming code

The pattern of violations of the parity checks is called the syndrome, and can be written as a binary vector (In Fig. b, the syndrome is $z = (1, 1, 0)$). (Syndrome: Happy (parity 0) and unhappy (parity 1))

Syndrome decoding

Find a unique bit that lies inside all the 'unhappy' circles and outside all the 'happy' circles. Flip this bit for correction.



Syndrome z	000	001	010	011	100	101	110	111
Unflip this bit	none	r_7	r_6	r_4	r_5	r_1	r_2	r_3

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I}_4 \\ \mathbf{P} \end{pmatrix},$$

where \mathbf{I}_4 is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix \mathbf{H} is given by $\mathbf{H} = [-\mathbf{P} \ \mathbf{I}_3]$. Since $-1 \equiv 1 \pmod{2}$, $\mathbf{H} = [\mathbf{P} \ \mathbf{I}_3]$.

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I}_4 \\ \mathbf{P} \end{pmatrix},$$

where \mathbf{I}_4 is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix \mathbf{H} is given by $\mathbf{H} = [-\mathbf{P} \ \mathbf{I}_3]$. Since $-1 \equiv 1 \pmod{2}$, $\mathbf{H} = [\mathbf{P} \ \mathbf{I}_3]$.

G and H

All the codewords $\mathbf{t} = \mathbf{G}^t \mathbf{s}$ of the code satisfy $\mathbf{Ht} = \mathbf{0}$, i.e., $\mathbf{HG}^t = \mathbf{0}$
In general, $\mathbf{G} = [\mathbf{I}_k | \mathbf{P}]$, $\mathbf{H} = [-\mathbf{P}^T | \mathbf{I}_{n-k}]$.

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I}_4 \\ \mathbf{P} \end{pmatrix},$$

where \mathbf{I}_4 is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix \mathbf{H} is given by $\mathbf{H} = [-\mathbf{P} \ \mathbf{I}_3]$. Since $-1 \equiv 1 \pmod{2}$, $\mathbf{H} = [\mathbf{P} \ \mathbf{I}_3]$.

G and H

All the codewords $\mathbf{t} = \mathbf{G}^t\mathbf{s}$ of the code satisfy $\mathbf{Ht} = \mathbf{0}$, i.e., $\mathbf{HG}^t = \mathbf{0}$

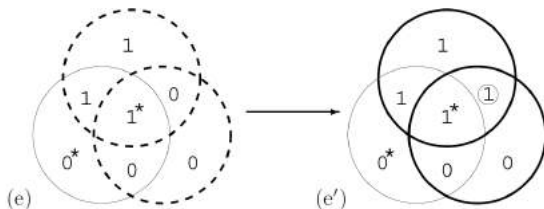
In general, $\mathbf{G} = [\mathbf{I}_k | \mathbf{P}]$, $\mathbf{H} = [-\mathbf{P}^T | \mathbf{I}_{n-k}]$.

Since the received vector $\mathbf{r} = \mathbf{G}^t\mathbf{s} + \mathbf{n}$, the syndrome-decoding problem is to find the most probable noise vector \mathbf{n} satisfying the equation

$$\mathbf{Hn} = \mathbf{z}$$

A decoding algorithm that solves this problem is called a *maximum-likelihood decoder*.

More than two bits are flipped



$$1000101 \rightarrow 10\textcolor{red}{1}0100$$

When two bits, r_3 and r_7 , are received flipped. The syndrome, 110, makes us suspect the single bit r_2 ; so our optimal decoding algorithm flips this bit. If we use the optimal decoding algorithm, any two-bit error pattern will lead to a decoded seven-bit vector that contains three errors.

$$1000101 \rightarrow 10\textcolor{red}{1}0100 \rightarrow 1\textcolor{red}{1}10100$$

There is a tradeoff between p_b and R . The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.

Performance of best code

There is a tradeoff between p_b and R . The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.

The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Performance of best code

There is a tradeoff between p_b and R . The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.

The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Capacity of BSC

$$C(f) = 1 - H_2(f)$$

Performance of best code

There is a tradeoff between p_b and R . The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.

The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Capacity of BSC

$$C(f) = 1 - H_2(f)$$

Exercise

1.3, 1.6, 2.28, 2.29

Decomposability of the entropy

The entropy of any probability distribution $p = \{p_1, p_2, \dots, p_I\}$ is that

$$H(p) = H(p_1, 1 - p_1) + (1 - p_1)H(p_2/(1 - p_1), \dots, p_I/(1 - p_1))$$

Generalizing further,

$$\begin{aligned} H(p) &= H(p_1 + \dots + p_m, p_{m+1} + \dots + p_I) \\ &\quad + (p_1 + \dots + p_m)H(p_1/(p_1 + \dots + p_m), \dots, p_m/(p_1 + \dots + p_m)) \\ &\quad + (p_{m+1} + \dots + p_I)H(p_{m+1}/(p_{m+1} + \dots + p_I), \dots, p_I/(p_{m+1} + \dots + p_I)) \end{aligned}$$

An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, \dots\}$, the number of flips?

An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, \dots\}$, the number of flips?

$$H(X) = H_2(f) + (1 - f)H(X)$$

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Recap: Error Correcting

- Channel model: BSC and its channel capacity
- Repetition code: encoder, decoder, information rate
- $(7, 4)$ Hamming code: encoder, decoder, information rate

Lecture 4: Lossy Source Coding

- Information content
- Shannon source coding theorem
- Typical set
- Fundamental tools

A mathematical brain-teaser



A mathematical brain-teaser



Oddball Problem

A mathematical brain-teaser



Oddball Problem

- Given 12 balls, all equal in weight except for **one** that is either heavier or lighter.

A mathematical brain-teaser



Oddball Problem

- Given 12 balls, all equal in weight except for **one** that is either heavier or lighter.
- You are also given a two-pan balance to use. In each use of the balance, there are three possible outcomes: equal, heavier, or lighter.

A mathematical brain-teaser



Oddball Problem

- Given 12 balls, all equal in weight except for **one** that is either heavier or lighter.
- You are also given a two-pan balance to use. In each use of the balance, there are three possible outcomes: equal, heavier, or lighter.
- Design a strategy to determine which is the odd ball and whether it is heavier or lighter in as few uses of the balance as possible.

(a) How can one measure information?

Questions

- (a) How can one measure information?
- (b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?

Questions

- (a) How can one measure information?
- (b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?
- (c) Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to be gained?

Questions

- (a) How can one measure information?
- (b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?
- (c) Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to be gained?
- (d) How much information is gained on the first step of the weighing problem if 6 balls are weighed against the other 6? How much is gained if 4 are weighed against 4 on the first step, leaving out 4 balls?

Ensemble and Information content

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, \dots, a_i, \dots, a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, \dots, p_I\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Ensemble and Information content

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, \dots, a_i, \dots, a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, \dots, p_I\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Shannon Information Content

The Shannon information content of the outcome $x = a_i$ is

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

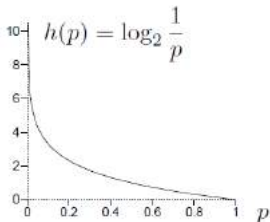
Ensemble and Information content

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, \dots, a_i, \dots, a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, \dots, p_I\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Shannon Information Content

The Shannon information content of the outcome $x = a_i$ is

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

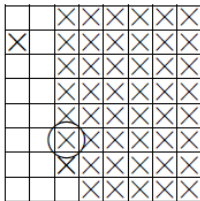


The game 'sixty-three'. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

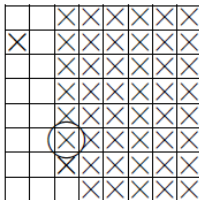
The game 'sixty-three'. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

- 1: is $x \geq 32$?
- 2: is $x \bmod 32 \geq 16$?
- 3: is $x \bmod 16 \geq 8$?
- 4: is $x \bmod 8 \geq 4$?
- 5: is $x \bmod 4 \geq 2$?
- 6: is $x \bmod 2 = 1$?

The game of submarine

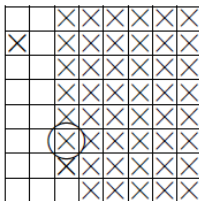


The game of submarine



The entropy is: $H(X) = \log_2 64 = 6$

The game of submarine

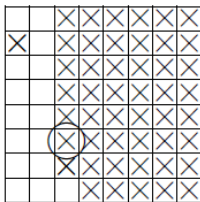


The entropy is: $H(X) = \log_2 64 = 6$

If we hit the submarine when there are n squares left to choose from, then the total information gained is:

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \dots + \log_2 \frac{n+1}{n} + \log_2 \frac{n}{1} = \log_2 64 = 6$$

The game of submarine



The entropy is: $H(X) = \log_2 64 = 6$

If we hit the submarine when there are n squares left to choose from, then the total information gained is:

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \dots + \log_2 \frac{n+1}{n} + \log_2 \frac{n}{1} = \log_2 64 = 6$$

Meaning of Shannon information content

Shannon information content measures the length of a binary file that encodes x .

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ -sufficient subset S_δ is the smallest subset of \mathcal{A}_X satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ -sufficient subset S_δ is the smallest subset of \mathcal{A}_X satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

The essential bit content of X is

$$H_\delta(X) = \log_2 |S_\delta|$$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ -sufficient subset S_δ is the smallest subset of \mathcal{A}_X satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

The essential bit content of X is

$$H_\delta(X) = \log_2 |S_\delta|$$

Denote by X^N the ensemble (X_1, X_2, \dots, X_N) , where X_i 's are independent identically distributed random variables. What is $H_\delta(X^N)$?

Shannon's source coding theorem

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon$$

Shannon's source coding theorem

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon$$

- Typicality: Law of large numbers

Shannon's source coding theorem

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon$$

- Typicality: Law of large numbers
- Some useful fundamental inequalities

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)\dots P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} \dots p_l^{p_l N}$$

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)\dots P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} \dots p_I^{p_I N}$$

The information content is

$$\log_2 \frac{1}{P(x)} \simeq N \sum_i p_i \log_2 \frac{1}{p_i} = NH$$

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)\dots P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} \dots p_I^{p_I N}$$

The information content is

$$\log_2 \frac{1}{P(x)} \simeq N \sum_i p_i \log_2 \frac{1}{p_i} = NH$$

Typical set

$$T_{N\beta} := \{x \in \mathcal{A}_X^N : \left| \frac{1}{N} \log_2 \frac{1}{P(x)} - H \right| < \beta\}$$

‘Asymptotic equipartition’ principle (AEP). With N sufficiently large, the outcome $x = (x_1, x_2, \dots, x_N)$ is almost certain to belong to a subset of \mathcal{A}_X^N having only $2^{NH(X)}$ members, each having probability ‘close to’ $2^{-NH(X)}$.

Several fundamental inequalities

Chebyshev's inequality 1

Let t be a non-negative real random variable, and let α be a positive real number. Then

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha},$$

where \bar{t} is the mean of t .

$P(t \geq \alpha) = \sum_{t \geq \alpha} P(t)$. We multiply each term by $t/\alpha \geq 1$ and obtain:
 $P(t \geq \alpha) \leq \sum_{t \geq \alpha} P(t)t/\alpha$. We add the (non-negative) missing terms and obtain: $P(t \geq \alpha) \leq \sum_t P(t)t/\alpha = \bar{t}/\alpha$

Several fundamental inequalities

Chebyshev's inequality 1

Let t be a non-negative real random variable, and let α be a positive real number. Then

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha},$$

where \bar{t} is the mean of t .

$P(t \geq \alpha) = \sum_{t \geq \alpha} P(t)$. We multiply each term by $t/\alpha \geq 1$ and obtain:
 $P(t \geq \alpha) \leq \sum_{t \geq \alpha} P(t)t/\alpha$. We add the (non-negative) missing terms and obtain: $P(t \geq \alpha) \leq \sum_t P(t)t/\alpha = \bar{t}/\alpha$

Chebyshev's inequality 2

Let x be a random variable, and let α be a positive real number. Then

$$P((x - \bar{x})^2 \geq \alpha) \leq \sigma_x^2/\alpha$$

Take $t = (x - \bar{x})^2$.

Weak Law of Large Numbers

Take x to be the average of N independent random variables h_1, h_2, \dots, h_N , having common mean \bar{h} and common variance σ_h^2 . Then

$$P((x - \bar{h})^2 \geq \alpha) \leq \sigma_h^2 / \alpha N$$

Take $\bar{x} = \bar{h}$ and $\sigma_x^2 = \sigma_h^2 / N$.

Proof of Source Coding Theorem

- $\frac{1}{N} H_{\delta}(X^N) < H + \epsilon$
- $\frac{1}{N} H_{\delta}(X^N) > H - \epsilon$

Reading: Ch. 4 (MacKay)

Exercise

4.9, 4.10, 4.11, 4.12

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Recap: Lossy Source Coding

- Shannon source coding theorem
- Probability inequalities
- Typicality: Application of Law of large numbers

Lecture 5: Symbol Codes

- Introduction
- Kraft inequality
- Optimality of symbol codes
- Huffman Coding

A (binary) symbol code \mathcal{C} for an ensemble X is a mapping from the range of x , $A_X = \{a_1, \dots, a_I\}$, to $\{0, 1\}^+$. $c(x)$ will denote the codeword corresponding to x , and $l(x)$ will denote its length, with $l_i = l(a_i)$.

A (binary) symbol code \mathcal{C} for an ensemble X is a mapping from the range of x , $A_X = \{a_1, \dots, a_I\}$, to $\{0, 1\}^+$. $c(x)$ will denote the codeword corresponding to x , and $l(x)$ will denote its length, with $l_i = l(a_i)$.

$$A_X = \{a, b, c, d\}, P_X = \{1/2, 1/4, 1/8, 1/8\}$$

a_i	$c(a_i)$	l_i
a	1000	4
b	0100	4
c	0010	4
d	0001	4

A (binary) symbol code \mathcal{C} for an ensemble X is a mapping from the range of x , $A_X = \{a_1, \dots, a_I\}$, to $\{0, 1\}^+$. $c(x)$ will denote the codeword corresponding to x , and $l(x)$ will denote its length, with $l_i = l(a_i)$.

$$A_X = \{a, b, c, d\}, P_X = \{1/2, 1/4, 1/8, 1/8\}$$

a_i	$c(a_i)$	l_i
a	1000	4
b	0100	4
c	0010	4
d	0001	4

The extended code C^+ is a mapping from \mathcal{A}_X^+ to $\{0, 1\}^+$ obtained by concatenation, without punctuation, of the corresponding codewords:

$$c^+(x_1, x_2, \dots, x_N) = c(x_1)c(x_2)\dots c(x_N)$$

$$c^+(acdbac) = 100000100001010010000010$$

Unique decoding

The decoding result should be unique

The decoding result should be unique

Prefix code

A symbol code is called a prefix code if no codeword is a prefix of any other codeword.

$\{0, 101\}$, $\{1, 110\}$

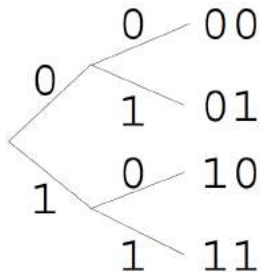
Unique decoding

The decoding result should be unique

Prefix code

A symbol code is called a prefix code if no codeword is a prefix of any other codeword.

$\{0, 101\}$, $\{1, 110\}$



Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

$$\sum_{i=1}^I 2^{-l_i} \leq 1$$

where $I = |\mathcal{X}|$

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

$$\sum_{i=1}^I 2^{-l_i} \leq 1$$

where $I = |\mathcal{A}_X|$

Completeness. If a uniquely decodeable code satisfies the Kraft inequality with equality then it is called a complete code.

Unique decodeability

Question: Given a list of positive integers $\{l_i\}$, does there exist a uniquely decodeable code with those integers as its codeword lengths?

Kraft inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0, 1\}$, the codeword lengths must satisfy:

$$\sum_{i=1}^I 2^{-l_i} \leq 1$$

where $I = |\mathcal{A}_X|$

Completeness. If a uniquely decodeable code satisfies the Kraft inequality with equality then it is called a complete code.

Kraft inequality and prefix codes. Given a set of codeword lengths that satisfy the Kraft inequality, there exists a uniquely decodeable prefix code with these codeword lengths.

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_X} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_X} p_i l_i$$

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_X} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_X} p_i l_i$$

Source coding theorem for symbol codes

For an ensemble X there exists a prefix code C with expected length satisfying

$$H(X) \leq L(C, X) < H(X) + 1$$

Expected Length

The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in \mathcal{A}_X} P(x) l(x)$$

We may also write this quantity as

$$L(C, X) = \sum_{x \in \mathcal{A}_X} p_i l_i$$

Source coding theorem for symbol codes

For an ensemble X there exists a prefix code C with expected length satisfying

$$H(X) \leq L(C, X) < H(X) + 1$$

$$l_i = \log_2(1/p_i)$$

Huffman coding

Reading: Ch. 5 (David MacKay)

Exercise

5.14, 5.19, 5.20, 5.21, 5.27

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Recap: Huffman Coding

- Symbol code
- Unique decodeability and suffix code
- Kraft inequality and its proof
- Huffman coding

Lecture 6: Noisy Channel Coding

- Noisy channel
- Useful model channels
- Channel coding theorem



- Entropy and mutual information

Review

- Entropy and mutual information
- Noisy world

Review

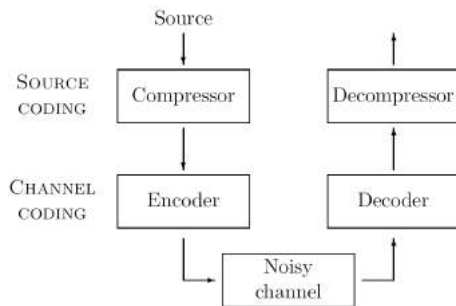
- Entropy and mutual information
- Noisy world
- Lossy source coding

Review

- Entropy and mutual information
- Noisy world
- Lossy source coding
- Channel capacity of BSC

Review

- Entropy and mutual information
- Noisy world
- Lossy source coding
- Channel capacity of BSC



system model

A simple example

The joint distribution of X and Y is

$P(x, y)$		x				$P(y)$
		1	2	3	4	
y	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

A simple example

The joint distribution of X and Y is

$P(x, y)$		x				$P(y)$
		1	2	3	4	
y	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

The conditional entropy of X given each y

$P(x y)$		x				$H(X y)/\text{bits}$
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

A simple example

The joint distribution of X and Y is

$P(x, y)$		x				$P(y)$
		1	2	3	4	
y	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

The conditional entropy of X given each y

$P(x y)$		x				$H(X y)/\text{bits}$
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

Channel: fixed $P(Y|X)$

DMC

A **discrete memoryless channel** Q is characterized by an input alphabet \mathcal{A}_X , an output alphabet \mathcal{A}_Y , and a set of conditional probability distributions $P(y|x)$, one for each $x \in \mathcal{A}_X$. These transition probabilities may be written in a matrix

$$Q_{j|i} = P(y = b_j | x = a_i).$$

DMC

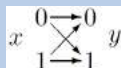
A **discrete memoryless channel** Q is characterized by an input alphabet \mathcal{A}_X , an output alphabet \mathcal{A}_Y , and a set of conditional probability distributions $P(y|x)$, one for each $x \in \mathcal{A}_X$. These transition probabilities may be written in a matrix

$$Q_{j|i} = P(y = b_j | x = a_i).$$

Channel may have **memory**, like tapes

Useful model channels

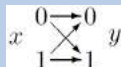
Binary symmetric channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, 1\}$.



$$\begin{aligned} P(y=0 | x=0) &= 1 - f; & P(y=0 | x=1) &= f; \\ P(y=1 | x=0) &= f; & P(y=1 | x=1) &= 1 - f. \end{aligned}$$

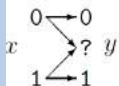
Useful model channels

Binary symmetric channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, 1\}$.



$$\begin{aligned} P(y=0 | x=0) &= 1-f; & P(y=0 | x=1) &= f; \\ P(y=1 | x=0) &= f; & P(y=1 | x=1) &= 1-f. \end{aligned}$$

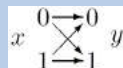
Binary erasure channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, ?, 1\}$.



$$\begin{aligned} P(y=0 | x=0) &= 1-f; & P(y=0 | x=1) &= 0; \\ P(y=? | x=0) &= f; & P(y=? | x=1) &= f; \\ P(y=1 | x=0) &= 0; & P(y=1 | x=1) &= 1-f. \end{aligned}$$

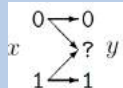
Useful model channels

Binary symmetric channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, 1\}$.



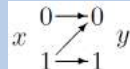
$$\begin{aligned} P(y=0 | x=0) &= 1 - f; & P(y=0 | x=1) &= f; \\ P(y=1 | x=0) &= f; & P(y=1 | x=1) &= 1 - f. \end{aligned}$$

Binary erasure channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, ?, 1\}$.



$$\begin{aligned} P(y=0 | x=0) &= 1 - f; & P(y=0 | x=1) &= 0; \\ P(y=? | x=0) &= f; & P(y=? | x=1) &= f; \\ P(y=1 | x=0) &= 0; & P(y=1 | x=1) &= 1 - f. \end{aligned}$$

Z channel. $\mathcal{A}_x = \{0, 1\}$. $\mathcal{A}_y = \{0, 1\}$.



$$\begin{aligned} P(y=0 | x=0) &= 1; & P(y=0 | x=1) &= f; \\ P(y=1 | x=0) &= 0; & P(y=1 | x=1) &= 1 - f. \end{aligned}$$

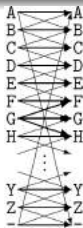




Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters $\{ A, B, \dots, Z, - \}$. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability $1/3$ each; when the input is C, the output is B, C or D; and so forth, with the final letter '-' adjacent to the first letter A.



Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters $\{ A, B, \dots, Z, - \}$. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability $1/3$ each; when the input is C, the output is B, C or D; and so forth, with the final letter '-' adjacent to the first letter A.



$$\begin{aligned}
 & \vdots \\
 P(y=F \mid x=G) &= 1/3; \\
 P(y=G \mid x=G) &= 1/3; \\
 P(y=H \mid x=G) &= 1/3; \\
 & \vdots
 \end{aligned}$$

The capacity of a channel Q is:

$$C(Q) = \max_{\mathcal{P}_X} I(X; Y).$$

The distribution \mathcal{P}_X that achieves the maximum is called the optimal input distribution, denoted by \mathcal{P}_X^* . [There may be multiple optimal input distributions achieving the same value of $I(X; Y)$.]

Real channels and Gaussian Channel

Consider a physical (electrical, say) channel with inputs and outputs that are continuous in time. We put in $x(t)$, and out comes $y(t) = x(t) + n(t)$. Our transmission has a power cost. The average power of a transmission of length T may be constrained thus

$$\int_0^T [x(t)]^2 / T dt \leq P.$$

Real channels and Gaussian Channel

Consider a physical (electrical, say) channel with inputs and outputs that are continuous in time. We put in $x(t)$, and out comes $y(t) = x(t) + n(t)$. Our transmission has a power cost. The average power of a transmission of length T may be constrained thus

$$\int_0^T [x(t)]^2 / T dt \leq P.$$

Gaussian Channel

The Gaussian channel has a real input x and a real output y . The conditional distribution of y given x is a Gaussian distribution:

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y-x)^2/2\sigma^2)$$

This channel is sometimes called the additive white Gaussian noise (AWGN) channel.

Capacity of Gaussian Channels

For a Gaussian channel with power constraint ν and variance of noise σ^2 , its channel capacity is

$$C = \max_{X: \text{Var}(X) \leq \nu} I(X; X + Z) = \frac{1}{2} \log\left(1 + \frac{\nu}{\sigma^2}\right)$$

signal to noise ration: $\frac{\nu}{\sigma^2}$.

Capacity of Gaussian Channels

For a Gaussian channel with power constraint ν and variance of noise σ^2 , its channel capacity is

$$C = \max_{X: \text{Var}(X) \leq \nu} I(X; X + Z) = \frac{1}{2} \log\left(1 + \frac{\nu}{\sigma^2}\right)$$

signal to noise ration: $\frac{\nu}{\sigma^2}$.

If we fixed ν , $C(Z)$ is minimized iff Z is gaussian.

Capacity of Gaussian Channels

For a Gaussian channel with power constraint ν and variance of noise σ^2 , its channel capacity is

$$C = \max_{X: \text{Var}(X) \leq \nu} I(X; X + Z) = \frac{1}{2} \log\left(1 + \frac{\nu}{\sigma^2}\right)$$

signal to noise ration: $\frac{\nu}{\sigma^2}$.

If we fixed ν , $C(Z)$ is minimized iff Z is gaussian.

Entropy Power inequality (Shannon 1948) For any independent random variables X and Y

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}$$

Equality holds iff X, Y are gaussian.

Reading: Ch. 8, Ch. 9, Ch. 10, Ch. 11 (David MacKay)

Exercise

8.5, 8.7, 9.2, 9.4, 9.15, 9.17, 10.8, 10.9, 10.12

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

- Bayesian Inference
- Occam's razor
- Monte Carlo Methods
- Ising model

Discussion: Coin Prediction

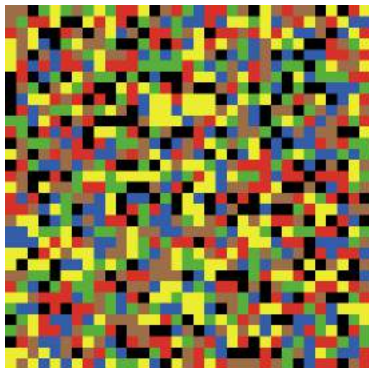
Toss a coin seven times, we obtain that: TTFFTTT. What is the result of next toss.

Discussion: Coin Prediction

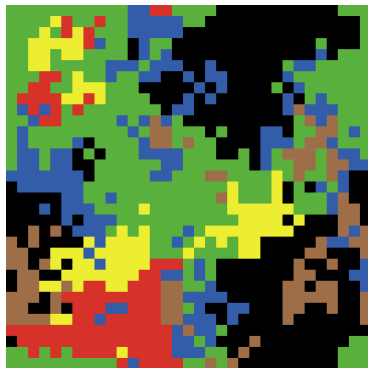
Toss a coin seven times, we obtain that: TTFFTTT. What is the result of next toss.

Given a sequence, -1, 3, 7, 11, what is the next number.

Discussion: Ising model



(a) Initial



(b) Final

The rules of probability ensure that if two people make the same assumptions and receive the same data then they will draw identical conclusions.

The rules of probability ensure that if two people make the same assumptions and receive the same data then they will draw identical conclusions.

you cannot do inference without making assumptions.

Forward probabilities

An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, N times.

- (a) What is the probability distribution of the number of times a black ball is drawn, n_B ?
- (b) What is the expectation of n_B ? What is the variance of n_B ? What is the standard deviation of n_B ? Give numerical answers for the cases $N = 5$ and $N = 400$, when $B = 2$ and $K = 10$.

Forward probabilities

An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, N times.

- (a) What is the probability distribution of the number of times a black ball is drawn, n_B ?
- (b) What is the expectation of n_B ? What is the variance of n_B ? What is the standard deviation of n_B ? Give numerical answers for the cases $N = 5$ and $N = 400$, when $B = 2$ and $K = 10$.

Forward probability problems involve a generative model that describes a process that is assumed to give rise to some data; the task is to compute the probability distribution or expectation of some quantity that depends on the data.

Inverse probabilities

An urn contains K balls, of which B are black and $W = K - B$ are white. We define the fraction $f_B := B/K$. Fred draws N times from the urn, obtaining n_B blacks, and computes the quantity

$$z = \frac{(n_B - f_B N)^2}{N f_B (1 - f_B)}$$

What is the expectation of z ? In the case $N = 5$ and $f_B = 1/5$, what is the probability distribution of z ? What is the probability that $z \leq 1$?

Inverse probabilities

An urn contains K balls, of which B are black and $W = K - B$ are white. We define the fraction $f_B := B/K$. Fred draws N times from the urn, obtaining n_B blacks, and computes the quantity

$$z = \frac{(n_B - f_B N)^2}{N f_B (1 - f_B)}$$

What is the expectation of z ? In the case $N = 5$ and $f_B = 1/5$, what is the probability distribution of z ? What is the probability that $z \leq 1$?

Like forward probability problems, inverse probability problems involve a generative model of a process, but instead of computing the probability distribution of some quantity produced by the process, we compute the conditional probability of one or more of the unobserved variables in the process, given the observed variables. This invariably requires the use of Bayes' theorem.

We need to compute:

$$P(u, n_B|N) = \frac{P(u)P(n_B|u, N)}{P(n_B|N)}$$

We call the marginal probability $P(u)$ the prior probability of u , and $P(n_B, u|N)$ is called the likelihood of u . It is important to note that the terms likelihood and probability are not synonyms. The quantity $P(n_B, u|N)$ is a function of both n_B and u . For fixed u , $P(n_B, u; N)$ defines a probability over n_B . For fixed n_B , $P(n_B, u; N)$ defines the likelihood of u .

Never say 'the likelihood of the data'. Always say 'the likelihood of the parameters'. The likelihood function is not a probability distribution.

The conditional probability $P(u|n_B, N)$ is called the posterior probability of u given n_B . The normalizing constant $P(n_B, N)$ has no u -dependence so its value is not important if we simply wish to evaluate the relative probabilities of the alternative hypotheses u . However, in most data-modelling problems of any complexity, this quantity becomes important, and it is given various names: $P(n_B, N)$ is known as the evidence or the marginal likelihood.

If θ denotes the unknown parameters, D denotes the data, and H denotes the overall hypothesis space, the general equation:

$$P(\theta|D, H) = \frac{P(D|\theta, H)P(\theta|H)}{P(D|H)}$$

is written:

$$\text{posterior} = \text{likelihood} \times \text{prior} / \text{evidence}$$

The likelihood principles

Q1: Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

Q2: Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the urn is urn A?

The likelihood principles

Q1: Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

Q2: Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the urn is urn A?

Likelihood principle

The likelihood principle: given a generative model for data d given parameters θ , $P(d|\theta)$, and having observed a particular outcome d_1 , all inferences and predictions should depend only on the function $P(d_1|\theta)$.

Reading: Ch. 2.2-2.3, Ch. 9 (David MacKay)

CS258: Information Theory

Fan Cheng

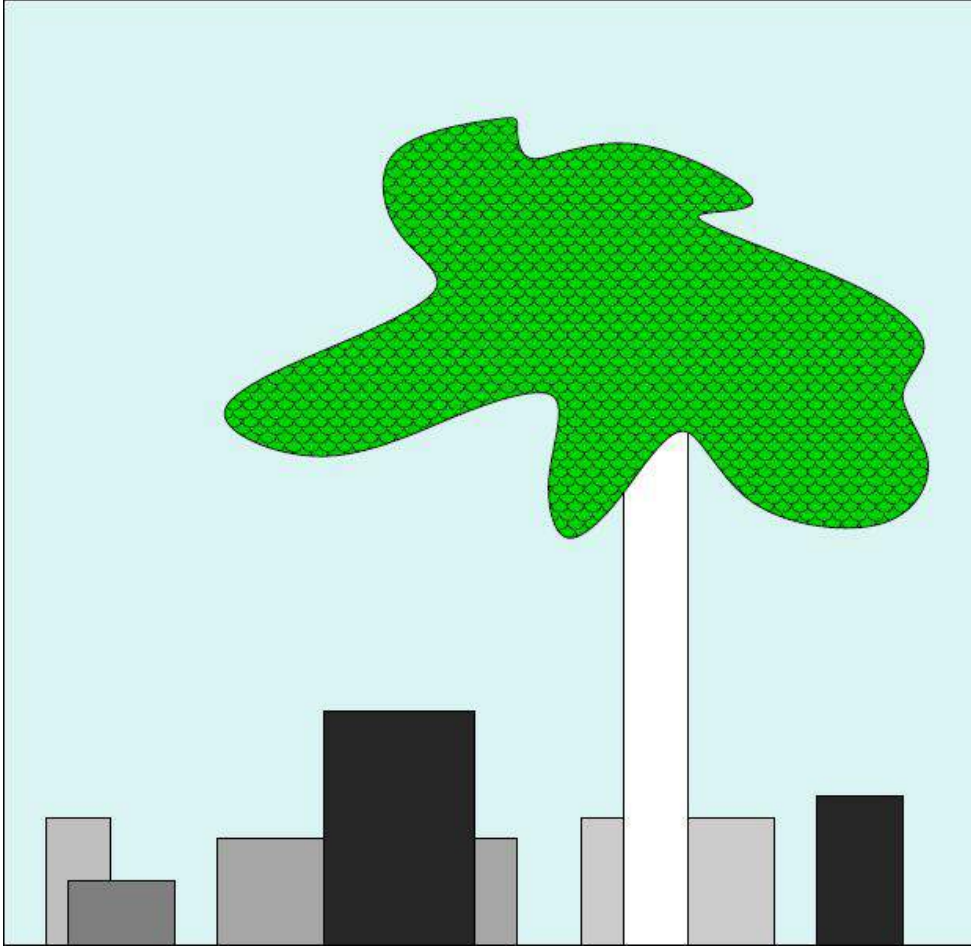


Spring, 2018. chengfan@sjtu.edu.cn

Lecture 8: Model Comparison and Simulation

- Model comparison
- Occam's Razor
- Monte Carlo Methods

Occam's Razor

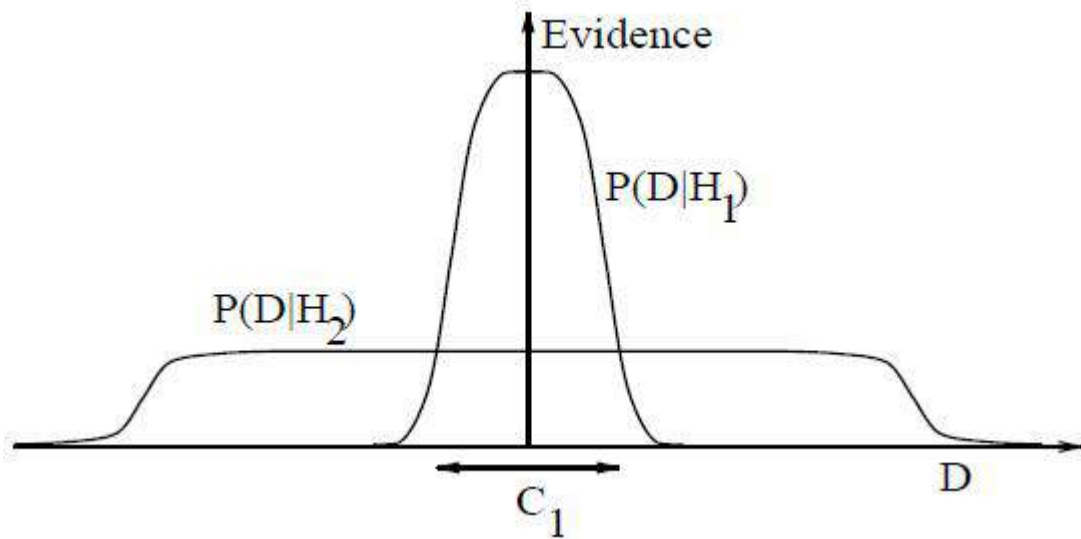


How many boxes in the picture?

Intuition: Accept the simplest explanation that fits the data

A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data

Model Comparison



Evaluate the plausibility of two alternative theories H_1 and H_2 in the light of data D .

$$\frac{P(\mathcal{H}_1 | D)}{P(\mathcal{H}_2 | D)} = \frac{P(\mathcal{H}_1) P(D | \mathcal{H}_1)}{P(\mathcal{H}_2) P(D | \mathcal{H}_2)}$$

Occam's razor:

- H_1 is a simpler model than H_2 .
- Complex models, by their nature, are capable of making a greater variety of predictions.
- H_2 must spread its predictive probability $P(D|H_2)$ more thinly over the data space than H_1 . Thus, in the case the simpler H_1 will turn out more probable than H_2 , without our having to express any subjective dislike for complex models.

Occam's Razor: An Example

- Question: Here is a sequence of numbers:

-1, 3, 7, 11.

The task is to predict the next two numbers, and infer the underlying process that gave rise to this sequence.

- Two general theories:
 - H_a – the sequence is an arithmetic progression, 'add n ', where n is an integer.
 - H_c – the sequence is generated by a cubic function of the form $x \rightarrow cx^3 + dx^2 + e$, where c , d and e are fractions. ($c = -1/11, d = 9/11, e = 23/11$)
- H_a depends on the added integer n , and the first number in the sequence. Let us say that these numbers could each have been anywhere between -50 and 50. Then

$$P(D|H_a) = \frac{1}{101} \frac{1}{101}$$

$$\begin{aligned} P(D|\mathcal{H}_c) &= \left(\frac{1}{101}\right) \left(\frac{4}{101} \frac{1}{50}\right) \left(\frac{4}{101} \frac{1}{50}\right) \left(\frac{2}{101} \frac{1}{50}\right) \\ &= 0.00000000000025 = 2.5 \times 10^{-12}. \end{aligned}$$

Minimum Description Length

- A complementary view of Bayesian model comparison: replacing probabilities of events by the lengths in bits of messages that communicate the events without loss to a receiver.
- $L(X): P(x) = 2^{-L(x)}, L(x) = -\log_2 P(x)$
- The MDL principle (Wallace and Boulton, 1968) : one should prefer models that can communicate the data in the smallest number of bits.

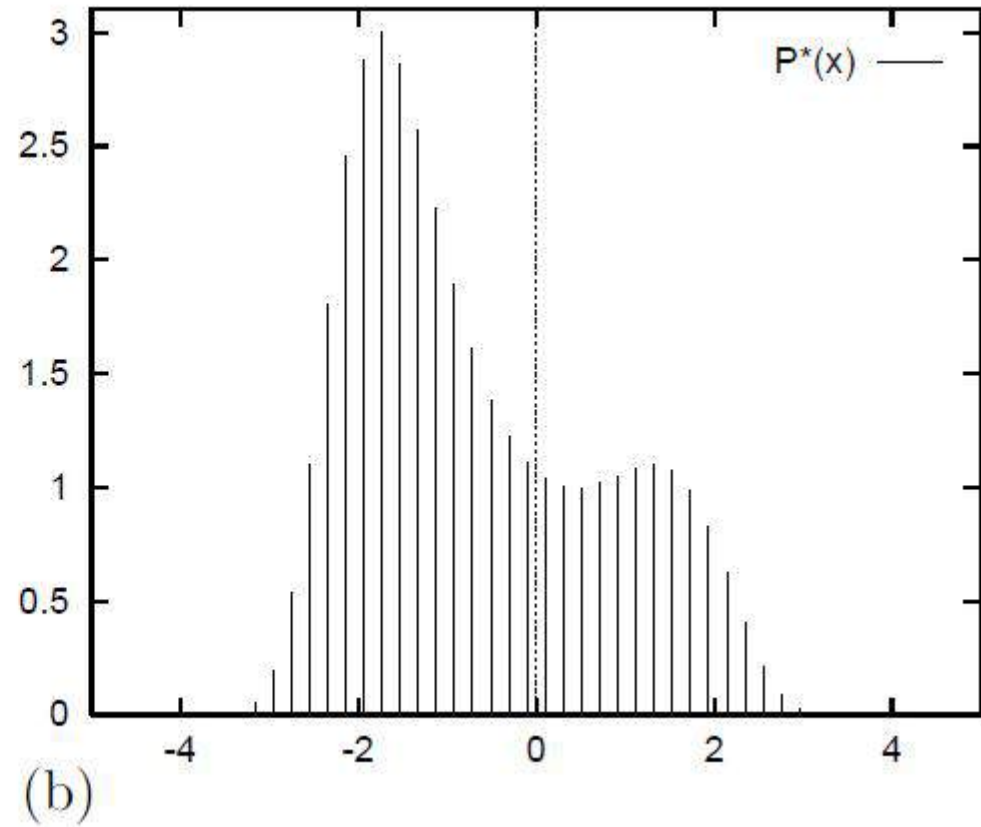
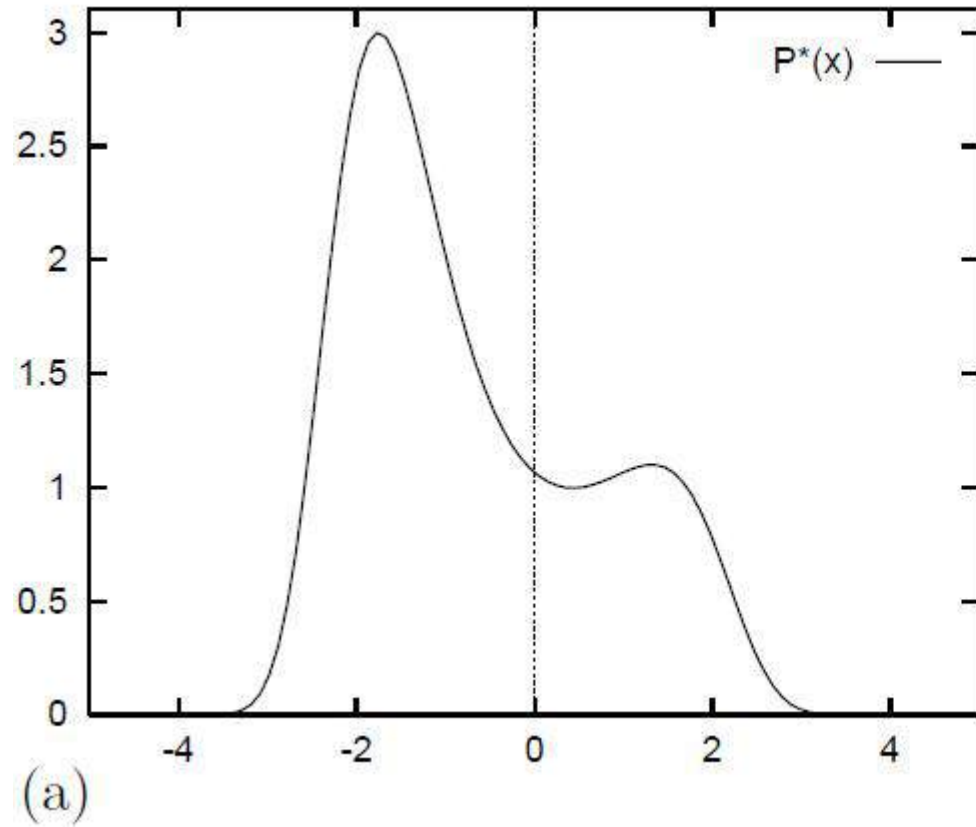
$$L(D, H) = L(H) + L(D|H)$$

- Models with a small number of parameters have only a short parameter block but do not fit the data well, and so the data message (a list of large residuals) is long. As the number of parameters increases, the parameter block lengthens, and the data message becomes shorter.

Monte Carlo Methods

- The aim of MC:
 - To generate samples from a given probability distribution
 - To estimate expectation of functions under given distribution. $\Phi(x) = \langle \phi(x) \rangle$
- History:
 - In the Buffon's needle experiment, in which π can be estimated by dropping needles on a floor made of parallel and equidistant strips
 - In the 1930s, Enrico Fermi first experimented with the Monte Carlo method while studying neutron diffusion, but did not publish anything on it
 - The modern version of the Markov Chain Monte Carlo method was invented in the late 1940s by Stanislaw Ulam, while he was working on nuclear weapons projects at the Los Alamos National Laboratory
 - Immediately after Ulam's breakthrough, John von Neumann understood its importance and programmed the ENIAC computer to carry out Monte Carlo calculations.

Life is hard



$$P(x) = \exp(0.4(x - 0.4)^2 - 0.08x^4)$$

Uniform Sampling

- Solve the problem by drawing random samples uniformly from the state space and evaluating $P(x)$ at those points.
- Take typical set for example, the possibility is $2^H/2^N$
- Thus uniform sampling is utterly useless for the study of Ising models of modest size.
- And in most high-dimensional problems, if the distribution $P(x)$ is not actually uniform, uniform sampling is unlikely to be useful.

Importance Sampling

- Assume $P(x) = \frac{P^*(x)}{Z}$.
- $P(x)$ is too complicated a function for us to be able to sample from it directly. We now assume that we have a simpler density $Q(x)$ from which we can generate samples and which we can evaluate to within a multiplicative constant ($Q(x) = Q^*(x)/Z_Q$).

$$w_r \equiv \frac{P^*(x^{(r)})}{Q^*(x^{(r)})}$$

$$\hat{\Phi} \equiv \frac{\sum_r w_r \phi(x^{(r)})}{\sum_r w_r}.$$

A practical difficulty with importance sampling is that it is hard to estimate how reliable the estimator $\hat{\Phi}$ is.

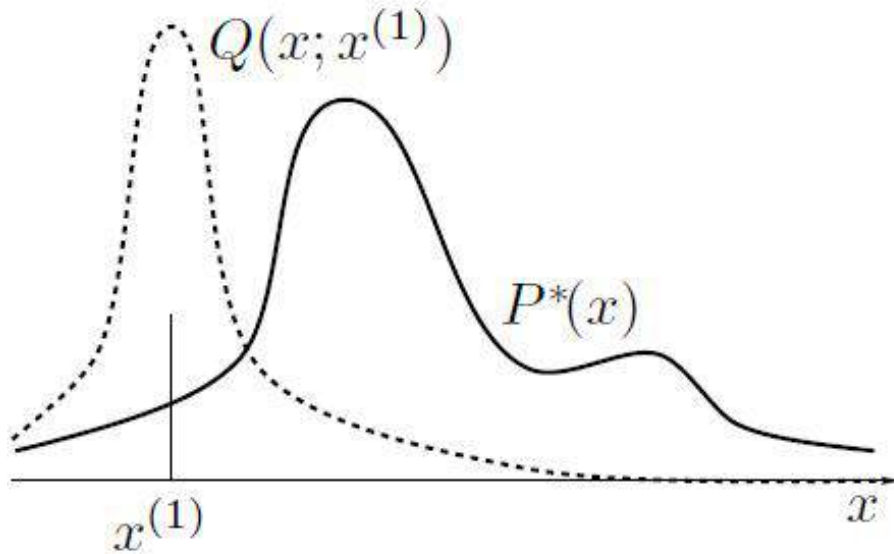
Rejection Sampling

- Assume that we know the value of a constant c such that

$$cQ^*(x) > P^*(x), \text{ for all } x$$

- We generate two random numbers. The first, x , is generated from the proposal density $Q(x)$. We then evaluate $cQ^*(x)$ and generate a uniformly distributed random variable u from the interval $[0, cQ^*(x)]$. These two random numbers can be viewed as selecting a point in the two-dimensional plane.
- We now evaluate $P^*(x)$ and accept or reject the sample x by comparing the value of u with the value of $P^*(x)$. If $u > P^*(x)$ then x is rejected; otherwise it is accepted.

Metropolis-Hasting method



Alg: The density $Q(x'; x^{(t)})$ might be simple distribution
An tentative x' is generated from the proposed density $Q(x'; x^{(t)})$. To Decide whether to accept the new state, we compute ,

$$a = \frac{P^*(x')}{P^*(x^{(t)})} \frac{Q(x^{(t)}; x')}{Q(x'; x^{(t)})}$$

If $a \geq 1$ then the new state is accepted.

Otherwise, the new state is accepted with probability a .

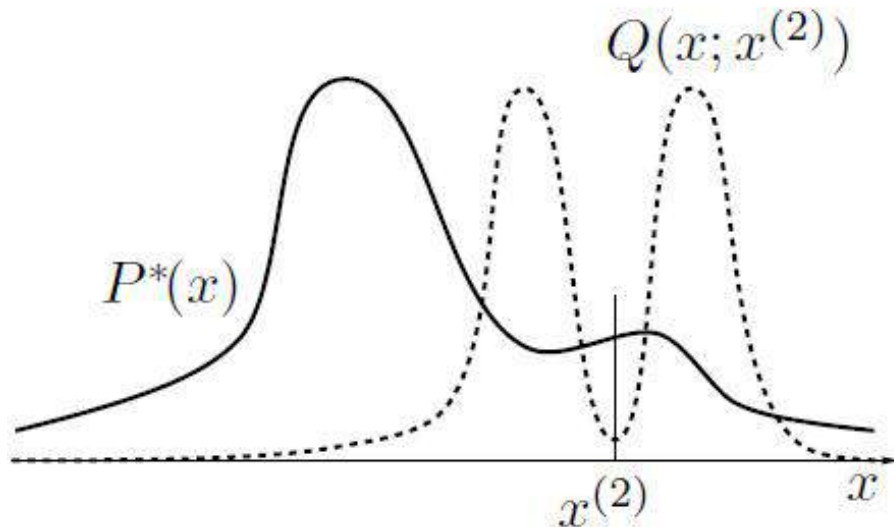
If the step is accepted, we set $x^{(t+1)} = x'$.

If the step is rejected, then we set $x^{(t+1)} = x^{(t)}$.

Nicholas Constantine Metropolis

Conclusion: For any positive Q , as $t \rightarrow \infty$, the probability distribution of $x^{(t)}$ tends to $P^*(x)/Z$.

An example of Markov chain Monte Carlo method



Gibbs Sampling

- Gibbs sampling can be viewed as a Metropolis method in which a sequence of proposal distributions Q are defined in terms of the conditional distributions of the joint distribution $P(\mathbf{x})$.

$$\begin{aligned}x_1^{(t+1)} &\sim P(x_1 \mid x_2^{(t)}, x_3^{(t)}, \dots, x_K^{(t)}) \\x_2^{(t+1)} &\sim P(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \dots, x_K^{(t)}) \\x_3^{(t+1)} &\sim P(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_K^{(t)})\end{aligned}$$

Reading

- Ch. 28, Ch. 29
- Exercise: 28.1, 28.2, 29.3, 29.4, 29.13

CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

Metropolis Algorithm

- Start with any initial value θ_0 satisfying $f(\theta_0) > 0$
- Using current θ value, sample a candidate point θ^* from some jumping distribution $q(\theta_1, \theta_2)$, which is the probability of returning a value of θ_2 given a previous value of θ_1 . This distribution is also referred to as the proposal or candidate-generating distribution. The only restriction on the jump density in the Metropolis algorithm is that it is symmetric, i.e., $q(\theta_1, \theta_2) = q(\theta_2, \theta_1)$.
- Given the candidate point θ^* , calculate the ratio of the density at the candidate (θ^*) and current (θ_{t-1}) points,

$$\alpha = \frac{p(\theta^*)}{p(\theta_{t-1})} = \frac{f(\theta^*)}{f(\theta_{t-1})}$$

- If the jump increases the density ($\alpha > 1$), accept the candidate point (set $\theta_t = \theta^*$) and return to step 2. If the jump decreases the density ($\alpha < 1$), then with probability α accept the candidate point, else reject it and return to step 2.

Metropolis Algorithm

Q_1 : show it is a special case of Metropolis-Hasting algorithms

Q_2 : Define

$$Pr(x \rightarrow y) = q(x, y)\alpha(x, y),$$

show that

$$Pr(x \rightarrow y)p(x) = Pr(y \rightarrow x)p(y)$$

Discuss: $q(x, y)p(x) \Rightarrow < q(y, x)p(y)$

Gibbs is a special case of MH

$$\alpha(x_n^{cand}, x_{-n}^{(i-1)} | x_n^{(i-1)}, x_{-n}^{(i-1)}) = 1$$

Proof.

$$\min\left\{1, \frac{q(x_n^{(i-1)}, x_{-n}^{(i-1)} | x_n^{cand}, x_{-n}^{(i-1)}) p(x_n^{cand}, x_{-n}^{(i-1)})}{q(x_n^{cand}, x_{-n}^{(i-1)} | x_n^{(i-1)}, x_{-n}^{(i-1)}) p(x_n^{(i-1)}, x_{-n}^{(i-1)})}\right\} \quad (1)$$

$$= \min\left\{1, \frac{p(x_n^{(i-1)} | x_{-n}^{(i-1)}) p(x_n^{cand}, x_{-n}^{(i-1)})}{p(x_n^{cand} | x_{-n}^{(i-1)}) p(x_n^{(i-1)}, x_{-n}^{(i-1)})}\right\} \quad (2)$$

$$= \min\left\{1, \frac{p(x_n^{(i-1)} | x_{-n}^{(i-1)}) p(x_n^{cand} | x_{-n}^{(i-1)}) p(x_{-n}^{(i-1)})}{p(x_n^{cand} | x_{-n}^{(i-1)}) p(x_n^{(i-1)} | x_{-n}^{(i-1)}) p(x_{-n}^{(i-1)})}\right\} \quad (3)$$

$$= 1 \quad (4)$$



CS258: Information Theory

Fan Cheng



Spring, 2018. chengfan@sjtu.edu.cn

1. Maximum entropy

The normal maximizes the entropy for a given variance

If $\text{Var}(X) = a$ is fixed, then

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

1. Maximum entropy

The normal maximizes the entropy for a given variance

If $\text{Var}(X) = a$ is fixed, then

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

Proof.

$D(f||g) \geq 0$, where $g \sim \mathcal{N}(0, a)$

2. Entropy of a disjoint mixture

Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \dots, m\}$ and $\mathcal{X}_2 = \{m+1, \dots, n\}$. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha, \\ X_2, & \text{with probability } 1 - \alpha. \end{cases}$$

- Find $H(X)$ in terms of $H(X_1)$, $H(X_2)$, and α .

2. Entropy of a disjoint mixture

Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \dots, m\}$ and $\mathcal{X}_2 = \{m+1, \dots, n\}$. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha, \\ X_2, & \text{with probability } 1 - \alpha. \end{cases}$$

- Find $H(X)$ in terms of $H(X_1)$, $H(X_2)$, and α .

By definition.

Question: what if $\mathcal{X}_1 = \mathcal{X}_2$?

3. Entropy of a sum

Let X and Y be random variables that take on values x_1, x_2, \dots, x_r and y_1, y_2, \dots, y_s , respectively. Let $Z = X + Y$.

- (a) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$. Thus, the addition of independent random variables adds uncertainty.
- (b) Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.
- (c) Under what conditions does $H(Z) = H(X) + H(Y)$?

4. Entropy and pairwise independence

Let X, Y, Z be three binary $Bernoulli(\frac{1}{2})$ random variables that are pairwise independent; that is, $I(X; Y) = I(X; Z) = I(Y; Z) = 0$.

- (a) Under this constraint, what is the minimum value for $H(X, Y, Z)$?
- (b) Give an example achieving this minimum.

4. Entropy and pairwise independence

Let X, Y, Z be three binary $Bernoulli(\frac{1}{2})$ random variables that are pairwise independent; that is, $I(X; Y) = I(X; Z) = I(Y; Z) = 0$.

- (a) Under this constraint, what is the minimum value for $H(X, Y, Z)$?
- (b) Give an example achieving this minimum.

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

5. Subset inequality

Prove that

$$\frac{1}{2}[H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)] \geq H(X_1, X_2, X_3)$$

5. Subset inequality

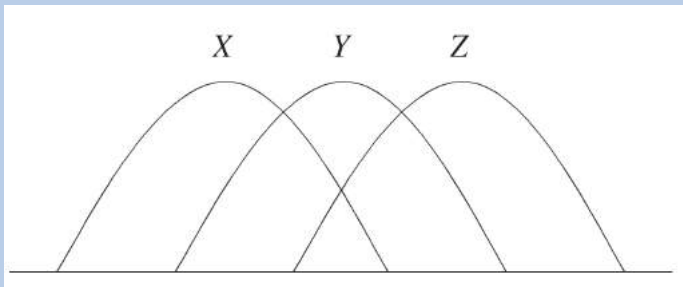
Prove that

$$\frac{1}{2}[H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)] \geq H(X_1, X_2, X_3)$$

Information diagram

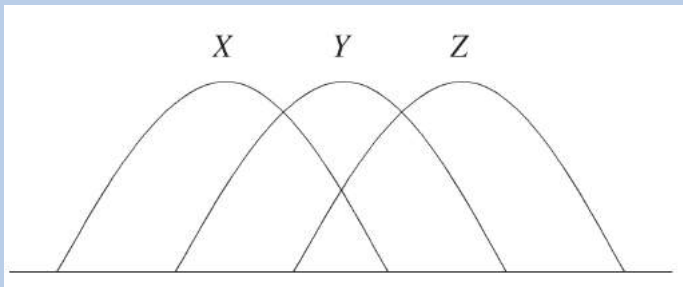
6. Information diagram for Markov chain

Verify that if $X \rightarrow Y \rightarrow Z$, then their information diagram can be simplified as



6. Information diagram for Markov chain

Verify that if $X \rightarrow Y \rightarrow Z$, then their information diagram can be simplified as



The results can be extended to $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$

7. Implication and Markov chain

- (a) Prove that under the constraint that $X \rightarrow Y \rightarrow Z$ forms a Markov chain, $X \perp Y|Z$ and $X \perp Z$ imply $X \perp Y$.
- (b) Prove that the implication in (a) continues to be valid without the Markov chain constraint.
- (c) Prove that $Y \perp Z|T$ implies $Y \perp Z|(X, T)$ conditioning on $X \rightarrow Y \rightarrow Z \rightarrow T$.

8. Markov chain with 4 random variables

Let $X \rightarrow Y \rightarrow Z \rightarrow T$ form a Markov chain. Determine which of the following inequalities always hold:

- (i) $I(X; T) + I(Y; Z) \geq I(X; Z) + I(Y; T)$
- (ii) $I(X; T) + I(Y; Z) \geq I(X; Y) + I(Z; T)$
- (iii) $I(X; Y) + I(Z; T) \geq I(X; Z) + I(Y; T)$

9. Imperfect secrecy theorem

Let X be the plain text, Y be the cipher text, and Z be the key in a secret key cryptosystem. Since X can be recovered from Y and Z , we have

$$H(X|Y, Z) = 0$$

We will show that this constraint implies

$$I(X; Y) \geq H(X) - H(Z)$$



Exercise 2

Exercise 1.3.[3, p.16] (a) Show that the probability of error of R_N , the repetition code with N repetitions, is

$$p_b = \sum_{n=(N+1)/2}^N \binom{N}{n} f^n (1-f)^{N-n}, \quad (1.24)$$

for odd N .

- (b) Assuming $f = 0.1$, which of the terms in this sum is the biggest? How much bigger is it than the second-biggest term?
- (c) Use Stirling's approximation (p.2) to approximate the $\binom{N}{n}$ in the largest term, and find, approximately, the probability of error of the repetition code with N repetitions.
- (d) Assuming $f = 0.1$, find how many repetitions are required to get the probability of error down to 10^{-15} . [Answer: about 60.]



- Exercise 1.6.^[2, p.17] (a) Calculate the probability of block error p_B of the $(7, 4)$ Hamming code as a function of the noise level f and show that to leading order it goes as $21f^2$.
- (b) ^[3] Show that to leading order the probability of bit error p_b goes as $9f^2$.


(b) The probability of bit error of the Hamming code is smaller than the probability of block error because a block error rarely corrupts all bits in the decoded block. The leading-order behaviour is found by considering the outcome in the most probable case where the noise vector has weight two. The decoder will erroneously flip a *third* bit, so that the modified received vector (of length 7) differs in three bits from the transmitted vector. That means, if we average over all seven bits, the probability that a randomly chosen bit is flipped is $3/7$ times the block error probability, to leading order. Now, what we really care about is the probability that

- ▷ Exercise 2.28.^[2, p.45] A random variable $x \in \{0, 1, 2, 3\}$ is selected by flipping a bent coin with bias f to determine whether the outcome is in $\{0, 1\}$ or $\{2, 3\}$; then either flipping a second bent coin with bias g or a third bent coin with bias h respectively. Write down the probability distribution of x . Use the decomposability of the entropy (2.44) to find the entropy of X . [Notice how compact an expression is obtained if you make use of the binary entropy function $H_2(x)$, compared with writing out the four-term entropy explicitly.] Find the derivative of $H(X)$ with respect to f . [Hint: $dH_2(x)/dx = \log((1-x)/x)$.]
- ▷ Exercise 2.29.^[2, p.45] An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, \dots\}$, the number of flips? Repeat the calculation for the case of a biased coin with probability f of coming up heads. [Hint: solve the problem both directly and by using the decomposability of the entropy (2.43).]

▷ Exercise 4.9.^[1] While some people, when they first encounter the weighing problem with 12 balls and the three-outcome balance (exercise 4.1 (p.66)), think that weighing six balls against six balls is a good first weighing, others say ‘no, weighing six against six conveys *no* information at all’. Explain to the second group why they are both right and wrong. Compute the information gained about *which is the odd ball*, and the information gained about *which is the odd ball and whether it is heavy or light*.

▷ Exercise 4.10.^[2] Solve the weighing problem for the case where there are 39 balls of which one is known to be odd.


- 
- ▷ Exercise 4.11.^[2] You are given 16 balls, all of which are equal in weight except for one that is either heavier or lighter. You are also given a bizarre two-pan balance that can report only two outcomes: ‘the two sides balance’ or ‘the two sides do not balance’. Design a strategy to determine which is the odd ball in as few uses of the balance as possible.
- ▷ Exercise 4.12.^[2] You have a two-pan balance; your job is to weigh out bags of flour with integer weights 1 to 40 pounds inclusive. How many weights do you need? [You are allowed to put weights on either pan. You’re only allowed to put one flour bag on the balance at a time.]
- 



(4.12.) By symmetry, if we could weigh x pounds of flour, we can also weigh $-x$. The range is -40 to 40 .


For x pound, we have $x = \sum_{i=1}^n a_i g_i$, where g_i 's are the weights and $a_i \in \{-1, 0, 1\}$.

Hence $3^n \geq 81$





Exercise 3



Exercise 8.5.^[4] The ‘entropy distance’ between two random variables can be defined to be the difference between their joint entropy and their mutual information:

$$D_H(X, Y) \equiv H(X, Y) - I(X; Y). \quad (8.12)$$

Prove that the entropy distance satisfies the axioms for a distance – $D_H(X, Y) \geq 0$, $D_H(X, X) = 0$, $D_H(X, Y) = D_H(Y, X)$, and $D_H(X, Z) \leq D_H(X, Y) + D_H(Y, Z)$. [Incidentally, we are unlikely to see $D_H(X, Y)$ again but it is a good function on which to practise inequality-proving.]

Prove it by information diagram.


Exercise 8.7.^[2, p.143] Consider the ensemble XYZ in which $\mathcal{A}_X = \mathcal{A}_Y = \mathcal{A}_Z = \{0, 1\}$, x and y are independent with $\mathcal{P}_X = \{p, 1-p\}$ and $\mathcal{P}_Y = \{q, 1-q\}$ and

$$z = (x + y) \bmod 2. \quad (8.13)$$

- (a) If $q = 1/2$, what is \mathcal{P}_Z ? What is $I(Z; X)$?
- (b) For general p and q , what is \mathcal{P}_Z ? What is $I(Z; X)$? Notice that this ensemble is related to the binary symmetric channel, with $x =$ input, $y =$ noise, and $z =$ output.

(a). By definition, compute the distribution of $P(Z)$.

(b). $I(Z; X) = H(Z) - H(Z|X) = H(Z) - H(X+Y|X) = H(Z) - H(Y|X) = H(Z) - H(Y)$



Exercise 9.2. [1, p.157] Now assume we observe $y = 0$. Compute the probability of $x = 1$ given $y = 0$.

Exercise 9.4. [1, p.157] Alternatively, assume we observe $y = 0$. Compute $P(x = 1 \mid y = 0)$.

By definition, it is trivial to compute these two probabilities.

Exercise 9.15. [3, p.159] Refer back to the computation of the capacity of the Z channel with $f = 0.15$.

- (a) Why is p_1^* less than 0.5? One could argue that it is good to favour the 0 input, since it is transmitted without error – and also argue that it is good to favour the 1 input, since it often gives rise to the highly prized 1 output, which allows certain identification of the input! Try to make a convincing argument.
- (b) In the case of general f , show that the optimal input distribution is

$$p_1^* = \frac{1/(1-f)}{1 + 2^{(H_2(f)/(1-f))}}. \quad (9.19)$$

- (c) What happens to p_1^* if the noise level f is very close to 1?

How to compute the capacity of Z-channel

Example 9.11. Consider the Z channel with $f=0.15$. Identifying the optimal input distribution is not so straightforward. We evaluate $I(X;Y)$ explicitly for $\mathcal{P}_X = \{p_0, p_1\}$. First, we need to compute $P(y)$. The probability of $y=1$ is easiest to write down:

$$P(y=1) = p_1(1-f). \quad (9.13)$$

Then the mutual information is:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(p_1(1-f)) - (p_0 H_2(0) + p_1 H_2(f)) \\ &= H_2(p_1(1-f)) - p_1 H_2(f). \end{aligned} \quad (9.14)$$

This is a non-trivial function of p_1 , shown in figure 9.3. It is maximized for $f=0.15$ by $p_1^* = 0.445$. We find $C(Q_Z) = 0.685$. Notice the optimal input distribution is not $\{0.5, 0.5\}$. We can communicate slightly more information by using input symbol 0 more frequently than 1.

Solution to exercise 9.15 (p.155). In example 9.11 (p.151) we showed that the mutual information between input and output of the Z channel is

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y | X) \\ &= H_2(p_1(1 - f)) - p_1 H_2(f). \end{aligned} \quad (9.47)$$

We differentiate this expression with respect to p_1 , taking care not to confuse \log_2 with \log_e :

$$\frac{d}{dp_1} I(X; Y) = (1 - f) \log_2 \frac{1 - p_1(1 - f)}{p_1(1 - f)} - H_2(f). \quad (9.48)$$

Setting this derivative to zero and rearranging using skills developed in exercise 2.17 (p.36), we obtain:

$$p_1^*(1 - f) = \frac{1}{1 + 2^{H_2(f)/(1-f)}}, \quad (9.49)$$

so the optimal input distribution is

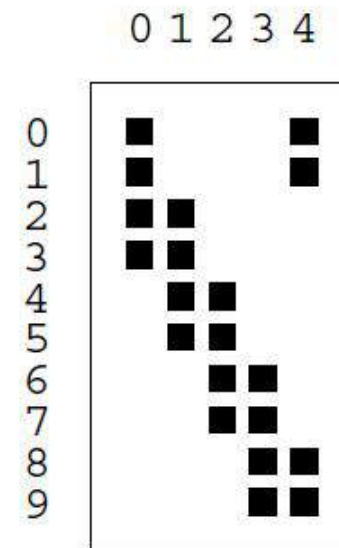
$$p_1^* = \frac{1/(1 - f)}{1 + 2^{(H_2(f)/(1-f))}}. \quad (9.50)$$

The intuition why $p_1^* \leq 0.5$ is because $x = 0$ and $x = 1$ can be treated as two channels, where $x = 0$ is zero error and $x = 1$ is noisy.

Symmetric Channel

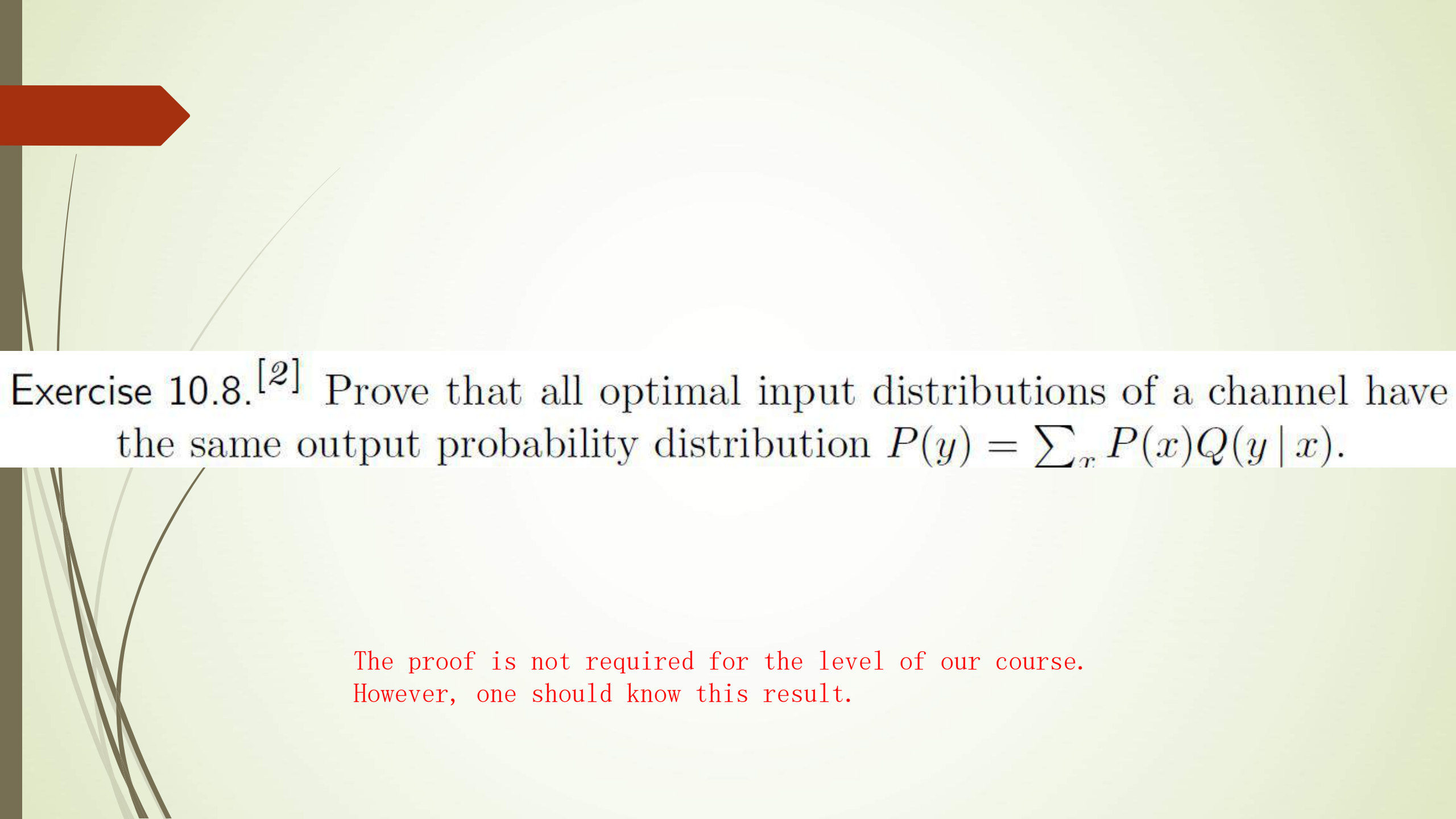
Exercise 9.17.^[2] What is the capacity of the five-input, ten-output channel whose transition probability matrix is

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix}$$



?

(9.20)



Exercise 10.8.^[2] Prove that all optimal input distributions of a channel have the same output probability distribution $P(y) = \sum_x P(x)Q(y | x)$.

The proof is not required for the level of our course.
However, one should know this result.

Example 10.9. This channel

$$\begin{aligned} P(y=0 \mid x=0) &= 0.7; & P(y=0 \mid x=1) &= 0.1; \\ P(y=? \mid x=0) &= 0.2; & P(y=? \mid x=1) &= 0.2; \\ P(y=1 \mid x=0) &= 0.1; & P(y=1 \mid x=1) &= 0.7. \end{aligned} \tag{10.23}$$

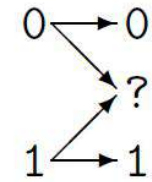
is a symmetric channel because its outputs can be partitioned into $(0, 1)$ and $?$, so that the matrix can be rewritten:

$P(y=0 \mid x=0)$	$= 0.7;$	$P(y=0 \mid x=1)$	$= 0.1;$	(10.24)
$P(y=1 \mid x=0)$	$= 0.1;$	$P(y=1 \mid x=1)$	$= 0.7;$	
$P(y=? \mid x=0)$	$= 0.2;$	$P(y=? \mid x=1)$	$= 0.2.$	

Exercise 10.10.^[2] Prove that for a symmetric channel with any number of inputs, the uniform distribution over the inputs is an optimal input distribution.

Exercise 10.12.^[2] A binary erasure channel with input x and output y has transition probability matrix:

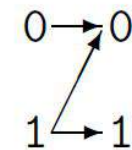
$$\mathbf{Q} = \begin{bmatrix} 1-q & 0 \\ q & q \\ 0 & 1-q \end{bmatrix}$$



Find the *mutual information* $I(X;Y)$ between the input and output for general input distribution $\{p_0, p_1\}$, and show that the *capacity* of this channel is $C = 1 - q$ bits.

A Z channel has transition probability matrix:

$$\mathbf{Q} = \begin{bmatrix} 1 & q \\ 0 & 1-q \end{bmatrix}$$



Show that, using a $(2, 1)$ code, **two** uses of a Z channel can be made to emulate **one** use of an erasure channel, and state the erasure probability of that erasure channel. Hence show that the capacity of the Z channel, C_Z , satisfies $C_Z \geq \frac{1}{2}(1 - q)$ bits.

Explain why the result $C_Z \geq \frac{1}{2}(1 - q)$ is an inequality rather than an equality.

For BEC, the alphabet of output y is $\{0, 1, ?\}$. When $P_X = \{p_0, p_1\}$, the output distribution is $\{p_0(1 - q), p_1(1 - q), q\}$.

$$H(Y) = H_2(q) + (1 - q)H_2(p_0).$$

$$H(Y|X) = p_0H(Y|X = 0) + p_1H(Y|X = 1) = p_0H_2(q) + p_1H_2(q) = H_2(q)$$

$$I(X; Y) = (1 - q)H_2(p_0)$$

We design the following code for Z-Channel. For two uses of z-channel, $\{01, 10\}$, we have a BEC with output alphabet $\{01, 00, 10\}$ and erasure probability q . Hence the channel of BEC is $(1-q)/2$. The channel capacity of Z-Channel is at least $(1-q)/2$.

In the code above, we discard several channels with input like 00, 11. That's why the lower bound is hardly possible to attain.