Probability and Entropy [*]

Shigeki Aida

1 Introduction

Suppose we are given a set of numbers $A = \{1, \ldots, N\} \subset \mathbb{N}$. We call A the alphabet and we call each element of A 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 $P(\{i\}) = p_i$.

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 independent random sequences $\{X_i\}_{i=1}^{\infty}$ with the same distribution P. We consider coding problem of the random sentences.

Basic observation: (1) Suppose that $P(\{1\}) = 1$ and $P(\{i\}) = 0$ $(2 \le i \le N)$. 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 this case, 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 $N \ge 3$ and consider a probability measure such that $P(\{1\}) = P(\{2\}) = 1/2$ and $P(\{i\}) = 0$ for $3 \le i \le N$. Then note that the sequences contain $i(\ge 3)$ 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 $N^k \ge 2^n$ then all possible sentences whose lengths are n can be encoded into the sentences of A whose lengths are $k(\le n)$. Also the decode is also possible. Note that

$$N^k \ge 2^n \Longleftrightarrow \frac{k}{n} \ge \log_N 2$$

The number $\log_N 2$ is the entropy of the probability distribution P in (2). In the case of (1), the entropy of the probability P is 0. Hence $k = 1$ is possible.

In general, we define the entropy of P by using the logarithmic function to the base N:

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

We summarize what we prove in the case of (2).

Coding result in the case of (2) If $\dfrac{k}{n} \ge H(P)$, then there exists an encoder $\varphi : A^n \to A^k$ and a decoder $\psi : A^k \to A^n$ such that

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \ne (X_1, \ldots, X_n)\Big) = 0. \tag{1.2}$$

The probability $P\Big(\psi(\varphi(X_1, \ldots, X_n)) \ne (X_1, \ldots, X_n)\Big)$ is called the error probability. For general P, we can prove the following theorem [2].

[*] This is one of lectures of "Mathematics B" in Graduate School of Science in Tohoku University in 2011.

Theorem 1.1 (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*

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n)\Big) \leq \varepsilon. \tag{1.3}$$

Remark 1.2. (1) *k/n is called the coding rate.*

(2) *φ is called an encoder and ψ is called a decoder.*

About entropy, we have

Theorem 1.3. *For every P, $0 \leq H(P) \leq 1$ 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.*

Clearly, the uniform distribution is most "random probability" and the probability concentrates one letter is most "not random probability". That is, we may say that the entropy stands for the uncertainty of probability.

The weak law of large number (actually an estimate by Chebyshev's inequality) is necessary in the proof of Shannon-McMillan's theorem and elementary probability is enough for the understanding of the proof. However, I think, it is not bad to learn "probability theory based on measure theory".

The plan of this lecture:

(I) Elementary probability

(II) Modern probability theory based on Lebesgue integration

(III) Proof of Shannon-McMillan theorem

2 Elementary probability theory

We recall several notion in elementary probability theory:

Sample Space, Event, Elementary Event, Probability, Random Variable, Expectation, Independent Event.....

Definition 2.1. *Let Ω be a set and suppose that for each subset $A \subset \Omega$ a non-negative number $P(A)$ is given such that*
(1) *$0 \leq P(A) \leq 1$ for any A and $P(\Omega) = 1$.*
(2) *When $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$.*

Then P is called a probability on Ω. Ω is called a **sample space**. *Each element $\omega \in \Omega$ is called an* **elementary event**. *Any subset of Ω is called an* **event**. *The sample space Ω itself is called a* **total event**.

Example 2.2 (Rolling Dice n times)**.** *Let*

$$\Omega_n = \{\omega = (x_1, \dots, x_n) \mid x_i = 1, \dots, 6\}.$$

For $A \subset \Omega_n$, define

$$P(A) = \frac{\sharp A}{6^n}.$$

Definition 2.3. *A probability P on $\mathbb{R}$ is called a* **probaility distribution** (**probability law**) *on $\mathbb{R}$.*

Definition 2.4. (1) *Let $\{a_i\}_{i=1}^N \subset \mathbb{R}$. Let p_i $(1 \le i \le N)$ be non-negative numbers such that*

$$\sum_{i=1}^N p_i = 1.$$

For $A \subset \mathbb{R}$, define

$$P(A) = \sum_{\{i|\omega_i \in A\}} p_i.$$

This probability distribution P is called a **discrete type probability (distribution)**.
(2) *Let $f(x)$ be a non-negative function on $\mathbb{R}$ such that $\int_{\mathbb{R}} f(x)dx = 1$. For $A \subset \mathbb{R}$, let*

$$P(A) = \int_A f(x)dx.$$

This probability P is called a **continuous type probability (distribution)** *with the* **(probability) density function** *f.*

Definition 2.5 (Random variable)**.** *Let (Ω, P) be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a random variable. Let us define a probability distribution P_X on $\mathbb{R}$ by*

$$P_X(A) = P(X \in A).$$

P_X is called the **probability distribution (probability law)** *of X.*

We define the expectation of a random variable.

Definition 2.6. *For a random variable X, we define the expectation $E[X]$ as follows.*

(i) *The case where X is a discrete-type random variable and takes values $a_1, \dots, a_N$:*

$$E[X] = \sum_{i=1}^N a_i P(X = a_i). \tag{2.1}$$

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

$$E[X] = \int_{\mathbb{R}} x f(x)dx. \tag{2.2}$$

The expectation $E[X]$ depends only on the distribution of P_X of X. So we call $E[X]$ the expectation (or **mean***) of P_X.*

Proposition 2.7 (Linearity of expectation)**.** *Let X, Y be random variables. Then for any $\alpha, \beta \in \mathbb{R}$,*

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

Definition 2.8. *We define the variance of X, say $V[X]$, by*

$$V[X] = E[(X-m)^2],$$

where $m = E[X]$.

Lemma 2.9. *In the case of* (i) *in Definition* 2.6*,*

$$V[X] = \sum_{i=1}^{N} (a_i - m)^2 P(X = a_i)$$

and in the case of (ii) *in Definition* 2.6*,*

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

In the next section, we give the modern definition of the expectation based on Lebesgue integration.

We introduce the notion of **independence of events** and **independence of random variables**.

Definition 2.10 (Independence of events)**.** *Let $A_1, \dots, A_n$ be events of Ω. We say that $A_1, \dots, A_n$ are independent if for any $1 \le i_1 < \cdots < i_k \le n$,*

$$P\left(\cap_{l=1}^{k} A_{i_l}\right) = \prod_{l=1}^{k} P(A_{i_l}).$$

Definition 2.11 (Independence of random variables)**.** *Let $\{X_i\}_{i=1}^N$ be random variables on a probability space (Ω, P). N is a natural number or $N = \infty$. $\{X_i\}_{i=1}^N$ are said to be independent if for any $m \le N$ (when $N = \infty$, m is any natural number) the following hold: For any intervals I_k $(1 \le k \le m)$, the events*

$$\{X_1 \in I_1\}, \dots, \{X_m \in I_m\}$$

are independent. That is the following hold: , the following hold:

$$P(X_1 \in I_1, \cdots, X_m \in I_m) = \prod_{i=1}^{m} P(X_i \in I_i). \tag{2.3}$$

Theorem 2.12. *Let X, Y be independent random variables. Then $E[XY] = E[X]E[Y]$.*

Proof. We prove the case where X and Y are discrete type random variables. Let $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$ be the values of X and Y respectively. Let $E_i=\{\omega \mid X(\omega)=x_i\}$, $F_j=\{\omega \mid X(\omega)=y_j\}$. Then

$$X(\omega)=\sum_{i=1}^{n} x_i 1_{E_i}(\omega), \quad Y(\omega)=\sum_{j=1}^{m} y_j 1_{F_j}(\omega),$$

where 1_A is defined such that $1_A(\omega)=1$ for $\omega\in A$ and $1_A(\omega)=0$ for $\omega\in A^c$. Therefore

$$\begin{aligned} E[XY] &= E\left[\left(\sum_{i=1}^{n} x_i 1_{E_i}\right)\left(\sum_{j=1}^{m} y_j 1_{F_j}(\omega)\right)\right] \\ &= \sum_{1\le i\le n, 1\le j\le m} x_i y_j E[1_{E_i}1_{F_j}] \\ &= \sum_{1\le i\le n, 1\le j\le m} x_i y_j P(E_i\cap F_j) && (2.4) \\ &= \sum_{1\le i\le n, 1\le j\le m} x_i y_j P(E_i)P(F_j) && (2.5) \\ &= \left(\sum_{1\le i\le n} x_i P(E_i)\right)\left(\sum_{1\le j\le m} y_j P(F_j)\right) = E[X]E[Y]. && (2.6) \end{aligned}$$

In (2.4) and (2.5), we have used respectively

$$E[1_{E_i}1_{F_j}]=E[1_{E_i\cap F_j}]=P(E_i\cap F_j),$$

$$P(E_i\cap F_j)=P(\{X=x_i, Y=y_j\})=P(X=x_i)P(Y=y_j)=P(E_i)P(F_j).$$

□

Exercise 1. *Let X_i $(1\le i\le n)$ be independent random variables such that $P(X_i=1)=p$, $P(X_i=0)=1-p$, $(0<p<1)$. Let $S_n=\sum_{i=1}^{n} X_i$.*
(1) Prove that $P(S_n=k)={}_n\mathrm{C}_k p^k(1-p)^{n-k}$ $(0\le k\le n)$[†].
(2) Calculate $E[X_i]$ and show that $E[S_n]=np$.
(3) Show that $V[S_n]=np(1-p)$.

Exercise 2. *Let us consider Ω_2 in Example 2.2. That is*

$$\Omega_2=\{\omega=(x_1,x_2) \mid 1\le x_1,x_2\le 6\}, \quad P(A)=\frac{\sharp A}{36}.$$

Let $X_1(\omega)=x_1$, $X_2(\omega)=x_2$ when $\omega=(x_1,x_2)$. Show that X_1,X_2 are independent. Find the distributions of $\max(X_1,X_2)$ *and* $\min(X_1,X_2)$ *and their expectations.*

Exercise 3. (1) *Let P be the probability distribution which has the density function*

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

[†] This distribution of S_n is called the Bernoulli distribution $B(n,p)$

Show that the mean of P is m and the variance of X is σ^2. The distribution P is called the normal distribution with mean m and variance σ^2 and denoted by $N(m, \sigma^2)$. Suppose that the law of the random variable X is $N(0,1)$(=standard normal distribution). Find the density function of X^2.

(2) *Let P be the Poisson distribution with parameter λ (> 0), that is, P is a discrete type probability such that*

$$P(\{k\}) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad k = 0, 1, \ldots.$$

Find the expectation and the variance of the Poisson distribution.

3 Probability theory based on measure theory

We already defined a probability space for shaking dice n times. How about the probability space for shaking the dice infinitely many times ? The sample space should be

$$\Omega_\infty = \{\omega = (x_1, \ldots, x_n, \ldots) \mid 0 \le x_i \le 6\}$$

This set is infinite set and the probability cannot be defined in a similar way to Ω_n $(n < \infty)$. To study this kind of probability, we need measure theory.

First, we introduce the notion of probability space based on measure theory.

Definition 3.1. (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) $\Omega, \emptyset \in F$.

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

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

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

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

Exercise 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 $P(A) \le P(B)$.*

What is $\mathcal{F}$? Let us consider $\Omega = [0,1]$. One may think that the length of $A \subset [0,1]$, say $|A|$, is natural candidate of the probability $P(A)$ in $[0,1]$. However what is the length of the set A? Of course if $A = [a,b] \subset [0,1]$, $|A| = b-a$. Also if

$$A = I_1 \cup \cdots \cup I_n, \quad I_i = [a_i, b_i], \quad I_i \cap I_j = \emptyset \ (i \neq j)$$

then $|A| = \sum_{i=1}^n (b_i - a_i)$. Actually, it is not possible to define the length for all subsets. A subset of $[0,1]$ for which the length (**Lebesgue measure**) is defined is called a **Lebesgue measurable subset**. We denote all Lebesgue measurable subsets by $\mathfrak{M}_L$. Then $\mathfrak{M}_L$ satisfies

(i) If $A_1, A_2, \ldots, A_i, \ldots \in \mathfrak{M}_L$, then $\cup_{i=1}^\infty A_i \in \mathfrak{M}_L$,

(ii) If $A \in \mathfrak{M}_L$, then $A^c \in \mathfrak{M}_L$.

(iii) If $A_1, \ldots, A_n, \ldots \in \mathfrak{M}_L$ and $A_i \cap A_j = \emptyset \ (i \neq j)$, then

$$|\cup_{i=1}^\infty A_i| = \sum_{i=1}^\infty |A_i|.$$

So $\mathfrak{M}_L$ is also a σ-field and $([0,1], \mathfrak{M}_L, |\cdot|)$ is a probability space.

We give more example of probability spaces.

Example 3.2. (1) *We consider the probability space for rolling dice n times. Then the sample space is $\Omega_n = \{\omega = (x_1, \ldots, x_n) \mid 1 \leq x_i \leq 6\}$. Also $\mathcal{F}$ = all subsets of Ω, and $P(A) = \frac{\sharp A}{6^n}$.*
(2) *We consider the probability space for rolling dice infinitely many times. Clearly the sample space is*

$$\Omega_\infty = \{\omega = (x_1, \ldots, x_n, \ldots) \mid 1 \leq x_i \leq 6\}.$$

Take a sequence $(a_1, \ldots, a_n) \in \Omega_n$. We define

$$C(a_1, \ldots, a_n) = \{\omega = (x_1, \ldots, x_n, \ldots) \mid x_1 = a_1, \ldots, x_n = a_n\} \subset \Omega_\infty.$$

This set is called a cylinder set. It is natural to define the probability of $C(a_1, \ldots, a_n)$ by

$$P(C(a_1, \ldots, a_n)) = \frac{1}{6^n}. \tag{3.1}$$

Let

$$\mathcal{F} = \textit{the smallest } \sigma\textit{-field including all cylinder sets.}$$

The we can prove that the probability can be defined for all sets in $\mathcal{F}$ extending (3.1). Note that $\mathcal{F} \subsetneqq 2^\Omega$.

Now we give the notion of random variables as measurable functions.

Definition 3.3. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X : \Omega \to \mathbb{R}$ be a real-valued function on Ω. We say that X is a measurable function if for any intervals $I \subset \mathbb{R}$,*

$$X^{-1}(I)(:= \{\omega \in \Omega \mid X(\omega) \in I\} \in \mathcal{F}.$$

Here we mean by interval the sets:

$$[a,b], \quad [a,b), \quad (a,b], \quad (a,b), \quad (-\infty,b], \quad (-\infty,b), \quad (a,\infty), \quad [a,\infty)$$

We call a measurable function on Ω a random variable.

Exercise 5. *Let X be a measurable function on $(\Omega, \mathcal{F}, P)$. Then*

$X^+(\omega) = \max(X(\omega), 0)$(= *positive part of* X), $X^-(\omega) = \max(-X(\omega), 0)$(= *negative part of* X),

*and $|X|$ (the function of the absolute value of X) are also measurable function. (***Actually we can prove that if X is maesurable then $\varphi(X)$ is also measurable for any continuous function φ on $\mathbb{R}$***).*

Exercise 6. *LetX_n $(n = 1, 2, \dots)$ be measurable functions. Assume that* $\lim_{n\to\infty} X_n(\omega)$ *converges for all $\omega \in \Omega$. We denote the limit by $Y(\omega)$. Then Y is also a measurable function.*

The notion of independence of events and random variables are the same as in the previous section.

Example 3.4. *Let us consider the probability space $(\Omega_\infty, \mathcal{F}, P)$ in Example 3.2. Let*

$$X_k(\omega) = x_k \quad \textit{if } \omega = (x_1, \dots, x_k, \dots,).$$

Then $\{X_k\}_{k=1}^\infty$ are independent random variables.

Exercise 7. *Let X_k be the random variables in Example* 3.4. *Let*

$$S = \left\{ \omega \in \Omega_\infty \mid \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = 3.5 \right\}.$$

Show that $S \in \mathcal{F}$[‡]*.*

We define the expectation of X as the integration of X over Ω in the Lebesgue sense.

Definition 3.5 (Lebesgue integral)**.** *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$.*
(1) [The case where $X \geq 0$]
(i) *The case where X is a discrete type random variable: That is,*

$$\{X(\omega) \mid \omega \in \Omega\} = \{a_1, \dots, a_N\}.$$

In this case, we define the expectation of X in a similar way as in the previous section.

$$E[X] := \sum_{i=1}^N a_i P(X = a_i).$$

(ii) *The case where $X \geq 0$:*
Let

$$X_n(\omega) = \begin{cases} 0 & \textit{if } 0 \leq X(\omega) < \frac{1}{2^n} \\ \frac{k}{2^N} & \textit{if } \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n},\ 0 < k \leq 2^n n - 1 \\ n & \textit{if } X(\omega) \geq n \end{cases} \tag{3.2}$$

Then $X_n$is also measurable function and non-negative discrete type random variable. So we have already defined $E[X_n]$. We define

$$E[X] := \lim_{n\to\infty} E[X_n].$$

[‡] The strong law of large number asserts that $P(S) = 1$.

Note that $E[X]$ maybe ∞.

(2) [General case] *We consider real valued measurable function X (So X may take positive values and negative values). Define*

$$X^+(\omega) = \max(X(\omega), 0)(= \textit{positive part of } X),\ X^-(\omega) = \max(-X(\omega), 0)(= \textit{negative part of } X).$$

Note that

$$X(\omega) = X^+(\omega) - X^-(\omega) \quad \textit{for all } \omega.$$

When $E[X^+] < \infty, E[X^-] < \infty$, we define the expectation of X by

$$E[X] = E[X^+] - E[X^-].$$

We may denote $E[X]$ by $\int_\Omega X(\omega)P(d\omega)$. Also we define

$$L^1(\Omega, \mathcal{F}, P) = \{X : \Omega \to \mathbb{R} \mid X \textit{ is a random variable such that } E[X^+] < \infty \textit{ and } E[X^-] < \infty\}.$$

Remark 3.6. (1) *By Exercise 5, $X^+, X^-, |X|$ are mesurable functions. The condition $E[X^+] < \infty$ and $E[X^-] < \infty$ is equivalent to $E[|X|] < \infty$.*

(2) *We may denote $L^1(\Omega, \mathcal{F}, P)$ by $L^1(\Omega, P)$ or $L^1(\Omega)$ simply.*

The following is very basic properties of the expectation.

Theorem 3.7. (1)[Linearity of expectation] *Let X, Y be random variables. Then for any $\alpha, \beta \in \mathbb{R}$, $\alpha X + \beta Y$ is also a measurable function and*

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

(2) *Let us define*

$$L^p(\Omega, \mathcal{F}, P) = \{X : \Omega \to \mathbb{R} \mid \int_\Omega |X(\omega)|^p P(d\omega) < \infty\}.$$

We use the notation

$$\|X\|_{L^p} = \left(\int_\Omega |X(\omega)|^p P(d\omega)\right)^{1/p}.$$

If $X, Y \in L^p(\Omega, \mathcal{F}, P)$, then $X + Y \in L^p(\Omega, \mathcal{F}, P)$ and

$$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p} \qquad (\textit{Minkowski's inequality})$$

(3) [Hölder's inequality] *Let $p > 1, q > 1$ be positive numbers with $\frac{1}{p} + \frac{1}{q} = 1$. For any $X \in L^p(\Omega, \mathcal{F}, P)$, $Y \in L^q(\Omega, \mathcal{F}, P)$, we have*

$$\|XY\|_{L^1} \le \|X\|_{L^p}\|Y\|_{L^q}.$$

The following limit theorem in Lebesgue integral is very important.

Theorem 3.8 (Monotone Convergence Theorem)**.** *Let X_n be random variables such that*

(i) $0 \le X_1(\omega) \le X_2(\omega) \le \cdots X_n(\omega) \le \cdots$ *for all ω,*

(ii) $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ *for all* ω.

Then

$$\lim_{n\to\infty} E[X_n] = E[X].$$

Theorem 3.9 (Lebesgue's dominated convergence theorem)**.** *Let* X_n, Y *be random variables such that*

(i) $|X_n(\omega)| \le Y(\omega)$ *for all* ω *and* n.

(ii) $Y \in L^1(\Omega, P)$.

(iii) $\lim_{n\to\infty} X_n(\omega) = Y(\omega)$ *for all* ω.

Then

$$\lim_{n\to\infty} E[X_n] = E[X].$$

Remark 3.10. *In the case where* $\Omega = [0,1], \mathcal{F} = \mathfrak{M}_L, P = |\cdot|)$*, we have two definitions of integration of a function* $X : [0,1] \to \mathbb{R}$*. That is, Riemann integral and Lebesgue integral. We can prove that if* $f : [0,1] \to \mathbb{R}$ *is a bounded Riemannian integrable function then* f *is Lebesgue integrable and the two integrals coincide. That is, the Riemannian integral*

$$\lim_{|\Delta|\to 0} \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)$$

$$x_i \le \xi_i \le x_{i+1}, \quad \Delta = \{0 = x_0 < \cdots < x_n = 1\}, \quad |\Delta| = \max_i (x_{i+1} - x_i).$$

is equal to the Lebesgue integral

$$\lim_{n\to\infty} \sum_{k=-\infty}^{\infty} \frac{k}{n} \left| \left\{ x \mid \frac{k}{n} \le f(x) < \frac{k+1}{n} \right\} \right|.$$

But the converse is not true. That is, the Lebesgue integrable function may not be Riemannian integrable.

Now we explain two law of large numbers. One is "weak law of large numbers"(=WLLN) and the second is "strong law of large numbers"(SLLN). First, we explain WLLN.

Lemma 3.11. *Let* $\{Z_i\}_{i=1}^n$ *be independent random variables whose means and variances are finite. Moreover we assume that the means and the variances coincide,* $E[Z_i] = m$ *and* $V[Z_i] = \sigma^2$*. Then*

$$P\left(\left| \frac{Z_1 + \cdots Z_n}{n} - m \right| \ge \delta \right) \le \frac{\sigma^2}{n\delta^2}. \tag{3.3}$$

Proof. We use the Chebyshev inequality: For any random variable X and $r > 0$,

$$P(|X| \ge r) \le \frac{E[|X|^2]}{r^2}.$$

Using $E[(Z_i - m)(Z_j - m)] = 0$, and applying the Chebyshev inequlity in the case where

$$X = \frac{(Z_1 - m) + \cdots + (Z_n - m)}{n}, \quad r = \delta$$

we get the theorem. □

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

Theorem 3.12. *Assume that $\{Z_i\}_{i=1}^{\infty}$ are independent random variables and their means and variances are finite and $E[Z_i] = m$, $V[Z_i] = \sigma^2$. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{Z_1 + \cdots Z_n}{n} - m\right| \geq \delta\right) = 0. \tag{3.4}$$

Next we state SLLN.

Theorem 3.13 (Kolmogorov)**.** *Let $\{Z_i\}_{i=1}^{\infty}$ be i.i.d. (=independent identically distributed) random variables. Assume that their mean is finite $E[X_i] = m$. Then*

$$P\left(\left\{\omega \mid \lim_{n\to\infty} \frac{Z_1(\omega) + \cdots Z_n(\omega)}{n} = m\right\}\right) = 1.$$

The proof of the above theorem is not easy. But the proof of the following is not so difficult.

Theorem 3.14. *Let $\{Z_i\}_{i=1}^{\infty}$ be independent random variables such that there exists $0 < K < \infty$ such that*

$$E[Z_i] = m, \quad E[|Z_i|^k] \leq K \qquad \textit{for all } 1 \leq k \leq 4, i = 1, 2, \ldots$$

Then

$$P\left(\left\{\omega \mid \lim_{n\to\infty} \frac{Z_1(\omega) + \cdots Z_n(\omega)}{n} = m\right\}\right) = 1.$$

Why strong and weak? This is because of the following result.

Proposition 3.15. *Assume that*

$$P\left(\left\{\omega \mid \lim_{n\to\infty} Y_n(\omega) = m\right\}\right) = 1.$$

Then for any $\delta > 0$,

$$\lim_{n\to\infty} P\left(\{\omega \mid |Y_n(\omega) - m| \geq \delta\}\right) = 0.$$

But the converse is not necessarily true.

Exercise 8. *Prove Proposition* 3.15 *applying Theorem* 3.9 *to functions $X_n(\omega) = 1_{A_n}(\omega)$, where $A_n = \{\omega \mid |Y_n(\omega) - m| \geq \delta\}$.*

4 Entropy

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

Definition 4.1 (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*

$$H(P) = -\sum_{i=1}^{N} P(\{A_i\}) \log P(\{A_i\}). \tag{4.1}$$

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, $\log_e$.

Example 4.2. (1) **Coin tossing:**
$E = \{H, T\}$ and $P_1(\{H\}) = P_1(\{T\}) = 1/2$. We have $H(P_1) = \log 2$.
(2) **Dice:** *$E = \{1, 2, 3, 4, 5, 6\}$. $P_2(\{i\}) = 1/6$ $(1 \leq i \leq 6)$. Then we have $H(P_2) = \log 6$.*
(3) **Unfair Dice:** *$E = \{1, 2, 3, 4, 5, 6\}$. $P_3(\{1\}) = 9/10, P_3(\{i\}) = 1/50$ $(2 \leq i \leq 6)$.*

$$H(P_3) = \log\left[\left(\frac{10}{9}\right)^{9/10}(50)^{1/10}\right] \leq \log\left(\frac{10}{9} \cdot \frac{3}{2}\right) < \log 2 = H(P_1) \tag{4.2}$$

Exercise 9. *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 4.3. *Suppose that $|E| = N$. Then for any probability distribution P, we have*

$$0 \leq H(P) \leq \log N. \tag{4.3}$$

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 $1 \leq i \leq N$.

We refer the proof to the proof of Theorem 5.1 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.

Exercise 10. *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.*

5 Shannon and McMillan's theorem

Suppose we are given a set of numbers $A = \{1, \dots, N\} \subset \mathbb{N}$. We call A the alphabet and the element is called a letter. A finite sequence $\{\omega_1, \omega_2, \dots, \omega_n\}$ ($\omega_i \in A$) is called a sentence with the length n. The set of the sentences whose length are n is the product space $A^n := \{(\omega_1, \dots, \omega_n) \mid \omega_i \in A\}$. 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:

$$H(P) = -\sum_{i=1}^{N} P(\{i\}) \log_N P(\{i\}). \tag{5.1}$$

We can prove that

Theorem 5.1. *For every P, $0 \leq H(P) \leq 1$ 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 5.2. *Let $g(x) = x \log x$, or $g(x) = -\log x$. Then for any $\{m_i\}_{i=1}^N$ with $m_i \geq 0$ and $\sum_{i=1}^N m_i = 1$ and nonnegative sequence $\{x_i\}_{i=1}^N$, we have*

$$g\left(\sum_{i=1}^{N} m_i x_i\right) \leq \sum_{i=1}^{N} m_i g(x_i). \tag{5.2}$$

Furthermore, when $m_i > 0$ for all i, the equality of (5.2) *holds if and only if $x_1 = \cdots = x_N$.*

We define for a nonnegative sequence $\{p_i\}_{i=1}^N$,

$$H(p_1, \dots, p_N) = -\sum_{i=1}^{N} p_i \log p_i. \tag{5.3}$$

Proof of Theorem 5.1. First, we consider the lower bound. Applying (5.2) to the case where $m_i = x_i = p_i$ and $g(x) = -\log x$, we have

$$\begin{aligned} H(p_1, \dots, p_N) &\geq -\log\left(\sum_{i=1}^{N} p_i^2\right) \\ &\geq -\log 1 = 0. \end{aligned} \tag{5.4}$$

Clearly, in (5.4), the equality holds if and only if $p_i = 1$ for some i. Next, we consider the upper bound. By applying Lemma 5.2 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

$$g\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \leq \frac{1}{N}\sum_{i=1}^{N} g(p_i). \tag{5.5}$$

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 5.2, the equality holds iff $p_i = 1/N$ for all i. □

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.

Theorem 5.3 (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*

$$P\Big(\psi(\varphi(X_1, \dots, X_n)) \neq (X_1, \dots, X_n)\Big) \leq \varepsilon. \tag{5.6}$$

The map $\varphi : A^n \to A^k$ is called an encoder and the map $\psi : A^k \to A^n$ is called a decoder. The probability $P\Big(\psi(\varphi(X_1, \dots, X_n)) \neq (X_1, \dots, X_n)\Big)$ is called the error probability. k/n is called the coding rate.

Proof of Theorem 5.3. 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 $\{a_i\}_{i=1}^{n}$,

$$P_n\left(\{\omega_1 = a_1, \dots, \omega_n = a_n\}\right) = \prod_{i=1}^{n} P\left(\{a_i\}\right). \tag{5.7}$$

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

$$\begin{aligned} m &= E[Z_i] = -\sum_{i=1}^{n} P\left(\{\omega_i\}\right) \log_n P\left(\{\omega_i\}\right) = H(P) \\ \sigma^2 &= E[(Z_i - E[Z_i])^2] = \sum_{i=1}^{n} \left(\log_N p_i\right)^2 p_i - H(P)^2. \end{aligned} \tag{5.8}$$

Take $\delta > 0$ such that $R > H(P) + \delta$. By Lemma 3.11,

$$P_n\left(\frac{1}{n}\sum_{i=1}^{n}(-\log_N P(\{\omega_i\})) \geq H(P) + \delta\right) \leq \frac{\sigma^2}{n\delta^2}. \tag{5.9}$$

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

$$P_n\left(\frac{1}{n}\sum_{i=1}^{n}(-\log_N P(\{\omega_i\})) \geq H(P) + \delta\right) \leq \varepsilon \qquad \text{for all } n \geq M. \tag{5.10}$$

Noting

$$\begin{aligned} &\left\{(\omega_1, \dots, \omega_n) \;\middle|\; \frac{1}{n}\sum_{i=1}^{n}(-\log_N P(\{\omega_i\})) < H(P) + \delta\right\} \\ &= \left\{(\omega_1, \dots, \omega_n) \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) > N^{-n(H(P)+\delta)}\right\} \\ &\subset \left\{(\omega_1, \dots, \omega_n) \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) \geq N^{-nR}\right\} =: C_n, \end{aligned} \tag{5.11}$$

by (5.10), we have, for $n \geq M$,

$$\begin{aligned} & P\left((X_1, \ldots, X_n) \in C_n\right) \\ & \quad = P_n\left(\left\{(\omega_1 \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \\ & \quad \geq P_n\left(\left\{(\omega_1, \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) \geq N^{-n(H(P)+\delta)}\right\}\right) \geq 1 - \varepsilon \end{aligned} \tag{5.12}$$

On the other hand, we have

$$|C_n| N^{-nR} \leq P_n\left(\left\{(\omega_1, \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \leq 1 \tag{5.13}$$

Hence we have

$$|C_n| \leq N^{nR}. \tag{5.14}$$

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

$$P\left(\psi(\varphi(X_1, \ldots, X_n)) = (X_1, \ldots, X_n)\right) \;\;\geq\;\; P\left((X_1, \ldots, X_n) \in C_n\right) \geq 1 - \varepsilon. \tag{5.15}$$

This completes the proof. □

References

[1] P. Billingsley, Probability and measure, John Wiley & Sons, New York, 1979

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

[3] N. Abramson, Informantion theory and coding. McGraw-Hill Book Co., New York-Toronto-London 1963

[4] J.S. Rosenthal, A first look at rigorous probability theory, World Scientific, 2006.

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

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