An Introduction to Information Theory

Frank Guo
ntguojiarui@pku.edu.cn

January 16, 2022

What the Author Would Like to Say

This introduction serves as reviewing materials of the course *Information Theory* by lecturer *Wang Liwei*. However, I do not want to simply list several conclusions or theorems without careful explanations. Hence, I will try to explain the logic behind the theorem so that they would be friendly to readers. In short, I write this introduction in admiration of the lecturer and millions of relevant scholars in the field of computer science. I would like to extend my greatest appreciation to them.

Due to time limitation, I leave out the part of the introduction of probability because I do not think anyone who is now reading this introduction shall lack relevant knowledge. So we will start from the concept of entropy.

1 Entropy

If we denote p_i as the possibility of the occurrence of the i^{th} character and l_i as the length of its code, then the average code length is defined as the expectation of the length of the code:

$$E(l) = \sum_i p_i l_i.$$

And we wish to minimize the average code length when it is uniquely decodable.

On this occasion, we hope that the code is prefix-free. However, this is a full but unnecessary requirement. But we can prove that for every non-prefix-free codes, there *always* exists a type of prefix-free codes which is not worse than it. As a result, we will only take prefix-free codes into consideration in the following parts.

Theorem 1.1 (Kraft Inequality)**.** *Assume that $c_1, \cdots, c_n$ are prefix-free codes and $l_1, \cdots, l_n$ are their lengths. Then*

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

with equality if and only if $l_i (1 \leq i \leq n)$ consist of a full binary tree.

Proof. By induction of the depth of the binary tree. □

It is trivial that we could get a binary tree if l_i satisfies 1.1. Hence, let us take a look at the following question: Assume $p_1, \cdots, p_n \geq 0$. Find out

$$\min_{(l_1, \cdots, l_n)} \sum_{i=1}^{n} p_i l_i$$

where

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

If we replace q_i with 2^{-l_1}, then we want to find out

$$\max \sum_{i=1}^{n} p_i \log q_i,$$

where

$$1 = \sum_{i=1}^{n} p_i \geq \sum_{i=1}^{n} q_i.$$

We already know that

$$\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i.$$

Hence, if we assume that l_i could be any positive number, then $l_i = -\log p_i$ and a lower bound for the source coding is

$$\sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$

(bits).

Definition 1.2 (Entropy)**.** *For random variable X and its probability distribution function $p = (p_1, \cdots, p_n)$, define its entropy as*

$$H(X) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}.$$

Let us take a look at the definition of entropy: As to random variable X and its probability distribution function

$$X = (p_1, p_2, \cdots, p_n = q_1 + q_2).$$

If we define two new random variable

$$Y = \left(\frac{q_1}{q_1 + q_2}, \frac{q_2}{q_1 + q_2} \right)$$

and

$$Z = (p_1, p_2, \cdots, p_{n-1}, q_1, q_2),$$

we can prove that (the addition of entropy):

$$H(X) + p_n H(y) = H(Z).$$

Since we have already get a infimum of the average code length, how to design such "optimal codes"?

According to the theorem above, we already know several properties of optimal codes:

- $p_1 \geq \cdots \geq p_n \Rightarrow l_1 \leq \cdots \leq l_n$.
- By Kraft Inequality, we know that all nodes form a complete binary tree.
- $l_n = l_{n-1}$. Without loss of generality, we can assume that the two nodes are siblings.
- If we merge p_{n-1} and p_n, then the new code $Y = (p_1, \cdots, p_{n-2}, p_{n-1} + p_n)$ shall also be optimal codes.

By the properties above, we conclude that the optimal code shall be Huffman Code.

2 Joint Entropy, Conditional Entropy, Relative Entropy, Mutual Information

2.1 Joint Entropy

Definition 2.1 (Joint Entropy)**.** *Let X, Y be random variable with their joint probability distribution $p(x, y)$, then the joint entropy is defined as*

$$H(X, Y) = -\sum_{i,j} P(X = i, Y = j) \log P(X = i, Y = j)$$

(*bits*).

2.2 Conditional Entropy

Given $Y = j$, the entropy of X is defined as

$$H(X|Y = j) = -\sum_i P(X = i|Y = j) \log P(X = i|Y = j).$$

Inspired by this fact, we will define conditional entropy.

Definition 2.2 (Conditional Entropy)**.** *Let X, Y be random variable with conditional probability distribution $p(x|y)$, then the conditional entropy is defined as*

$$H(X|Y) = \sum_j P(Y = j)H(X|Y = j) = -\sum_{i,j} P(X = i, Y = j) \log P(X = i|Y = j)$$

It is easy to prove that $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(Y|X)$. Hence, by Jensen Inequality,

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

which tells us that $0 \leq H(Y|X) \leq H(Y)$. Moreover, if $H(Y|X) = 0$, then X determines Y; if $H(Y|X) = H(Y)$, then X and Y are independent.

2.3 Mutual Information

We use mutual information to describe how much information Y contains about X.

Definition 2.3 (Mutual Information)**.** *Mutual information of random variable X and Y is defined as*

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

By some computation, we know that

$$\begin{aligned} I(X; Y) &= \sum_{i,j} P(X = i, Y = j) \log \frac{P(X = i, Y = j)}{P(X = i)P(Y = j)} = I(Y; X) \\ &= H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y), \end{aligned}$$

which shows that mutual information is symmetric.

With the definition above, we can define joint entropy, conditional entropy and mutual information with three or more random variables.

Theorem 2.4 (Chain Rule)**.** *Multi-variable entropy has the following property:*

$$H(X_1, \cdots, X_n) = H(X_1)+H(X_2, \cdots, X_n|X_1) = H(X_1)+H(X_2|X_1)+\cdots+H(X_n|X_1, \cdots, X_{n-1}).$$

2.4 Relative Entropy

If we measure the probability distribution function of X to be $P = (p_1, \cdots, p_n)$, then we could design optimal code according to P. However, our measurement could not be exact and the real distribution function might be $Q = (q_1, \cdots, q_n)$. So the error is

$$\sum_i p_i \log \frac{p_i}{q_i}.$$

Definition 2.5 (Relative Entropy, KL-divergence)**.** *Let $P = (p_1, \cdots, p_n)$, $Q = (q_1, \cdots, q_n)$ be two probability distribution function. The relative entropy is defined as*

$$D(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

It is trivial that relative entropy is always non-negative. Also, let $P = (p_1, \cdots, p_n)$ and $U = (\frac{1}{n}, \cdots, \frac{1}{n})$, then relative entropy

$$D(P\|U) = \sum_{i=1}^{n} p_i \log p_i + \log n = \log n - H(P).$$

Hence, this is a trick to define entropy using relative entropy.

Theorem 2.6 (Data Processing Inequality)**.** *Let $X \to Y \to Z$ denote three random variables. If X, Y and Z satisfies Marcov Property, that is: $P(Z|X, Y) = P(Z|Y)$, then $I(X; Z) \leq I(X; Y)$. In other words, the information of X cannot increase during this process.*

3 Entropy Rate

Take the case below into consideration: random variable $X \sim B(1, 0.01)$. The average code length of X is 1 bit. However, the entropy of X is approximately 0(0.07 to be precise). What leads to this inconsistency? How to define the efficiency of the code?

Remember that efficiency shall be defined in ratios. A possible solution to the phenomenon above can be to encode in groups of n each time. Since $\text{ACL} \leq H(X) + 1$, on this occasion we get

$$H(x_1, \cdots, x_n) = 0.07n, \text{ACL} \leq 0.07n + 1.$$

As n increases, the value $\frac{\text{ACL}}{H}$ converges to 1. In this way, we prove that H is really an infimum.

However, this method will trigger delay in time. Also, we potentially assume that $x_1, \cdots, x_n$ are independent, which is unrealistic and not often the case. Nevertheless, if we continue to encode the symbol in groups of t, then

$$\text{ACL(per symbol)} \leq \frac{H(x_1, \cdots, x_t) + 1}{t}.$$

If t is big enough, right hand side is approximately $\frac{H(x_1, \cdots, x_t)}{t}$. If this value converges, then it will be meaningful.

Definition 3.1 (Entropy Rate)**.** *Entropy rate is defined as*

$$\lim_{t \to \infty} \frac{H(x_1, \cdots, x_t)}{t} \tag{1}$$

if it is convergent.

Another way of understanding entropy rate is to focus on the additional information caused by x_t, so it can also be defined as

$$\lim_{t \to \infty} H(x_t | x_1, \cdots, x_{t-1}). \tag{2}$$

By Stolz Theorem, we conclude that 1 and 2 have the same value. Moreover, if X satisfies Marcov Property, then 2 could be written as

$$\lim_{t \to \infty} H(x_t | x_{t-1}).$$

4 Differential Entropy

In this section, we will consider continuous random variables.

If a continuous random variable has probability density function $f(x)$, it cannot be encoded with finite bits. However, we still define its entropy as

$$h(X) = -\int f(x) \log f(x) \mathrm{d}x.$$

This is called the differential entropy of X. But what does it imply? Does

$$H(X_\Delta) = -\sum_i p_i \log p_i$$

converges to $h(X)$?

Trivially $h(X)$ diverges to $+\infty$. However,

$$\log \frac{1}{\Delta} + h(X) = H(X_\Delta).$$

Similarly, define joint entropy as

$$h(X, Y) = -\iint f(x, y) \log f(x, y) \mathrm{d}x \mathrm{d}y.$$

Define conditional entropy as

$$h(X|Y) = -\int \mathrm{d}y \int \mathrm{d}x f(x,y) \log f(x|y).$$

Define mutual information as

$$\begin{aligned}\tilde{I}(X;Y) &= h(X) - h(X|Y) = h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y) = \iint f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \mathrm{d}x\mathrm{d}y.\end{aligned}$$

In this way, we get that $\tilde{I}(X;Y) = I(\tilde{X}_\Delta; \tilde{Y}_\Delta)$.

$$\begin{array}{llll} h(X) & X \longrightarrow & \tilde{X}_\Delta & H(\tilde{X}_\Delta) \\ h(Y) & Y \longrightarrow & \tilde{Y}_\Delta & H(\tilde{Y}_\Delta) \end{array}$$

Define relative entropy as

$$\widetilde{KL}(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

Similarly, $\widetilde{KL}(f\|g) = KL(\tilde{X}_\Delta \| \tilde{Y}_\Delta)$ when $\Delta \to 0$.

5 Kolmogorov Complexity

As to a deterministic variable, what is its minimum description length? Here, minimum description length means the minimum length of a program which is required to output the string. And what Kolmogorov used here is a Turing machine.

Definition 5.1 (Kolmogorov Complexity)**.** *Given a universal Turing machine u, for any string $x \in \{0,1\}^*$, the Kolmogorov complexity of x with respect to u is*

$$K_u(x) = \min_{P:u(P)=x} |P|.$$

In order to compute Kolmogorov complexity, the language of the program shall be given in advance. But, does the language matter?

Theorem 5.2. *$\forall u, u'$, there exists a constant C, such that: $\forall x$,*

$$K_{u'}(x) \le K_u(x) + C(u,u'),$$

where C is independent from x. (Here, C represents the program we need to translate u into u').

However, the problem above is incomputable. Even the halting problem is incomputable.

Theorem 5.3. *Halting problem is incomputable. In other words, let P be the program and I be the input. There is no Turing machine M, such that $M(P,I)$ decides whether P would halt on input I.*

Proof. Let I be the program itself. We define a procedure $U(P,P)$:

Algorithm $U(P,P)$

Input: a program P
Output: 0 or 1

```
1: if M(P,P) = 1 then
2:     return 0;
3: else
4:     while (1) do
5:         endless loop
```

Let us see what will happen when we input U:

- If $M(U,U) = 1$, then U will not halt. On this occasion, $M(U,U) = 0$, implying that U will halt.
- If $M(U,U) = 0$, then U will halt. On this occasion, $M(U,U) = 1$, implying that U will not halt.

On any case we all get a contradiction, so halting problem is incomputable. □

6 Maximum Entropy Principle

We will only list two facts in this section.

Theorem 6.1. *Let random variable X satisfies $EX = 0, DX = 1$. Then when X satisfies normal Gaussian distribution $N(0,1)$,*

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

the entropy of X gets its maximum value.

Theorem 6.2. *Let X be a discrete random variable with only positive integer value. If $EX = \mu$, then when X satisfies geometry distribution*

$$p_i = P(X = i) = \frac{1}{\mu - 1}\left(\frac{\mu - 1}{\mu}\right)^i,$$

its entropy gets its maximum value.

7 Channel Coding, Channel Capacity

7.1 Channel Coding

Usually we get a noisy channel in practice. How to eliminate the noise? Definitely we can reduce noise by repetition. However, we could maximize the "distance" between the code.

Consider a map:

$$\{0,1\}^n \to \{0,1\}^m$$
$$c_1, \cdots, c_{2^n}.$$

where $m > n$. We will maximize $d_H(c_i, c_j)$. We already know that the number of the ball with its radium less than $\frac{r}{2}$ is

$$\sum_{k=0}^{\frac{r}{2}} \binom{m}{k}.$$

We can use Chernoff bound to estimate the value above. Moreover, by Weak Law of Large Numbers, we know that for random variable $X = X_1, \cdots, X_n$ with independent and identical distribution and its expection $EX = p$, the following fact holds:

$$P\left(\left|\frac{1}{n}\sum_i X_i - p\right| \geq \varepsilon\right) \leq \exp(-O(n)) \approx \exp(-2n\varepsilon^2).$$

Actually, right hand side is $\exp\left(-nD_e^B(p+\varepsilon\|p)\right)$, where

$$D_e^B(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

Also, if we want to correct t bits of error, then $\forall i \neq j$, $d_H(c_i, c_j) \geq 2t+1$. If we take efficiency into consideration, the efficiency of the codes above is defined as $\frac{m}{n}$.

Now, we aim to find 2^n codes in $\{0,1\}^m$ which satisfy the following requirement:

$$\forall 1 \leq i \neq j \leq 2^n, d_H(c_i, c_j) \geq \delta m$$

where $\delta \in \left(0, \frac{1}{2}\right)$ is a constant.

Theorem 7.1 (Gilbert-Vashamov Bound)**.** *If m and n are defined as above, then the following inequality holds:*

$$m \geq \frac{2n}{1 - H(\delta)}.$$

Here $H(\delta) = -[\delta\log\delta + (1-\delta)\log(1-\delta)]$ is its entropy.

Proof. This proof is a classic proof using probability. Actually, for $c_1, \cdots, c_{2^n} \in \{0,1\}^m$, $\forall i \neq j$,

$$P(d_H(c_i, c_j) < \delta m) \leq \exp\left(-mD_e^B\left(1 - \delta \| \frac{1}{2}\right)\right).$$

Hence,

$$P(\exists i \neq j, d_H(c_i, c_j) < \delta m) \leq \binom{2^n}{2} \exp\left(-mD_e^B\left(1 - \delta \| \frac{1}{2}\right)\right) \leq 2^{2n-1-mD^B\left(1-\delta\|\frac{1}{2}\right)}$$

If right hand side is smaller than 1, we conclude that such m is enough. So we have

$$2n - 1 - mD^B\left(1 - \delta \| \frac{1}{2}\right) = 2n - m(1 - H(\delta)) \leq 0 \Rightarrow m \geq \frac{2n}{1 - H(\delta)}.$$

□

But all theorems above only show the existence of such codes. How to find them in practice? In other words, how to design an algorithm for decoding and encoding?

Take Hamming Codes as an example: given an encoded code s, we could compute the Hamming distance of s and each decoded code c. However, its time complexity is exponential. When n is big enough, the algorithm above is unacceptable. If we use a table for memory, such decoding process could be done in constant time using hashing table. Nevertheless, space complexity also matters.

Let $m = 7$ and $n = 4$. In this example, our algorithm could correct 1 bit error, which means that $\forall i \neq j$, $d_H(c_i, c_j) \geq 3$. Let

$$\boldsymbol{H} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix},$$

and define the addition and multiplication in Galois field $GF(2)$. According to the equation below:

$$\dim \boldsymbol{H} = \dim \ker \boldsymbol{H} + \dim \operatorname{im} \boldsymbol{H},$$

we know that $\dim \ker \boldsymbol{H} = 4$ so $\ker \boldsymbol{H}$ has 16 elements. What is the minimum Hamming distance between the elements in $\ker \boldsymbol{H}$?

First of all, $\ker \boldsymbol{H}$ is a linear space so the addition of two elements in $\ker \boldsymbol{H}$ is still an elements in $\ker \boldsymbol{H}$. By simple enumeration we can find out that each element (except 0) has at least 3 '1', so

$$\min_{c_i, c_j \in \ker \boldsymbol{H}} d_H(c_i, c_j) \geq 3.$$

How to correct the 1-bit error? Trivially a mistaken code must have the form of $c + e_i$, where $c \in \ker \boldsymbol{H}$, the i^{th} bit of e_i is 1. Hence,

$$\boldsymbol{H}(c + e_i) = \boldsymbol{H}c + \boldsymbol{H}e_i = \boldsymbol{H}e_i$$

is the i^{th} column of the matrix $\boldsymbol{H}$.

How to encode? We have to find a set of base of $\ker \boldsymbol{H}$. Let

$$c_i = \sum_{j=1}^{4} x_j \varepsilon_j.$$

What we aim to do is to find a set of base efficiently. Re-arrange $\boldsymbol{H}$, then

$$\boldsymbol{H} = \begin{pmatrix} & & & 1 & \\ \boldsymbol{P}_{3\times 4} & & 1 & & \\ & & & & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_{3\times 4} & \boldsymbol{I}_{3\times 3} \end{pmatrix}.$$

Let

$$\boldsymbol{G}_{7\times 4} = \begin{pmatrix} \boldsymbol{I}_{4\times 4} \\ \boldsymbol{P}_{3\times 4} \end{pmatrix}, .$$

It is easy to verify that

$$\boldsymbol{HG} = 2\boldsymbol{P}_{3\times 4} = \boldsymbol{O}_{3\times 4},$$

hence $\boldsymbol{G}$ is a base of $\ker \boldsymbol{H}$. Finally, let $m \in \{0,1\}^4$ be a 4×1 vector, then

$$\boldsymbol{G}m = \begin{pmatrix} m \\ \boldsymbol{P}m \end{pmatrix}.$$

We could extend Hamming code to other k, m, n. Still, it only has the ability to correct 1 bit of error.

7.2 Channel Capacity

Now, we concentrate on how to describe a noisy channel. That is, given input X and output Y, we have to know $P(Y|X)$. We use mutual information to measure the information brought by X.

Definition 7.2 (Channel Capacity)**.** *Channel capacity is defined as the maximum mutual information in the channel. In other words,*

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

This subsection will mainly discuss Channel Coding Theorem given by Shannon.

Theorem 7.3 (Shannon: Channel Coding Theorem)**.** *Let R be the average amount of information of the input and C be the channel capacity.*

1. *If $R < C$, then $\forall \varepsilon > 0$, there exists a type of channel coding, such that its probability of error is less than ε.*
2. *If $R > C$, then $\exists \varepsilon_0 > 0$, such that there do not exist a type of channel coding of which the probability of error is less than ε_0.*

However, before we start our proof of this theorem, we will clarify several basic facts first.

By Weak Law of Large Numbers, let $X = X_1, \cdots$ be random variables with independent and identical distribution. Then for $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - EX\right| > \varepsilon\right) \to 0 (n \to +\infty).$$

Let g be a continuous function, then

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(X)\right| > \varepsilon\right) \to 0 (n \to +\infty).$$

Let $p(x)$ be the probability distribution function of random variable X and $g(X) = \log p(x)$. Hence,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) - E\log p(X)\right| > \varepsilon\right) \to 0 (n \to +\infty).$$

It is worth noticing that $-E\log p(x)$ is just $H(X)$. The inequality above implies that

$$P\left(2^{-n(H(X)+\varepsilon)} < P(X_1, \cdots, X_n) < 2^{-n(H(X)-\varepsilon)}\right) > 1 - \delta$$

when n is big enough.

Definition 7.4. *Define a set*

$$A_n = \left\{(x_1, \cdots, x_n) \in \{0,1\}^n : P(x_1, \cdots, x_n) \in 2^{-n(H(X)\pm\varepsilon)}\right\}.$$

For any sequence $(x_1, \cdots, x_n)$, if it is an element of A_n, we call it a typical sequence.

Typical sequences have some properties(we call them asymptotic equipartition properties):

- $P(A_n) \geq 1 - \delta$.
- Unstrictly speaking, for any typical sequence $(x_1, \cdots, x_n)$, $P(x_1, \cdots, x_n) \approx 2^{-nH(X)}$. So we could estimate the size of A_n as follows: $|A_n| \approx 2^{nH(n)}$.

Similarly, we could define jointly typical sequence.

Definition 7.5. *Let $(X, Y) \sim P(X, Y)$. A sequence $X_1, Y_1; \cdots, X_n, Y_n$ is called a jointly typical sequence, if*

- $P(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \in 2^{-n(H(X,Y)\pm\varepsilon)}$;
- $P(x_1, \cdots, x_n) \in 2^{-n(H(X)\pm\varepsilon)}$;
- $P(Y_1, \cdots, Y_n) \in 2^{-n(H(Y)\pm\varepsilon)}$.

Similarly, if we define A_n as the set of jointly typical sequence, then $|A_n| \approx 2^{nH(X,Y)}$. The problem is, if we randomly choose a typical sequence of X and a typical sequence of Y, what is the probability of (X, Y) be a jointly typical sequence?

Since there are approximately $2^{nH(X)}$ typical sequences of X and $2^{nH(Y)}$ typical sequences of Y, there could be $2^{nH(X)} \cdot 2^{nH(Y)}$ jointly sequences in all. Nevertheless, only $2^{nH(X,Y)}$ of them are jointly typical sequences, so the probability is

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

Finally, we will introduce an inequality before the proof starts.

Theorem 7.6 (Fano's Inequality)**.** *Let X, Y be discrete random variable and $X \in H$. We want to estimate X using $Y : \hat{X} = g(Y)$. Then the estimation error*

$$p_e = P(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log(|H| - 1)}.$$

Proof. Define a random variable E as

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

It is trivial that $H(E|\cdot) \leq 1$. Hence, by chain rule,

$$H(X, E|Y) = H(X|Y) + H(E|X, Y) = H(X|Y); \tag{3}$$

$$\begin{aligned} H(X, E|Y) &= H(E|Y) + H(X|E, Y) \leq 1 + H(X|E = 0, Y)(1 - p_e) + H(X|E = 1, Y)p_e \\ &= 1 + H(X|E = 1, Y)p_e \leq 1 + \log(|H| - 1)p_e. \end{aligned} \tag{4}$$

By comparing 3 and 4, the following inequality holds:

$$p_e \geq \frac{H(X|Y) - 1}{1 + \log(|H| - 1)} \geq \frac{H(X|Y) - 1}{\log(|H| - 1)}$$

□

Now we will start our proof.

Proof. • If $R < C$, let $P(X)$ be the input such that $I(X; Y)$ has its maximum value. As to this X, similarly define $P(X|Y)$. Let c_{ij} with independent and identical distribution generated from $P(X)$.

$$\begin{cases} c_1 = (c_{11}, \cdots, c_{1n}); \\ c_2 = (c_{21}, \cdots, c_{2n}); \\ \cdots \\ c_{2^{nR}} = (c_{2^{nR},1}, \cdots, c_{2^{nR},n}). \end{cases}$$

If the receiver accept a sequence $y_1, \cdots, y_n$, the sequence will be decoded as c_i, where c_i and $y_1, \cdots, y_n$ are jointly typical sequence. We want to prove that once n is big enough, the average error rate could be diminished to an arbitrarily small degree.

First of all, only in two cases will the channel coding report an error: no jointly typical sequence or not unique jointly typical sequence.

- No jointly typical sequence: According to Weak Law of Large Numbers, when n is big enough, the error rate could be arbitrarily small.
- The jointly typical sequence is not unique: Let (c_i, y) and (c_j, y) denote jointly typical sequences. Then c_i and c_j must be typical sequence. Hence, the probability of this event is $p_{ij} = 2^{-nC}$. Since i, j range from 1 to nR, the probability of the existence of nonunique typical sequence

$$p = (2^{nR} - 1) \cdot 2^{-nC} \approx 2^{n(R-C)} \to 0.$$

So we complete the first part of the proof.

- If $R > C$, similarly define $M = (c_1, \cdots, c_{2^{nR}})$ and Y be the message we receive. Then

$$nR = H(M) = H(M|Y) + I(M;Y) \leq H(M|Y) + nC.$$

Hence, $H(M|Y) \geq n(R - C)$. By applying 7.6, we know that

$$p_e \geq \frac{n(R-C) - 1}{\log(2^{nR} - 1)} \approx \frac{R - C}{R} > 0.$$

And this is the second part of the theorem.

□

8 Communication Complexity

Now, we wish to compute $f(x, y)$, where $x, y \in \{0, 1\}^n$ and $f \in \{0, 1\}$. But Alice only has x and Bob only has y. What is the lower bound of the cost of communication in order to compute $f(x, y)$?

If we show f in the form of a matrix, then every bit of communication will divide the rectangle into two parts and choose one part since we have a protocol in advance. In short, if there are M rectangles in the matrix with each rectangle all '0' or all '1', then we need at least $O(\log M)$ bits.

Definition 8.1 (Communication Complexity)**.** *Communication complexity is defined as the lower bound of the bits required to compute $f(x, y)$. It is usually written as $CC(f)$.*

Definition 8.2. *As to a fixed $f(x, y)$, let M_f be its matrix and $\chi(f)$ be the number of rectangles of the best division.*

Trivially $CC(f) \geq \log \chi(f)$. Also, since $\operatorname{rank}(A + B) \leq \operatorname{rank} A + \operatorname{rank} B$, we know that $\operatorname{rank} M_f \leq \chi(f)$.

Actually, as to communication complexity, *Andrew Yao* proved that

$$\log \chi(f) \leq CC(f) \leq O(\log^2 \chi(f)).$$

And now, people have other guess of the bound and there are other conjectures about communication complexity.

Conjecture 8.3 (Log-rank Conjecture)**.**

$$CC(f) \leq \mathit{poly}(\log(\operatorname{rank} M_f)).$$

Now, let us end this section with an interesting problem. Suppose that there is a graph $G = (V, E)$, $|V| = n$ and both Alice and Bob can see the graph. Alice has a clique C in this graph and Bob has an independent set I in this graph. Define a function $f(C, I)$ as follows:

$$f(C, I) = \begin{cases} 1, & \text{if } C \cap I \neq \varnothing; \\ 0, & \text{otherwise.} \end{cases}$$

How many bits do we need in order to compute f? Trivially $CC(f) \leq n$. We will prove that $O(\log^2 n)$ bits is enough.

Proof. We will try to list all vertexes in a matrix. This matrix has at most n rectangles. We conclude that if two matrices share at least one row, then there exists an edge between two vertexes. So Alice will have a set of rectangles of which every two share at least one row. Bob will have a set of rectangles of which every two share at least one column. Since these rectangles do not share any common space, the rectangles in Bob's set consists of an independent set in the graph. Also, the rectangles in Alice's set consists of a clique. Similarly define the concept of degree.

In every step, Alice will choose a rectangle R_x^* with the smallest degree in his set, and Bob will choose a rectangle R_y^* with the largest degree in his set. After they send the number of the rectangle in the matrix, we conclude that some rectangles are "invalid".

First of all, if $R \cap R_x^* = \varnothing$, then R shall be ignored. Also, if $R \cap R_x^* \neq \varnothing$ but R has a degree larger than R_x^*, then R shall be ignored as well. We will estimate how many rectangles are ignored each step.

Since Bob will choose after Alice, it is trivial that $d_x(R_x^*) \leq d_x(R_y^*)$. Respectively, Alice removed at least $n - d_x(R_x^*)$ rectangles and Bob removed at least $n - d_y(R_y^*)$ rectangles. Hence,

$$d_x(R_x^*) + d_y(R_y^*) \leq d_x(R_y^*) + d_y(R_y^*) \leq n.$$

The inequality above holds because two different rectangles cannot have common row and common column simultaneously. Finally, we get

$$\max\{n - d_x(R_x^*), n - d_y(R_y^*)\} \geq \frac{1}{2}(2n - d_x(R_x^*) - d_y(R_y^*)) \geq \frac{n}{2}.$$

So we removed at least half of the rectangles. So within $2\log^2 n$ bits we can compute f correctly. □

9 Fisher Information, Cramer-Rao Inequality

Let us think about the following case: There is a sample $X = (X_1, \cdots, X_n)$ with independent and identical distribution and its probability distribution function $f(X; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$. We wish to estimate θ from X via a mapping $\varphi(X) = \hat{\theta}$ and our goal is to minimize $D(\hat{\theta})$.

Definition 9.1 (Scored Function)**.** *Let $X = (X_1, \cdots, X_n)$ satisfy $f(X; \theta)$ with independent and identical distribution. The scored function is defined as*

$$S(X; \theta) = \frac{\partial}{\partial \theta} \ln f(X; \theta).$$

It is easy to show that the expectation of scored function shall be 0.

$$\begin{aligned} ES(X; \theta) &= \int \frac{\partial}{\partial \theta} \ln f(x; \theta) \cdot f(x) \mathrm{d}x = \int \frac{f'_\theta}{f} \cdot f \mathrm{d}x \\ &= \frac{\partial}{\partial \theta} \int f \mathrm{d}x = \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

Definition 9.2 (Fisher Information)**.** *Fisher information is defined as*

$$I(\theta) = DS(X; \theta).$$

Similarly we can know that

$$I(\theta) = -E\left[\frac{\partial^2}{\partial^2\theta}\ln f(X;\theta)\right]$$

Definition 9.3 (Unbiased Estimator)**.** *Let $\varphi(X)$ be a mapping from X to $\hat{\theta}$ to estimate θ. If $E\varphi(X) = \theta$, then it is called an unbiased estimator.*

Cramer-Rao Inequality tells us that whatever unbiased estimator we choose, a certain amount of error *always* exists.

Theorem 9.4 (Cramer-Rao Inequality)**.** *For all unbiased estimator φ of θ, the following inequality holds:*

$$D\varphi(X) \geq \frac{1}{I(\theta)}.$$

Proof.

$$\begin{aligned}
D\varphi(X)\cdot I(\theta) &= D\varphi(X)\cdot D(S)\\
&\geq \left[E\left(\varphi(X) - E\varphi(X)\right)(S - ES)\right]^2\\
&= \left(E(\varphi(X)-\theta)S\right)^2 = \left(E\varphi(X)S\right)^2\\
&= \left(\int \varphi(x)\frac{f'_\theta}{f}\cdot f\mathrm{d}x\right)^2 = \left(\int \frac{\partial}{\partial\theta}\left(\varphi(x)f(x,\theta)\right)\mathrm{d}x\right)^2\\
&= \left(\frac{\partial}{\partial\theta}\int \varphi(x)f(x;\theta)\mathrm{d}x\right)^2 = \left(\frac{\partial}{\partial\theta}E\varphi(X)\right)^2 = 1.
\end{aligned}$$

□

The following definition and theorem can apply in multi-variable situation as well. If we want to estimate $\theta = (\theta_1, \cdots, \theta_d)$ using $X = (X_1, \cdots, X_n)$, then 9.1 would be

$$S(X;\theta) = \nabla_\theta \ln f(X;\theta) = \left(\frac{\partial}{\partial\theta_1}\ln f, \cdots, \frac{\partial}{\partial\theta_d}\ln f\right).$$

9.2 would be defined as

$$\boldsymbol{I}(\theta) = \operatorname{cov} S(X;\theta) = \left(-E\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(X;\theta)\right)_{dd}.$$

On this occasion, 9.4 would be

$$\operatorname{cov}(\varphi(X)) \geq \boldsymbol{I}(\theta)^{-1}.$$

Since both left hand side and right hand side are matrices, the inequality above means that $\operatorname{cov}(\varphi(X)) - \boldsymbol{I}(\theta)^{-1}$ is a positive semi-definite matrix.

Theorem 9.5. *Let $X = (X_1, \cdots, X_n)$ and $\theta = (\theta_1, \cdots, \theta_d)$. Let $q(\theta) \in \mathbb{R}$ and $\varphi(X)$ be any unbiased estimator of $q(\theta)$. Then*

$$D(\varphi(X)) \geq \nabla_\theta q(\theta)^T \boldsymbol{I}(\theta)^{-1}\nabla_\theta q(\theta).$$

Proof. First of all,

$$\begin{aligned}
\nabla_\theta q(\theta) &= \nabla_\theta \int \varphi(x)f(x;\theta)\mathrm{d}x = \int \varphi(x)\nabla_\theta f(x;\theta)\frac{f(x;\theta)}{f(x;\theta)}\mathrm{d}x\\
&= \int \varphi(x)\nabla_\theta \ln f(x;\theta)f(x;\theta)\mathrm{d}x\\
&= E\left[\varphi(X)\cdot\nabla_\theta \ln f(X;\theta)\right] = E\left[\varphi(X)\cdot S(X;\theta)\right]\\
&= E\left[\left(\varphi(X) - E\varphi(X)\right)S(X;\theta)\right].
\end{aligned}$$

Since $S(X;\theta)\cdot S(X;\theta)^T = \boldsymbol{I}(\theta)$, the right hand side of the inequality could be rewritten as follows:

$$\begin{aligned}\text{RHS} &= E\left[\nabla_\theta q(\theta)^T \boldsymbol{I}(\theta)^{-1}(\varphi(X)-E\varphi(X))S(X;\theta)\right]\\ &\le \left[E\left(\varphi(X)-E\varphi(X)\right)^2\right]^{\frac{1}{2}}\cdot\left[E\left(\nabla_\theta q(\theta)^T \boldsymbol{I}(\theta)^{-1}S(X;\theta)S(X;\theta)^T\boldsymbol{I}(\theta)^{-1}\nabla_\theta q(\theta)\right)\right]^{\frac{1}{2}}\\ &= (D\varphi(X))^{\frac{1}{2}}\cdot(E\nabla_\theta q(\theta)^T\boldsymbol{I}(\theta)^{-1}\nabla_\theta q(\theta))^{\frac{1}{2}}\\ &= (D\varphi(X))^{\frac{1}{2}}\cdot(\text{RHS})^{\frac{1}{2}}.\end{aligned}$$

Hence, we know that RHS $\le D\varphi(X)$. □

Definition 9.6 (Renyi Entropy)**.** *For discrete random variable X with its probability distribution function $(p_1,\cdots,p_n)$ and $r>0$, define its Renyi entropy as*

$$H_r(X)=\frac{1}{1-r}\log\left(\sum_{i=1}^n p_i^r\right).$$

We can know that:

- As $r\to 0$, $H_r(X)\to \log n$ however its probability distribution function.
- As $r\to 1$, $H_r(X)\to H(X)$ is just its entropy.
- As $r\to+\infty$, $H_r(X)$ would be determined by the event with the biggest probability.

$$\lim_{r\to+\infty} H_r(X) = -\log\max_{1\le i\le n} p_i.$$

We will end this section with an application of information theory. As to a positive define matrix $\boldsymbol{A}$, verify that $f(\boldsymbol{A})=\log\det\boldsymbol{A}$ is a concave function, which means that for two positive definite matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $0<\lambda<1$, the following inequality holds:

$$\log\det(\lambda\boldsymbol{A}+(1-\lambda)\boldsymbol{B})\ge\lambda\log\det\boldsymbol{A}+(1-\lambda)\log\det\boldsymbol{B}.$$

We can prove this property using mathematical techniques. Now we will try to prove it by using information theory.

Proof. Consider two continuous random variable $X_1\sim N(0,\boldsymbol{A})$ and $X_2\sim N(0,\boldsymbol{B})$: Define two random variables $Y\sim B(1,\lambda)$ and $Z=YX_1+(1-Y)X_2$. It is trivial to prove that the covariance matrix of Z is $\lambda\boldsymbol{A}+(1-\lambda)\boldsymbol{B}$.

Given the covariance matrix, Gaussian distribution has maximum entropy, so

$$h(Z)\le\frac{1}{2}\log(2\pi\mathrm{e})^n\det(\lambda\boldsymbol{A}+(1-\lambda)\boldsymbol{B}).$$

Take conditional entropy into consideration: Since entropy has the property of addition, we know that

$$h(Z|Y)=\frac{1}{2}\lambda\log(2\pi\mathrm{e})^n\det(\boldsymbol{A})+\frac{1}{2}(1-\lambda)\log(2\pi\mathrm{e})^n\det(\boldsymbol{B}).$$

By Jensen's Inequality: $h(Z)\ge h(Z|Y)$, so

$$\log\det(\lambda\boldsymbol{A}+(1-\lambda)\boldsymbol{B})\ge\lambda\log\det\boldsymbol{A}+(1-\lambda)\log\det\boldsymbol{B}.$$

□

10 Rate Distortion Theorem

Definition 10.1 (Hamming Distance)**.** *Given $x,y\in\{0,1\}^n$, the Hamming distance between two strings is defined as*

$$d(x,y)=\sum_{i=1}^n I(x_i\ne y_i)=\sum_{i=1}^n 1_{x_i\ne y_i}.$$

According to asymptotic equipartition properties, only a small proportion of $\{0,1\}^n$ covers up almost every possible cases.

Definition 10.2 (Distortion Code)**.** *Let φ be a mapping from X^n to X^n. Say φ is an $(2^{nR}, n)$ distortion code, if the image of φ consists of 2^{nR} elements.*

Definition 10.3. *The distortion associated with $(2^{nR}, n)$ code is defined as*

$$D = Ed(x_1, \cdots, x_n; y_1, \cdots, y_n).$$

Definition 10.4 (Rate Distortion Function)**.** *The rate distortion function $R(D)$ is defined as the minimum R so that $(2^{nR}, n)$ code exists with its distortion less than D.*

Definition 10.5 (Information-Theoretic Rate Distortion Function)**.** *Let I denote the information. Information-theoretic rate distortion function is defined as*

$$R^I(D) = \min_{P(Y|X):Ed(X,Y)\leq D} I(X;Y).$$

Theorem 10.6. $R(D) = R^I(D)$.

We will end the introduction of these concepts with an example: Let $X_1, \cdots X_n \sim B(1,p)$ with independent and identical distribution. What about $R(D)$? Without loss of generation, we assume that $0 \leq p \leq \frac{1}{2}$.

Trivially when D is big enough, then $R^I(D) = 0$. Let $Y = (0, 0, \cdots, 0)$, then $Ed(\cdot,\cdot) = np$. So when $D \geq p$, then $R^I(D) = 0$.

Actually, the answer to this problem is

$$R^I(D) = \begin{cases} H(p) - H(D), & \text{if } 0 \leq D < p; \\ 0, & \text{otherwise.} \end{cases}$$

However, due to time limitation, we will omit the proof of the remaining part here.