Probability and Entropy

Shigeki Aida
Osaka University
June 30, 2003

1 Introduction

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

Definition 1.1 (Shannon) *Let us consider a finite set $E = \{A_1, \dots, A_n\}$. A nonnegative function P on E is called a probability distribution if $\sum_{i=1}^n P(E_i) = 1$. Then for this probability distribution P, we define the entropy by*

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

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,* $\log_e$ *(***nat***). (*$\log_2$*...***bit***). We define for a nonnegative sequence $\{p_i\}_{i=1}^n$,*

$$H(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1.2}$$

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

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

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. Actually the discovery is older than that of the information theory, of course. In statistical mechanics, $S = k \log N$ (Boltzmann), where N stands for the number of all possible states

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

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

The equality holds if and only if P is the uniform distribution, namely, $P(E_i) = 1/n$ for all $1 \le i \le n$.

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

The notion of entropy is used to solve the following problem

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

In order to get into the detail, we prepare the notion of probability theory.

2 Basic notions in probability theory

In general, mathematically, a probability space is defined to be a measure space whose total measure equals 1. We refer the audiences to some text books for the precise definition. We give very rough definition of it.

Definition 2.1 *Let Ω be a set and $\mathcal{F}$ be a set of some subsets of Ω. A nonnegative function $P : \mathcal{F} \to \mathbb{R}$ is called a probability neasure on P if $P(\Omega) = 1$ holds. A function on Ω is called a random variable.*

For a random variable X, we can define the probability distribution μ_X on $\mathbb{R}$. Below, we consider the following cases only.

Definition 2.2 (1) *X is a discrete type random variable, that is, X takes finite number values $\{a_1, \ldots, a_n\}$. Then $p_i := P(\{\omega \in \Omega \mid X(\omega) = a_i\})$ satisfies that $\sum_{i=1}^{N} p_i = 1$. The probability distribution μ_X on $\{a_1, \cdots, a_n\}$ such that $\mu_X(\{a_i\}) = p_i$ is called the distribution of X.*
(2) *X is a continuous type random variable in the sense that 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 $[a, b]$. In this case, the distribution of X is the probability distribution μ_X on $\mathbb{R}$ which has the density function $f(x)$.

Definition 2.3 *For a random variable X, we denote the expectation and the variance by m and σ^2 respectively. Namely,*
(i) The case where X is a discrete type random variable and takes values $a_1, \ldots, a_n$:

$$m = E[X] = \sum_{i=1}^{n} a_i P(X = a_i), \tag{2.1}$$

$$\sigma^2 = E[(X - m)^2] \sum_{i=1}^{n} (a_i - m)^2 P(X = a_i). \tag{2.2}$$

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

$$m = E[X] = \int_{\mathbb{R}} xf(x)dx, \tag{2.3}$$

$$\sigma^2 = E[(X-m)^2] = \int_{\mathbb{R}} (x-m)^2 f(x)dx. \tag{2.4}$$

Definition 2.4 (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:*

$$P\left(X_{i_1} \in [a_1, b_1], \cdots, X_{i_m} \in [a_m, b_m]\right) = \prod_{i=1}^m P(X_{i_k} \in [a_k, b_k]). \tag{2.5}$$

Definition 2.5 *Let $A = \{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R}$ and consider a probability distribution μ on A. Let $\{X_i\}_{i=1}^\infty$ be independent random variables and the probability distribution of X_i is equal to μ for all i. Then $\{X_i\}$ is said to be i.i.d. (=independent and identically distributed) 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 $\{x_1, x_2, \ldots, x_n\}$ $(x_i \in A)$ is called a sentense with the length n. The set of the sentenses 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 any P, $0 \le H(P) \le 1$. The equality holds if and only if P is the uniform distribution, that is, $p_i = 1/N$ for all i.*

Lemma 3.2 *Let $f(x) = x \log x$. 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*

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

Furtheremore, 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. By Lemma 3.2, for any nonnegative probability distribution $\{p_i\}$, we have

$$f\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \le \frac{1}{N}\sum_{i=1}^{N} f(p_i). \tag{3.3}$$

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

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

Thus, $-\sum_{i=1}^{N} p_i \log p_i \le \log N$ and $-\sum_{i=1}^{N} p_i \log_N p_i \le 1$. 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 the problem coding the sequence of letters.

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. Actually, if we get just the first letter, we immediately know that subsequent all letters are 1. Namely, we can encode all sentenses, whatever the lentgh are, to just one letter. Note that the entropy of P is 0.

(2) Suppose that $N \ge 2$ 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 $N^k \ge 2^n (\Longleftrightarrow \frac{k}{n} \ge \log_N 2 = H(P))$, then all possible sentences whose lengths are n can be encoded into the sentenses whose lengths are $k(\le n)$. Also the decode is also possible. More precisely, we can prove the following claim.

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

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \ne (X_1, \ldots, X_n)\Big) = 0. \tag{3.4}$$

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.

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 $n \ge M$ and k satisfying that $\frac{k}{n} \ge R$, there exists $\varphi : A^n \to A^k$ and $\psi : A^k \to A^n$ such that*

$$P\Big(\psi(\varphi(X_1, \ldots, X_n)) \ne (X_1, \ldots, X_n)\Big) < \varepsilon. \tag{3.5}$$

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*

$$P\left(\left|\frac{Z_1 + \cdots Z_n}{n} - m\right| \geq \delta\right) \leq \frac{\sigma}{n\delta^2}, \tag{3.6}$$

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

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

Theorem 3.5 *Assume the same assumption on $\{Z_i\}$. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{Z_1 + \cdots Z_n}{n} - m\right| \geq \delta\right) = 0. \tag{3.7}$$

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 on A^n such that for any $\{a_i\}_{i=1}^n$,

$$P_n\left(\{\omega_1 = a_1, \ldots, \omega_n = a_n\}\right) = \prod_{i=1}^{n} P\left(\{a_i\}\right). \tag{3.8}$$

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

$$\begin{aligned} m &= E[Z_i] = -\sum_{i=1}^{n} P\left(\{\omega_i\}\right) \log_n P\left(\{\omega_i\}\right) = H(P) \\ \sigma &= E[(Z_i - E[Z_i])^2] = \sum_{i=1}^{n} \left(\log_N p_i\right)^2 p_i - H(P)^2. \end{aligned} \tag{3.9}$$

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

$$P_n\left(\frac{1}{n}\sum_{i=1}^{n}\left(-\log_N P(\{\omega_i\})\right) \geq H(P) + \delta\right) \leq \frac{\sigma}{n\delta^2}. \tag{3.10}$$

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

$$P_n\left(\frac{1}{n}\sum_{i=1}^{n}\left(-\log_N P(\{\omega_i\})\right) \geq H(P) + \delta\right) \leq \varepsilon \qquad \text{for all } n \geq M. \tag{3.11}$$

Noting

$$\begin{aligned} &\left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \frac{1}{n}\sum_{i=1}^{n}\left(-\log_N P(\{\omega_i\})\right) < H(P) + \delta\right\} \\ &\quad = \left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) > N^{-n(H(P)+\delta)}\right\} \\ &\quad \subset \left\{(\omega_1, \ldots, \omega_n) \;\middle|\; \prod_{i=1}^{n} P(\{\omega_i\}) \geq N^{-nR}\right\} =: S_n, \end{aligned} \tag{3.12}$$

by (3.11), we have, for $n \geq M$,

$$\begin{aligned} & P\left((X_1, \ldots, X_n) \in S_n\right) \\ & \quad = P_n\left(\left\{(\omega_1 \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \\ & \quad \geq P\left(\left\{(\omega_1, \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-n(H(P)+\delta)}\right\}\right) \geq 1 - \varepsilon \end{aligned} \tag{3.13}$$

On the other hand, we have

$$\begin{aligned} 1 &= P_n\left(\left\{(\omega_1, \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) \geq N^{-nR}\right\}\right) \\ &\quad + P\left(\left\{(\omega_1, \ldots, \omega_n) \in A^n \;\middle|\; \prod_{i=1}^n P(\{\omega_i\}) < N^{-nR}\right\}\right) \\ &\geq |S_n| N^{-nR}. \end{aligned} \tag{3.14}$$

Consequently we have

$$|S_n| \leq N^{nR}. \tag{3.15}$$

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

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

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

$$P\left(\psi(\varphi(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n)\right) \leq P\left((X_1, \ldots, X_n) \notin S_n\right) \leq \varepsilon. \tag{3.16}$$

This completes the proof. □

4 Entropy and central limit theorem

Let $\{X_i\}_{i=1}^\infty$ be i.i.d. such that $E[X_i] = 0$ and $E[X_i^2] = 1$. 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 with mean 0 and variance 1 and it is denoted by $N(0,1)$. A standard proof of CLT is the proof using the characteristic function of S_n, $E[e^{-1tS_n}]$. In this section, we show a proof using the entropy of probability distributions.

Below, we assume

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

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

In the previous sections, we define the entropy for the discrete type probability distribution. 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 L by

$$H(P) = H(X) = -\int_{\mathbb{R}} f(x)\log f(x)dx, \tag{4.2}$$

$$L(P) = L(X) = \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)}dx. \tag{4.3}$$

The following hold.

Theorem 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* $\dfrac{1}{a}\displaystyle\int_{\mathbb{R}} f\left(\frac{x}{a} - y\right) g(y)dy$

(2) (Gibbs's lemma) *Let $f(x)$ be a density of a probability whose mean* 0 *and the variance is* 1*, that is,*

$$\int_{\mathbb{R}} xf(x)dx = 0, \tag{4.4}$$

$$\int_{\mathbb{R}} x^2 f(x) = 1. \tag{4.5}$$

Then we have,

$$0 \le H(f) \le H(G). \tag{4.6}$$

The equality holds iff $f(x) = G(x)$.

(3) (Shannon-Stam's inequality) *Let X, Y be independent random variables whose density functions satisfy* (4.4) *and* (4.5)*. Then for $a, b \in \mathbb{R}$ with $a^2 + b^2 = 1$, we have*

$$a^2 H(X) + b^2 H(Y) \le H(aX + bY). \tag{4.7}$$

The equality holds iff the **laws** *of X and Y are $N(0,1)$.*

(4) (Fisher information inequality) *Let X, Y be independent random variables whose density functions satisfy* (4.4) *and* (4.5)*. Then for $a, b \in \mathbb{R}$ with $a^2 + b^2 = 1$,*

$$(a+b)^2 L(X+Y) \le a^2 L(X) + b^2 L(Y). \tag{4.8}$$

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

$$\left(\int_{\mathbb{R}} |f(x) - G(x)|dx\right)^2 \leq 2\left(H(G) - H(f)\right) \tag{4.9}$$

"Proof" (2) is proved by Jensen's inequality. (4) implies (3). By (1), S_n S_n has the density function, say, $f_n(x)$. By (4),

$$L(f_n) \leq L(G) = \frac{1}{4}. \tag{4.10}$$

Also, by (4.7),

$$H(S_n) \leq H(S_{n+1}) \leq H(G)(= \frac{1}{2}(1 + \log(2\pi))) \qquad \text{for all } n \tag{4.11}$$

By (4.9), $\lim_{n\to\infty} f_n(x) = f(x)$ exists for all x.

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

Kinetic theory of rarefied gases:

$(v_x^i(t), v_y^i(t), v_z^i(t))$:the velocity of the i-th molecule $(1 \leq i \leq N)$. N denotes the number of mlecules. 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, $f_t(v_x, v_y, v_z)dv_xdv_ydv_z$ and derived the following his H-theorem:

Theorem 5.1 (Boltzmann) *Let*

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

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 incraeses.

(2) Some people raised questions about the H-theorem.

(i) Newton's equation of motion for the particles $\mathbf{x}(t) = (x_i(t))_{i=1}^N$ moving in a potential U reads as follows:

$$\begin{aligned} m_i \frac{d^2}{dt^2} x_i(t) &= -\nabla U(\mathbf{x}(t)) && (5.1)\\ x_i(0) &= x_{i,0}. && (5.2)\\ \dot{x}_i(0) &= v_i. && (5.3) \end{aligned}$$

The time reversed curve $\mathbf{x}(-t)$ is the solution with initial velocity $-v_i$. Clearly, it is impossible that both entropy of $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}(-t)\}$ increase. This is a contradiction.

(ii) Poincaré's recurrence theorem:

The reason is in the statistical treatment.

From now on, we consider stochastic dynamics which is called Markov chains and prove that the entropy increases.

Example 5.3 There is a **datum** on the weather forecast in some local area.

today tomorrow

fine fine weather ...$\frac{2}{3}$

fine rain.....$\frac{1}{3}$

rain fine.....$\frac{1}{2}$

fine rain.....$\frac{1}{2}$

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 漸化式：

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

So we obtain

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

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 $\{x(n)\}_{n=0,1,2,\ldots}$ on S. Let $\{p_{ij}\}_{i,j\in E}$ be nonnegative numbers satisfying that*

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

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 $(i \in E)$ 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 time t distribution $\pi(n) = (\pi_1(n), \ldots, \pi_N(n))$ is determined 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) $\sum_{i=1}^{N} p_{ij} = 1$ *for all* $1 \le j \le N$.

(A2) *There exists* $n_0 \in \mathbb{N}$ *such that* $p_{ij}^{n_0} > 0$ *for all* $i, j \in S$.

Then for any initial distribution π, *we have*

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

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

(A1) holds if $p_{ij} = p_{ji}$ for all $i, j \in E$. This theorem can be proved by using the entropy of $\pi(n)$. Recall

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

Lemma 5.6 (1) *Assume* (A1). *Then for any* π *and* $n \in \mathbb{N} \cup \{0\}$, $H(\pi P^{n+1}) \ge H(\pi P^n)$ *for all* $n \in \mathbb{N}$.
(2) (Irreducibility of the Markov chain) *Assume* (A1) *and* (A2). *Then for any* $\pi \ne (1/N, \dots, 1/N)$ *and for all* $n \ge n_0$, *it holds that*

$$H(\pi P^n) > H(\pi). \tag{5.8}$$

Proof. (1) It suffices to prove the case $n = 0$.

$$\begin{aligned} H(\pi P) &= -\sum_{i=1}^{N} (\pi P)_i \log (\pi P)_i \\ &= -\sum_{i=1}^{N} \left(\sum_{k=1}^{N} \pi_k p_{ki} \right) \log \left(\sum_{k=1}^{N} \pi_k p_{ki} \right). \end{aligned} \tag{5.9}$$

Since $\sum_{k=1}^{N} p_{ki} = 1$, by Lemma 3.2,

$$\left(\sum_{k=1}^{N} \pi_k p_{ki} \right) \log \left(\sum_{k=1}^{N} \pi_k p_{ki} \right) \le \sum_{k=1}^{N} p_{ki} \pi_k \log \pi_k. \tag{5.10}$$

(5.9) and (5.10) implies $H(\pi P) \ge H(\pi)$. Now we consider (2). It is obvious that $p_{ij}^{(n)} > 0$ for any $n \ge n_0$ and any i, j. By noting this, (2) is proved by the same method and by the last assertion of Lemma 3.2. □

By using this lemma, we prove Theorem 5.5.

Proof of Theorem 5.5 We prove that $\lim_{n\to\infty} H(\pi(n)) = \log N (= H(1/N, \dots, 1/N))$. Since $\pi(n)$ moves in a bounded subset in $\mathbb{R}^N$, there exist the accumulation points. That is, there exist $x = (x_1, \dots, x_N)$ and a subsequence $\{\pi(n(k))\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \pi(n(k)) = x$. x is also a probability on E and satisfies that $H(x) = \lim_{k\to\infty} H(\pi(n(k)))$. Since $H(xP^{n_0}) = \lim_{k\to\infty} H(\pi(n(k) + n_0))$ and $\{H(\pi(n))\}_n$ is an increasing sequence, $H(x) = H(xP^{n_0})$. By Lemma 5.6 (2), we have $x = (1/N, \dots, 1/N)$. This completes the proof. □