Entropy and Limit theorems in Probability Theory [*]

Shigeki Aida

1 Introduction

Important Notice : Solve at least one problem from the following **Problems 1-8** and submit the report to me until June 29.

What is entropy? Entropy represents the uncertainty of probabilistic phenomena. The following definition is due to Shannon.

Definition 1.1 (Shannon) *Let us consider a finite set $E = \{A_1, \ldots, A_N\}$. A nonnegative function P on E is called a probability distribution if $\sum_{i=1}^N P(\{A_i\}) = 1$. Each A_i is called an elementary event. A subset of E is called an event. Then, for this probability distribution P, we define the entropy by*

$$H(P) = -\sum_{i=1}^N P(\{A_i\}) \log P(\{A_i\}). \tag{1.1}$$

Remark 1.2 *We use the convention, $0 \log 0 = 0$. If we do not mention about the base of the logarithmic function, we mean by $\log$ the natural logarithm, $\log_e$. We define for a nonnegative sequence $\{p_i\}_{i=1}^N$,*

$$H(p_1, \ldots, p_N) = -\sum_{i=1}^N p_i \log p_i. \tag{1.2}$$

Example 1.3 (1) **Coin tossing:**
$E = \{H, T\}$ *and* $P_1(\{H\}) = P_1(\{T\}) = 1/2$. *We have* $H(P_1) = \log 2$.
(2) **Dice:** $E = \{1, 2, 3, 4, 5, 6\}$. $P_2(\{i\}) = 1/6$ $(1 \le i \le 6)$. *Then we have* $H(P_2) = \log 6$.
(3) **Unfair Dice:** $E = \{1, 2, 3, 4, 5, 6\}$. $P_3(\{1\}) = 9/10, P_3(\{i\}) = 1/50$ $(2 \le i \le 6)$.

$$H(P_3) = \log\left[\left(\frac{10}{9}\right)^{9/10} (50)^{1/10}\right] \le \log\left(\frac{10}{9}\cdot\frac{3}{2}\right) < \log 2 = H(P_1) \tag{1.3}$$

Problem 1 *For unfair dice $E = \{1, 2, 3, 4, 5, 6\}$ with the probability $P_4(\{1\}) = 8/10$, $P_4(\{2\}) = 1/10$, $P_4(\{i\}) = 1/40$ $(i = 3, 4, 5, 6)$, calculate the entropy $H(P_4)$. Is $H(P_4)$ bigger than $H(P_1)$?*

In the above examples (1) and (2), the entropies are nothing but log(# all elementary events), because all elementary events have equal probabilities. The notion of entropy appeared in statistical mechanics also. Of course, the discovery is before that in the information theory. The following is a basic property of the entropy.

Theorem 1.4 *Suppose that $|E| = N$. Then for any probability distribution P, we have*

$$0 \le H(P) \le \log N. \tag{1.4}$$

Then the minimum value is attained by probability measures such that $P(\{A_i\}) = 1$ for some i. The maximum is attained by the uniform distribution P, namely, $P(A_i) = 1/N$ for all $1 \le i \le N$.

[*] This is one of lectures of "Mathematics B"in Graduate School of Science in Tohoku University in 2010.

We refer the proof to the proof of Theorem 3.1 in the next section.

The notion of entropy is used to solve the following problem:

Problem Here are eight gold coins and a balance. One of coins is an imitation and it is slightly lighter than the others. How many times do you need to use the balance to find the imitation?

Solution: In information theory, the entropy stands for the quantity of the information. In the above problem, we have eight equal possibilities such that each coin may be imitation. So the entropy is $\log 8$. We get some information by using the balance. By using the balance one time, we can get the following three informations: 1.The same weight, 2.The left one is lighter, 3.The right one is lighter. So it contains information $\log 3$. Thus, by using k-times of the balance, we get information which is amount of $k \log 3$. So if $k \log 3 < \log 8$, we do not get full information. So we need $k \geq 2$. Also it is not difficult to see that two times is enough. If the number of coins N satisfies $3^{n-1} < N \leq 3^n$, then n-times is enough. We refer the detail and other various examples to [15].

Problem 2 *In the above problem, how many times do you need to use the balance in the case where $n = 27$? Also present a method how to use the balance.*

In order to get into the detail, we recall basic notions in probability theory.

2 Basic notions in probability theory

Mathematically, probability space is defined to be a measure space whose total measure equals 1. We refer the audiences to some text books (e.g. [14], [13], [11]) for the precise definition.

Definition 2.1 (1) *A triplet $(\Omega, \mathcal{F}, P)$ is called a probability space if the following hold. Ω is a set and $\mathcal{F}$ is a family of some subsets of Ω satisfying that*

(i) *If $A_1, A_2, \ldots, A_i, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.*

(ii) *If $A \in F$, then $A^c \in F$.*

(iii) *$\Omega, \emptyset \in F$.*

For each $A \in \mathcal{F}$, a nonnegative number $P(A)$ is asssigned and satisfying that

(i) *$0 \leq P(A) \leq 1$ for all $A \in \mathcal{A}$.*

(ii) *$P(\Omega) = 1$.*

(iii) *(σ-additivity) If $A_1, A_2, \ldots, A_i, \ldots \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$ $(i \neq j)$, then*

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

The nonnegative function $P : \mathcal{F} \to [0,1]$ is called a probability measure on Ω. $A \in \mathcal{F}$ is called an event and $P(A)$ is called the probability of A.
(2) A function X on Ω is called a random variable (or measurable function) if for any $I = [a,b]$, $X^{-1}(I) := \{\omega \in \Omega \mid X(\omega) \in I\} \in \mathcal{F}$ holds. For random variable X, define

$$\mu_X(I) = P\left(X^{-1}(I)\right), \tag{2.1}$$

where I denotes an interval on $\mathbb{R}$. μ_X is also a probability on $\mathbb{R}$ and it is called the law of X or the distribution of X.

Problem 3 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Prove that if $A_i \in \mathcal{F}$ $(i = 1, 2, \ldots)$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Problem 4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $A, B \in \mathcal{F}$. Under the assumption that $A \subset B$, prove that $P(A) \leq P(B)$.*

In this course, we consider the following random variables.
(1) The case where X is a discrete-type random variable:
Namely, X takes finite number values $\{a_1, \ldots, a_n\}$. Then $p_i := P(\{\omega \in \Omega \mid X(\omega) = a_i\})(=: P(X = a_i))$ satisfies that $\sum_{i=1}^{n} p_i = 1$. The probability distribution μ_X of X is given by $\mu_X(\{a_i\}) = p_i$ $(1 \leq i \leq n)$.
(2) The case where X has a density function:
That is, there exists a nonnegative function $f(x)$ on $\mathbb{R}$ such that

$$P(\{\omega \in \Omega \mid X(\omega) \in [a,b]\}) = \int_a^b f(x)dx$$

for any interval $[a,b]$.

Definition 2.2 *For a random variable X, we denote the expectation, the variance by m and σ^2 respectively. Namely,*
(i) The case where X is a discrete-type random variable and takes values $a_1, \ldots, a_n$:

$$m = E[X] = \sum_{i=1}^{n} a_i P(X = a_i), \tag{2.2}$$

$$\sigma^2 = E[(X-m)^2] = \sum_{i=1}^{n} (a_i - m)^2 P(X = a_i). \tag{2.3}$$

(ii) The case where X is a continuous-type random variable which has the density function f:

$$m = E[X] = \int_{\mathbb{R}} x f(x)dx, \tag{2.4}$$

$$\sigma^2 = E[(X-m)^2] = \int_{\mathbb{R}} (x-m)^2 f(x)dx. \tag{2.5}$$

Definition 2.3 (Independence of random variables) *Let $\{X_i\}_{i=1}^N$ be random variables on a probability space $(\Omega, \mathcal{F}, P)$. N is a natural number or $N = \infty$. $\{X_i\}_{i=1}^N$ are said to be independent if for any $\{X_{i_k}\}_{k=1}^m \subset \{X_i\}_{i=1}^N$ $(m \in \mathbb{N})$ and $-\infty < a_k < b_k < \infty$, the following hold:*

$$P\left(X_{i_1} \in [a_1, b_1], \cdots, X_{i_m} \in [a_m, b_m]\right) = \prod_{i=1}^m P(X_{i_k} \in [a_k, b_k]). \tag{2.6}$$

Definition 2.4 *Let μ be a probability distribution on $\mathbb{R}$. Let $\{X_i\}_{i=1}^\infty$ be independent random variables and the probability distribution of X_i is equal to the same distribution μ for all i. Then $\{X_i\}$ is said to be i.i.d. (= independent and identically distributed) random variables with the distribution μ.*

3 Entropy and Law of large numbers (Shannon and McMillan's theorem)

Suppose we are given a set of numbers $A = \{1, \ldots, N\} \subset \mathbb{N}$. We call A the alphabet and the element is called a letter. A finite sequence $\{\omega_1, \omega_2, \ldots, \omega_n\}$ $(\omega_i \in A)$ is called a sentence with the length n. The set of the sentences whose length are n is the product space $A^n := \{(\omega_1, \ldots, \omega_n) \mid \omega_i \in A\}$. Let P be a probability distribution on A. We denote $P(\{i\}) = p_i$. In this section, we define the entropy of P by using the logarithmic function to the base N:

$$H(P) = -\sum_{i=1}^N P(\{i\}) \log_N P(\{i\}). \tag{3.1}$$

We can prove that

Theorem 3.1 *For every P, $0 \le H(P) \le 1$ holds. $H(P) = 0$ holds if and only if $P(\{i\}) = 1$ for some $i \in A$. $H(P) = 1$ holds if and only if P is the uniform distribution, that is, $p_i = 1/N$ for all i.*

Lemma 3.2 *Let $g(x) = x \log x$, or $g(x) = -\log x$. Then for any $\{m_i\}_{i=1}^N$ with $m_i \ge 0$ and $\sum_{i=1}^N m_i = 1$ and nonnegative sequence $\{x_i\}_{i=1}^N$, we have*

$$g\left(\sum_{i=1}^N m_i x_i\right) \le \sum_{i=1}^N m_i g(x_i). \tag{3.2}$$

Furthermore, when $m_i > 0$ for all i, the equality of (3.2) *holds if and only if $x_1 = \cdots = x_N$.*

Proof of Theorem 3.1**:** First, we consider the lower bound. Applying (3.2) to the case where $m_i = x_i = p_i$ and $g(x) = -\log x$, we have

$$\begin{aligned} H(p_1, \ldots, p_N) &\ge -\log\left(\sum_{i=1}^N p_i^2\right) \\ &\ge -\log 1 = 0. \end{aligned} \tag{3.3}$$

Clearly, in (3.3), the equality holds if and only if $p_i = 1$ for some i. Next, we consider the upper bound. By applying Lemma 3.2 to the case where $m_i = 1/N, x_i = p_i$ and $g(x) = x \log x$, for any nonnegative probability distribution $\{p_i\}$, we have

$$g\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \leq \frac{1}{N}\sum_{i=1}^{N} g(p_i). \tag{3.4}$$

Since $\sum_{i=1}^{N} p_i = 1$, this implies

$$-\frac{1}{N}\log N \leq \frac{1}{N}\sum_{i=1}^{N} p_i \log p_i.$$

Thus, $-\sum_{i=1}^{N} p_i \log p_i \leq \log N$ and $-\sum_{i=1}^{N} p_i \log_N p_i \leq 1$. By the last assertion of Lemma 3.2, the equality holds iff $p_i = 1/N$ for all i. □

Now we consider the following situation. Here is a (memoryless) information source S which sends out the letter according to the probability distribution P at each time independently. Namely, mathematically, we consider i.i.d. $\{X_i\}_{i=1}^{\infty}$ with the distribution P. We consider coding problem of the sequence of letters.

Basic observation: (1) Suppose that $P(\{1\}) = 1$ and $P(\{i\}) = 0$ $(2 \leq i \leq N)$. Then the random sequence X_i is, actually, a deterministic sequence $\{1, 1, \ldots, 1, \ldots\}$. Thus, the variety of sequence is nothing. In this case, we do not need to send the all sequences. In fact, if we know that the entropy of P is 0, then immediately after getting the first letter, we know that subsequent all letters are 1. Namely, we can encode all sentences, whatever the lengths are, to just one letter.
(2) Suppose that $N \geq 2$ and consider a probability measure such that $P(\{1\}) = P(\{2\}) = 1/2$ and $P(\{i\}) = 0$ for $3 \leq i \leq N$. Then note that the sequences contain $i(\geq 3)$ are not sent out. Thus the number of possible sequences under P whose lengths are n are 2^n. Note that the number of all sequences of alphabet A whose lengths are k is N^k. Thus, if $N^k \geq 2^n (\Longleftrightarrow \frac{k}{n} \geq \log_N 2 = H(P))$, then all possible sentences whose lengths are n can be encoded into the sentences whose lengths are $k(\leq n)$. Also the decode is also possible. More precisely, we can prove the following claim.

Claim If $\dfrac{k}{n} \geq H(P)$, then there exists an encoder $\varphi : A^n \to A^k$ and a decoder $\psi : A^k \to A^n$ such that

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n)\Big) = 0. \tag{3.5}$$

The probability $P\Big(\psi(\varphi(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n)\Big)$ is called the error probability. For general P, we can prove the following theorem [6].

Theorem 3.3 (Shannon and McMillan) *Take a positive number $R > H(P)$. For any $\varepsilon > 0$, there exists $M \in \mathbb{N}$ such that for all $n \geq M$ and k satisfying that $\frac{k}{n} \geq R$, there exists $\varphi : A^n \to A^k$ and $\psi : A^k \to A^n$ such that*

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n)\Big) \leq \varepsilon. \tag{3.6}$$

R is called the coding rate. We need the following estimates for the proof of the above theorem.

Lemma 3.4 *Let $\{Z_i\}_{i=1}^{\infty}$ be i.i.d. Suppose that $E[|Z_i|] < \infty$ and $E[|Z_i|^2] < \infty$. Then*

$$P\left(\left|\frac{Z_1 + \cdots Z_n}{n} - m\right| \geq \delta\right) \leq \frac{\sigma^2}{n\delta^2}, \tag{3.7}$$

where $m = E[Z_i], \sigma^2 = E[(Z_i - m)^2]$.

Problem 5 *Prove Lemma* 3.4 *using the Chebyshev lemma.*

This lemma immediately implies the following weak law of large numbers.

Theorem 3.5 *Assume the same assumption on $\{Z_i\}$ as in Lemma* 3.4*. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{Z_1 + \cdots Z_n}{n} - m\right| \geq \delta\right) = 0. \tag{3.8}$$

Proof of Theorem 3.3**:** Take $n \in \mathbb{N}$. The probability distribution of the i.i.d. subsequence $\{X_i\}_{i=1}^n$ is the probability distribution P_n defined on A^n such that for any $\{a_i\}_{i=1}^n$,

$$P_n\left(\{\omega_1 = a_1, \ldots, \omega_n = a_n\}\right) = \prod_{i=1}^n P\left(\{a_i\}\right). \tag{3.9}$$

Let us consider random variables on A^n, $Z_i(\omega) = -\log_N P\left(\{\omega_i\}\right)$ $(1 \leq i \leq n)$. Then $\{Z_i\}_{i=1}^n$ are i.i.d. and the expectation and the variance are finite. In fact, we have

$$\begin{aligned} m &= E[Z_i] = -\sum_{i=1}^n P\left(\{\omega_i\}\right) \log_n P\left(\{\omega_i\}\right) = H(P) \\ \sigma^2 &= E[(Z_i - E[Z_i])^2] = \sum_{i=1}^n \left(\log_N p_i\right)^2 p_i - H(P)^2. \end{aligned} \tag{3.10}$$

Take $\delta > 0$ such that $R > H(P) + \delta$. By Lemma 3.4,

$$P_n\left(\frac{1}{n}\sum_{i=1}^n (-\log_N P(\{\omega_i\})) \geq H(P) + \delta\right) \leq \frac{\sigma^2}{n\delta^2}. \tag{3.11}$$

Hence, for any $\varepsilon > 0$, there exists $M \in \mathbb{N}$ such that

$$P_n\left(\frac{1}{n}\sum_{i=1}^n (-\log_N P(\{\omega_i\})) \geq H(P) + \delta\right) \leq \varepsilon \qquad \text{for all } n \geq M. \tag{3.12}$$

Noting

$$\begin{aligned} &\left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \frac{1}{n}\sum_{i=1}^n (-\log_N P(\{\omega_i\})) < H(P) + \delta\right\} \\ &= \left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) > N^{-n(H(P)+\delta)}\right\} \\ &\subset \left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-nR}\right\} =: C_n, \end{aligned} \tag{3.13}$$

by (3.12), we have, for $n \geq M$,

$$\begin{aligned} & P\left((X_1,\ldots,X_n) \in C_n\right) \\ & \quad = P_n\left(\left\{(\omega_1\ldots,\omega_n) \in A^n \;\Big|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \\ & \quad \geq P_n\left(\left\{(\omega_1,\ldots,\omega_n) \in A^n \;\Big|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-n(H(P)+\delta)}\right\}\right) \geq 1-\varepsilon \end{aligned} \tag{3.14}$$

On the other hand, we have

$$|C_n|N^{-nR} \leq P_n\left(\left\{(\omega_1,\ldots,\omega_n) \in A^n \;\Big|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \leq 1 \tag{3.15}$$

Hence we have

$$|C_n| \leq N^{nR}. \tag{3.16}$$

By this estimate, if $k \geq nR$, then, there exists an injective map $\phi : C_n \to A^k$ and a map $\psi : A^k \to C_n$ such that

$$\psi(\phi(\omega_1,\ldots,\omega_n)) = (\omega_1,\ldots,\omega_n) \qquad \text{for any } (\omega_1,\ldots,\omega_n) \in C_n.$$

By taking a map $\varphi : A^n \to A^k$ which satisfies $\varphi|_{C_n} = \phi$, we have

$$P\left(\psi(\varphi(X_1,\ldots,X_n)) = (X_1,\ldots,X_n)\right) \;\; \geq \;\; P\left((X_1,\ldots,X_n) \in C_n\right) \geq 1-\varepsilon. \tag{3.17}$$

This completes the proof. □

4 Entropy and central limit theorem

Let $\{X_i\}_{i=1}^\infty$ be i.i.d. such that $m = E[X_i] = 0$ and $\sigma^2 = E[X_i^2] = 1$. Let

$$S_n = \frac{X_1 + \cdots X_n}{\sqrt{n}}.$$

Then we have

Theorem 4.1 (Central limit theorem=CLT) *For any $-\infty < a < b < \infty$,*

$$\lim_{n\to\infty} P\left(S_n \in [a,b]\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx. \tag{4.1}$$

The probability distribution whose density is $G(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$ is called a normal distribution (Gaussian distribution) with mean 0 and variance 1 and the notation $N(0,1)$ stands for the distribution. A standard proof of CLT is given by using the characteristic function of S_n, $\varphi(t) = E[e^{\sqrt{-1}tS_n}]$ $(t \in \mathbb{R})$. Below, we assume

Assumption 4.2 *The distribution of X_i has the density function f, namely,*

$$P\left(X_i \in [a,b]\right) = \int_a^b f(x)dx.$$

Then, we can prove that the distribution of S_n also has the density function $f_n(x)$ by Lemma 4.3 (1). In this case, (4.1) is equivalent to

$$\lim_{n\to\infty}\int_a^b f_n(x)dx = \int_a^b G(x)dx. \tag{4.2}$$

By using the entropy of S_n, we can prove a stronger result

$$\lim_{n\to\infty}\int_{\mathbb{R}} |f_n(x) - G(x)|dx = 0 \tag{4.3}$$

under additional assumptions on f. For the distribution P which has the density $f(x)$, and a random variable X whose distribution is P, we define the entropy H and Fisher's information I by

$$H(P) = -\int_{\mathbb{R}} f(x)\log f(x)dx, \tag{4.4}$$

$$I(P) = \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)}dx. \tag{4.5}$$

We may denote $H(P)$ by $H(X)$, $H(f)$ and may denote $I(P)$ by $I(X)$, $I(f)$.

We summarize basic results on H and I.

Lemma 4.3 (1) *If random variables X and Y have the density functions f and g respectively then $a(X+Y)$ $(a>0)$ has the density function $h(x) = \frac{1}{a}\int_{\mathbb{R}} f\left(\frac{x}{a} - y\right) g(y)dy$.*
(2) (Gibbs's lemma) *Let $f(x)$ be a density of a probability whose mean 0 and the variance is 1, that is,*

$$\int_{\mathbb{R}} xf(x)dx = 0, \tag{4.6}$$

$$\int_{\mathbb{R}} x^2 f(x) = 1. \tag{4.7}$$

Then we have,

$$H(f) \le H(G). \tag{4.8}$$

The equality holds if and only if $f(x) = G(x)$ for all x.
(3) (Shannon-Stam's inequality) *Let X, Y be independent random variables whose density functions satisfy* (4.6) *and* (4.7). *Then for $a, b \in \mathbb{R}$ with $a^2 + b^2 = 1$, we have*

$$a^2H(X) + b^2H(Y) \le H(aX + bY). \tag{4.9}$$

The equality holds iff the laws of X and Y are $N(0,1)$.
(4) (Blachman-Stam's inequality) *Let X, Y be independent random variables whose density functions satisfy* (4.6) *and* (4.7). *Then for $a, b \in \mathbb{R}$ with $a^2 + b^2 = 1$,*

$$I(aX + bY) \le a^2I(X) + b^2I(Y). \tag{4.10}$$

(5) (Csiszár-Kullback-Pinsker) *For a probability density function f satisfying* (4.6) *and* (4.7), *we have*

$$\left(\int_{\mathbb{R}} |f(x) - G(x)|dx\right)^2 \le 2\left(H(G) - H(f)\right). \tag{4.11}$$

(6) (*Stam's inequality*) *For a probability density function f, we have*

$$e^{-2H(f)} \leq \frac{1}{2\pi e} I(f). \tag{4.12}$$

(7) (*Gross's inequality*) *For a probability density function f, we have*

$$-2H(f) \leq I(f) - \log\left(2\pi e^2\right). \tag{4.13}$$

Problem 6 *Using Stam's inequality and an elementary inequality $\log x \leq x - 1$ $(x > 0)$, prove Gross's inequality.*

Problem 7 *Prove Stam's inequality by using Gross's inequality in the following way.*
(1) *Show that $f_t(x) = \sqrt{t} f(\sqrt{t}x)$is a probability density function for any $t > 0$. Next substituting $f_t(x)$ to Gross's inequality, get an upper bound estimate for $-2H(f)$.*
(2) *Optimize the estimate for $-2H(f)$ choosing suitable t and prove Stam's inequality.*

Problem 8 *Prove that Gross's inequality is equivalent to the following Gross's logarithmic Sobolev inequality:*

For all $u = u(x)$ with $\int_{\mathbb{R}} u(x)^2 d\mu(x) = 1$, we have

$$\int_{\mathbb{R}} u(x)^2 \log u(x)^2 d\mu(x) \leq 2 \int_{\mathbb{R}} |u'(x)|^2 d\mu(x), \tag{4.14}$$

where $d\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

Remark 4.4 $\mathcal{N}(f) = e^{2H(f)}$ is called the Shannon's entropy power functional. Stam [9] proved his inequality in 1959. This inequality reveals a relation between the Fisher information and the Shannon entropy. Later, Gross [5] proved his log-Sobolev inequality in the form (4.14) in 1975. He also proved that the log-Sobolev inequality is equivalent to the hypercontractivity of the corresponding Ornstein-Uhlenbeck semi-group. The hypercontractivity is very important in the study of quantum field theory. Gross's motivation is in the study of quantum field theory. After Gross's work, the importance of the log-Sobolev inequality was widely known. Meanwhile, the contribution of the Stam had been forgotton but some people noted his contribution like in [2, 3] and some people call the inequality as Stam-Gross logarithmic Sobolev inequality as in [10]. My recent work [1] is related with semiclassical problem of $P(\phi)_2$ Hamiltonians which appear in the constructive quantum field theory. In the work, I used logarithmic Sobolev inequality to determine the asymptotic behavior of the ground state energy (=the lowest eigenvalue of the Hamiltonian) under the semiclassical limit $\hbar \to 0$.

By Lemma 4.3 (1), S_n has a density function f_n. Also note that $H(S_{nm}) \geq H(S_n)$ for any $n, m \in \mathbb{N}$. We can prove f_n converges to G.

Theorem 4.5 *Assume that f is C^1-function and $I(f) < \infty$. Then for any n, f_n is a continuous function. Moreover we have*

$$\lim_{n\to\infty} f_n(x) = G(x) \qquad \textit{for all } x \tag{4.15}$$

$$\lim_{n\to\infty} H(f_n) = H(G), \tag{4.16}$$

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(x) - G(x)| dx = 0. \tag{4.17}$$

Sketch of Proof: By Shannon-Stam's inequality, we can prove that (4.16). This and Csiszár-Kullback-Pinsker inequality implies (4.17). On the other hand, Blachman-Stam inequality implies $I(f_n) \leq I(f)$. This and (4.17) implies (4.15). See [2], [7] and references in them for the detail. □

5 Boltzmann's H-theorem, Markov chain and entropy

We recall kinetic theory of rarefied gases. Let $(v_x^i(t), v_y^i(t), v_z^i(t))$ be the velocity of the i-th molecule at time t $(1 \leq i \leq N)$. N denotes the number of molecules. The velocities $v^i(t) = (v_x^1(t), v_y^i(t), v_z^i(t))$ obey Newton's equation of motion, but N is very big and it is almost meaningless to know each behavior of $v^i(t)$. Boltzmann considered the probability distribution of the velocity, say, $f_t(v_x, v_y, v_z)dv_x dv_y dv_z$ and derived the following his H-theorem:

Theorem 5.1 (Boltzmann) *Let*

$$H(t) = -\int_{\mathbb{R}^3} f_t(v_x, v_y, v_z) \log f_t(v_x, v_y, v_z) dv_x dv_y dv_z.$$

Then $\dfrac{d}{dt}H(t) \geq 0.$

Remark 5.2 (1) In statistical mechanics, the entropy is nothing but $kH(t)$, where k is the Boltzmann's constant (=$1.38 \times 10^{-23} J \cdot K^{-1}$). Therefore, the above theorem implies that the entropy increases.
(2) Some people raised questions about the H-theorem.
(**i**) Newton's equation of motion for the particles $\boldsymbol{x}(t) = (\boldsymbol{x_i}(t))_{i=1}^N$ moving in a potential U reads as follows:

$$\begin{aligned} m_i \ddot{\boldsymbol{x}}_{\boldsymbol{i}}(t) &= -\frac{\partial U}{\partial \boldsymbol{x_i}}(\boldsymbol{x}(t)) & (5.1)\\ \boldsymbol{x_i}(0) &= \boldsymbol{x_{i,0}}, & (5.2)\\ \dot{\boldsymbol{x}}_{\boldsymbol{i}}(0) &= \boldsymbol{v_i}, & (5.3) \end{aligned}$$

where $(\boldsymbol{x_{i,0}}), (\boldsymbol{v_i})$ are initial positions and the initial velocities. m_i is the mass of $\boldsymbol{x_i}$. The time reversed dynamics $\boldsymbol{x}(-t)$ is the solution of (5.1) with initial velocity $-\boldsymbol{v_i}$. Clearly, it is impossible that both entropies of $\boldsymbol{x}(t)$ and $\boldsymbol{x}(-t)$ increase. This is a contradiction.
(**ii**) The second question is based on Poincaré's recurrence theorem:
• There exists a time t_0 such that $\boldsymbol{x}(t_0)$ is close to $\boldsymbol{x}(0)$.

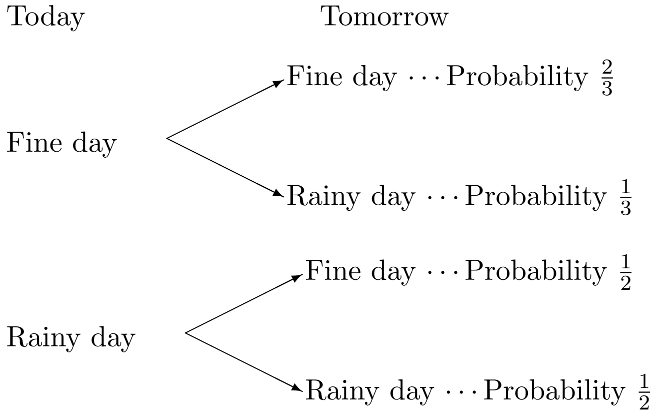
Then at that time t_0, the entropy also should be close to each other. Therefore, the monotone property of entropy does not hold.

The reason why the above paradox appeared is in the statistical treatment of rarefied gases. We refer the recent progress based on probabilistic models to [4].

Note: Poincaré's recurrence theorem holds under some additional assumptions on U.

From now on, we consider stochastic dynamics which are called Markov chains and prove that the entropy of the dynamics increases when time goes on.

Example 5.3 There are datum on the weather in a certain city.


Today
Tomorrow
Fine day
Fine day ··· Probability $\frac{2}{3}$
Rainy day ··· Probability $\frac{1}{3}$
Rainy day
Fine day ··· Probability $\frac{1}{2}$
Rainy day ··· Probability $\frac{1}{2}$

Today(=0-th day)is fine, then how much is the probability that the n-th day is also fine?

Solution:

Let p_k be the probability that the k-th day is fine and set $q_k = 1 - p_k$, that is the probability that k-th day is rainy day. Then (p_k, q_k) satisfies the following recurrence relation:

$$(p_k, q_k) \quad = \quad (p_{k-1}, q_{k-1}) \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{5.4}$$

So we obtain

$$(p_k, q_k) = (1, 0) \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}^k.$$

Definition 5.4 *Let $E = \{S_1, \ldots, S_N\}$ be a finite set. E is called a state space. We consider a random motion of a particle on S. Let $\{p_{ij}\}_{i,j \in E}$ be nonnegative numbers satisfying that*

$$\sum_{j=1}^{N} p_{ij} = 1 \qquad \text{for all } 1 \leq i \leq N. \tag{5.5}$$

p_{ij} denotes the probability that the particle moves from S_i to S_j. If the probability that the particle locates at S_i is π_i $(1 \leq i \leq N)$ at the time n, then the probability that the particle locates at S_j at the time $n+1$ is $\sum_{i=1}^{N} \pi_i p_{ij}$. That is, if the initial distribution of the particle is $\pi(0) = (\pi_1, \ldots, \pi_N)$, then the distribution $\pi(n) = (\pi_1(n), \ldots, \pi_N(n))$ at time n is given by

$$\pi(n) = \pi(0) P^n, \tag{5.6}$$

where P denotes the $N \times N$ matrix whose (i,j)-element is p_{ij}. Below, we denote by $p_{ij}^{(n)}$ the (i,j)-element of P^n.

We prove the following.

Theorem 5.5 *Assume that*

(A1) $\displaystyle\sum_{i=1}^{N} p_{ij} = 1$ *for all* $1 \leq j \leq N$.

(A2) (Mixing property of the Markov chain) *There exists $n_0 \in \mathbb{N}$ such that $p_{ij}^{(n_0)} > 0$ for any $i, j \in \{1, \ldots, N\},$.*

Then for any initial distribution $\pi = (\pi_1, \ldots, \pi_N)$, we have

$$\lim_{n\to\infty} \pi(n)_i = \frac{1}{N}. \qquad \textit{for all } 1 \le i \le N. \tag{5.7}$$

Note that $(\pi(n)_i)_{i=1}^N = \pi(n) = \pi P^n$.

Remark 5.6 (A1) is equivalent to

$$(1, \ldots, 1) = (1, \ldots, 1)P.$$

Also (A1) holds if $p_{ij} = p_{ji}$ for all $i, j \in E$.

This theorem can be proved by using the entropy of $\pi(n)$ and Lemma 3.2. Recall

$$H(\pi) = -\sum_{i=1}^N \pi_i \log \pi_i.$$

Lemma 5.7 *Let $Q = (q_{ij})$ be a (N, N)-probability matrix, that is, $q_{ij} \ge 0$ for all i, j and $\sum_{j=1}^N q_{ij} = 1$ for all i. Assume $\sum_{i=1}^N q_{ij} = 1$ for any j. Then for any π, $H(\pi Q) \ge H(\pi)$. In addition, if $q_{ij} > 0$ for all i, j, then, $H(\pi Q) > H(\pi)$ for all $\pi \ne (1/N, \ldots, 1/N)$.*

Proof.

$$\begin{aligned} H(\pi Q) &= -\sum_{i=1}^N (\pi Q)_i \log{(\pi Q)}_i \\ &= -\sum_{i=1}^N \left(\sum_{k=1}^N \pi_k q_{ki}\right) \log \left(\sum_{k=1}^N \pi_k q_{ki}\right). \end{aligned} \tag{5.8}$$

Since $\sum_{k=1}^N q_{ki} = 1$, by applying Lemma 3.2 to the case where $m_k = q_{ki}, x_k = \pi_k$,

$$\left(\sum_{k=1}^N \pi_k q_{ki}\right) \log \left(\sum_{k=1}^N \pi_k q_{ki}\right) \le \sum_{k=1}^N q_{ki} \pi_k \log \pi_k. \tag{5.9}$$

(5.8), (5.9) and $\sum_{i=1}^N q_{ki} = 1$ imply $H(\pi Q) \ge H(\pi)$. The last assertion follows from the last assertion of Lemma 3.2. □

By using this lemma, we prove Theorem 5.5.

Proof of Theorem 5.5 Since $\pi(n)$ moves in a bounded subset in $\mathbb{R}^N$, there exist the accumulation points. That is, there exist $x = (x_1, \ldots, x_N)$ and a subsequence $\{\pi(n(k))\}_{k=1}^\infty$ such that $\lim_{k\to\infty} \pi(n(k)) = x$. x is also a probability and satisfies that $H(x) = \lim_{k\to\infty} H(\pi(n(k)))$. Note that P^{n_0} is a probability matrix and all elements of P^{n_0} are positive. So if $x \ne (1/N, \ldots, 1/N)$, then $H(xP^{n_0}) > H(x)$. But $xP^{n_0} = \lim_{k\to\infty} \pi P^{n(k)+n_0}$ and $H(xP^{n_0}) = \lim_{k\to\infty} H(\pi P^{n(k)+n_0})$. Since $H(\pi(n))$ is an increasing sequence, $H(xP^{n_0}) = H(x)$. This is a contradiction. So $x = (1/N, \ldots, 1/N)$ which completes the proof. □

References

[1] S. Aida, Semi-classical limit of the lowest eigenvalue of a Schrdinger operator on a Wiener space. II. $P(\phi)_2$-model on a finite volume. J. Funct. Anal. 256 (2009), no. 10, 3342–3367.

[2] A. Barron, Entropy and the central limit theorem. Ann. Probab. 14 (1986), no. 1, 336–342.

[3] E. Carlen, Super additivity of Fisher's Information and logarithmic Sobolev inequalities, Journal of Functional Analysis, **101** (1991), 194–211.

[4] 内山耕平, 舟木直久著, ミクロからマクロへ, 2. 格子気体の流体力学極限, シュプリンガーフェアラーク東京, 2002 年

[5] L. Gross, Logarithmic Sobolev inequalities. Amer. J. Math. 97 (1975), no. 4, 1061–1083.

[6] A.I. Khinchin, Mathematical foundations of information theory, Dover books on advanced mathematics, 1957.

[7] P.L. Lions and G. Toscani, A strengthened central limit theorem for smooth densities, Journal of Functional Analysis, **129** (1995), 148–167.

[8] C.E. Shannon, The mathematical theory of communication, Bell Syst, Techn. Journ. **27**, 379–423, 623–656 (1948).

[9] A.J. Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon, Information and Control **2** (1959), 101–112.

[10] C. Villani, Topics in optimal transportation. Graduate Studies in Mathematics, 58. American Mathematical Society, Providence, RI, 2003.

[11] D. Williams, Probability with martingales, Cambridge Mathematical Textbooks, 1991.

[12] K. Yoshida, Markoff process with a stable distribution, Proc. Imp. Acad. Tokyo, **16** (1940), 43–48.

[13] 熊谷 隆, 確率論, 共立出版株式会社.

[14] 西尾真喜子, 確率論, 実教出版.

[15] ヤグロム, 情報理論入門, みすず書房.(邦訳)