J. KSIAM **Vol.2**, No.2, 125-135, 1998

ON MONOTONICITY OF ENTROPY

Youngsoo Lee

Abstract. In this paper we define the entropy rate and stationary Markov chain and we show the monotonicity of entropy per element and prove that the random tree T_n grows linearly with n.

1. Introduction

The asymptotic equipartition property (A.E.P) states that $\frac{1}{n}\log\frac{1}{p(X_1, X_2, \cdots, X_n)}$ is closed the entropy H, where $X_1, X_2, \cdots, X_n$ are independent identically distribution (i.i.d) random variables and $p(X_1, X_2, \cdots, X_n)$ is the probability of observing the sequence $X_1, X_2, \cdots, X_n$, $p(X_1, X_2, \cdots, X_n)$ is close to $\sum 2^{-nH}$ with high probability.

Let $B_\delta^{(n)} < æ^n$ be any set with $\Pr\{B_\delta^{(n)}\} \geq 1-\delta$ and let $X_1, X_2, \cdots X_n$ be i.i.d. Then the theorem 3.4 see that $\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta$ for n sufficiently large.

When the limit exists, we define two definitions of entropy rate for a stochastic process as follows

$H(æ) = \lim_{n\to\infty}\frac{1}{n}H(X_1, X_2, \cdots X_n)$,

$H'(æ) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \cdots, X_1)$.

In particular, for a stationary stochastic process,

$$H(æ) = H'(æ).$$

In this paper we will show that the thorem 4.4 is established. In detail, the contents of this paper is as follows. In 2, we explain the terminology of typical set and entropy. In 3, we define typical set and we prove the theorem 3.4. In 4, we define entorpy rate and we prove the theorem 4.4, theorem 4.5.

Key words and phrases : Entropy, information theory

This research was supported by Woosuk University Research Grant.

Typeset by $\mathcal{AMS}$-TeX

2. Preliminary

Let X be a discrete random variable with alphabet æ and probability mass function by $p(x)$. Then the entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in X} p(x) \log p(x).$$

We often denote the $H(X)$ as $H(p)$ and entropy is expressed in bits.

$H(X, Y)$ of a pair of discrete random variable (X, Y) with a joint distribution $p(x, y)$ is called the joint entropy and it is defined as

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y).$$

Also $H(Y|X)$ is called the conditional entropy as

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) H(Y|X = x) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x) \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x). \end{aligned}$$

The relative entropy between two probability mass function $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.$$

The relative entropy between the joint distributions and the product distributions $p(x), p(y)$ are called the mutual information and it is represented as

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

The following properties are well-known ([2],[3],[9])

(i) $H(X) \geq 0$

(ii) For any two random variables $X, Y, H(X|Y) \leq H(X)$

(iii) $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$.

The random variables X_i are independent iff equality holds.

(iv) $H(X) \leq \log |$ æ$|$ where X is uniformly distributed over æ iff equality holds.

The joint entropy and conditional entropy can make the chain rule as follows.

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

Indeed,

$$\begin{aligned} H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\ \log\, p(x,y) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\ \log\, p(x)p(y|x) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \\ &= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

Proposition 1.1. $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$.

Proof.

$$\begin{aligned} H(X,Y|Z) &= \sum_{z\in Z} p(z)H(X,Y|Z) \\ &= -\sum_{z\in Z} p(z)\sum_{x\in X}\sum_{y\in Y} p(x,y|z)\ \log\, p(x,y|z) \\ &= -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\ \log\, \{p(x|z)\cdot p(y|x,z)\} \\ &= -\sum_{x\in X}\sum_{z\in Z} p(x,z)\log\, p(x|z) \\ &\quad -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\, p(y|x,z) \\ &= H(X|Z) + H(Y|X,Z). \end{aligned}$$

Let $I(X;Y)$ be a mutual information. Then by the definition,

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\ \log\, \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\ \log\, \frac{p(y)p(x|y)}{p(x)p(y)} \\ &= -\sum_{x,y} p(x,y)\ \log\, p(x) + \sum_{x,y} p(x,y)\ \log\, p(x|y) \\ &= -\sum_{x} p(x)\ \log\, p(x) - (-\sum_{x,y}\ \log\, p(x,y)\ \log\, p(x|y)) \\ &= H(X) - H(X|Y). \end{aligned}$$

By symmetry, $I(X;Y) = H(Y) - H(Y|X)$. Since $H(X,Y) = H(Y) - H(Y|X), I(X;Y) = H(X) + H(Y) - H(X,Y)$.

Finally we obtain that

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

The relationship between $H(X), H(Y), H(X,Y), H(X|Y), H(Y|X)$ and $I(X;Y)$ is expressed in a Venn diagram.

$$\begin{aligned}&I(X;Y) = H(X) - H(X|Y), \quad I(X;Y) = H(Y) - H(Y|X),\\ &I(X;Y) = H(X) + H(Y) - H(X,Y),\\ &I(X;Y) = I(Y;X), \quad I(X;X) = H(X).\end{aligned}$$

3. The smallest probable set

The asymptotic equipartition property (AEP) is a direct consequence of the weak law of large numbers. If $X_1, X_2, \cdots, X_n$ are independent, identically distributed (i.i.d.) random variables and $p(X_1, X_2, \cdots, X_n)$ is the probability of observing the sequence $X_1, X_2, \cdots, X_n$, then the AEP states that $\frac{1}{n}\log\frac{1}{p(X_1, X_2, \cdots, X_n)}$ is close to the entropy H. Indeed, since the X_i are i.i.d. So are log $p(X_i)$.

Hence $-\frac{1}{n}$ log $p(X_1, X_2, \cdots, X_n) = -\frac{1}{n}\sum_i$ log $p(X_i)$
$= -E$ log $p(X)$ in probability $= H(X)$.

We define that the typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \cdots, x_n) \in$ æ with the following properties :

$$2^{-n(H(x)+\epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(x)-\epsilon)}.$$

We obtain that the typical set $A_\epsilon^{(n)}$ has the following properties ([3], [6], [9]).

Proposition 3.1. *1. If $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$, then*

$$H(x) - \epsilon \leq -\frac{1}{n} \ \mathit{log}\ p(x_1, x_2, \cdots, x_n) \leq H(x) + \epsilon.$$

Also $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for n sufficiently large.

2. $|A_\epsilon^{(n)}| \leq 2^{n(H(x)+\epsilon)}$, where $|A|$ denotes the number of elements in the set A.

$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(x)-\epsilon)}$ *for* n *sufficiently large.*

Proof. 1. Since $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$,

$$2^{-n(H(x)+\epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(x)-\epsilon)}.$$

Taking the log with base 2 to both sides,

$$-n(H(x)+\epsilon) \leq \ \log{}_2 p(x_1, x_2, \cdots, x_n) \leq -n(H(x)-\epsilon).$$

Therefore $H(x) - \epsilon \leq -\dfrac{1}{n} \log p(X_1, X_2, \cdots, X_n) \leq H(x) + \epsilon$ since the probability of the event $(X_1, X_2, \cdots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. For any $\delta > 0$ there exist an n_0, such that for all $n \geq n_0$,

$$\text{Pr } \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \cdots, X_n) - H(X) \right| < \epsilon \right\} > 1 - \delta.$$

2. $1 = \sum_{x \in X^n} P(x) \geq \sum_{x \in A_\epsilon^{(n)}} P(x) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(x)+\epsilon)}$
$= 2^{-n(H(x)+\epsilon)} |A_\epsilon^{(n)}|$,
Finally, since Pr $\{A_\epsilon^{(n)}\} > 1 - \epsilon$, $\ 1 - \epsilon <$ Pr $\{A_\epsilon^{(n)}\}$
$\leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(x)-\epsilon)} = 2^{-n(H(x)-\epsilon)} |A_\epsilon^{(n)}|$.
Hence $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(x)-\epsilon)}$.

Now we divide all sequences in æ^n into two sets :

One is the typical set $A_\epsilon^{(n)}$ and the other is complement $A_\epsilon^{(n)c}$ and we order all elements in each set according to lexicographic order.

Then we can represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set.

Giving the index of the sequence in the set, we can represent each sequence of $A_\epsilon^{(n)}$. Since there are $\leq 2^{n(H-\epsilon)}$ sequences in $A_\epsilon^{(n)}$, the indexing requires no more than $n(H+\epsilon)+1$ bits because $n(H+\epsilon)$ may not be an integer.

We prefix all their sequences by a 0, giving a total length of $\leq n(H+\epsilon)+2$ bits to represent each sequence in $A_\epsilon^{(n)}$.

We denote æ^n as a sequence $X_1, X_2, \cdots, X_n$. Let $I(x^n)$ be the length of the code word corresponding to x^n.

Lemma 3.2. *Let X^n be independent identically distribution (i. i.d.) with probability $p(x)$. Let $\epsilon > 0$. Then there exists a code which maps sequences x^n of length n into binary strings such that the mapping is one to one and $E[\frac{1}{n}l(X^n)] < H(X) + \epsilon$, for n sufficiently large.*

Proof. $E(l(X^n)) = \sum_{x^n} P(x^n)l(x^n)$
$= \sum_{x^n \in A_\epsilon^{(n)}} P(x^n)l(x^n) + \sum_{x^n \in A_\epsilon^{(n)c}} P(x^n)l(x^n)$
$\leq \sum_{x^n \in A_\epsilon^{(n)}} P(x^n)[n(H+\epsilon)+2] + \sum_{x^n \in A_\epsilon^{(n)}} P(x^n)(n\log|\,\text{æ}| + 2)$
$= Pr\{A_\epsilon^{(n)}\}\{n(H+\epsilon)+2\} + Pr\{A_\epsilon^{(n)c}\}(n\log|\,\text{æ}|) + 2$
$\leq n(H+\epsilon) + \epsilon_n(\,\log\,|\,\text{æ}\,|) + 2 = n(H+\epsilon^1)$
where $\epsilon^1 = \epsilon + \epsilon(\,\log\,|\,\text{æ}|) + \frac{2}{n}, \{B_\delta^{(n)}\} \geq 1-\delta.$

Definition 3.3. For each $n = 1, 2, \cdots$, let $B_\delta^{(n)} \subset \text{æ}^n$ be any set with $Pr\{B_\delta^{(n)}\} \geq 1-\delta$ must have significant intersection with $A_\epsilon^{(n)}$ and therefore must have about as many elements.

Theorem 3.4. *Let $X_1, X_2, \cdots, X_n$ be i.i.d. with $p(x)$. For $\delta < \frac{1}{2}$ and any $\delta^1 > 0$, if Pr $\{B_\delta^{(n)}\} > 1-\delta$, then $\frac{1}{n}$ log $|B_\delta^{(n)}| > H - \delta^1$ for n sufficiently large. Thus $B_\delta^{(n)}$ must have at least 2^{nH} elements, to first order in the exponent. But $A_\epsilon^{(n)}$ has $2^{n(H\pm\epsilon)}$ elements.*

Proof. Let any two sets A, B such that $Pr(A) > 1-\delta_1$ and $Pr(B) > 1-\epsilon_2$. Then this shows that $Pr(A \cap B) > 1-\epsilon_1-\epsilon_2$, hence $Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) > 1-\epsilon-\delta$. Indeed, since $X_1, X_2, \cdots, X_n$ are i.i.d. with $p(x)$, if we fix $\epsilon < \frac{1}{2}$, then

$$Pr(A \cap B) = Pr(A) \cdot Pr(B) > (1-\epsilon_1)(1-\epsilon_2) = 1-\epsilon_1-\epsilon_2.$$

Accordingly, $Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) = Pr(A_\epsilon) \cdot Pr(B_\delta) \geq (1-\epsilon)(1-\delta) = 1-\epsilon-\delta$ by Proposition 3.1.(1).

Next by the chain rule of inequalities,

$$\begin{aligned} 1-\epsilon-\delta &< Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \\ &= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} P(x^n) \leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \\ &= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} \leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}, \end{aligned}$$

$$|B_\delta^{(n)}| \geq (1-\epsilon-\delta)2^{n(H-\epsilon)}.$$

Taking the logarithm with base 2 to both sides,

$$\log{}_2|B_\delta^{(n)}| \geq \log\,(1-\epsilon-\delta) + n(H+\epsilon),$$

$$\frac{1}{n}\log\,|B_\delta^{(n)}| > \frac{1}{n}\log\,(1-\epsilon-\delta) + (H-\epsilon).$$

Accordingly for n sufficiently large, we obtain

$$\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta^1.$$

We denote the notation $a_n \doteq b_n$ as follows.

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} = 0.$$

Then we can now restate the theorem 3.4. as

$$|B_\delta^{(n)}| \doteq |A_\epsilon^{(n)}| \doteq 2^{nH}.$$

4. Monotonicity of entropy

Let the joint distribution of any subset of the sequence of random variables to be invariant with respect to shifts in the time index, i.e.
$Pr\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$
$= Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \cdots, X_{n+l} = x_n\}$
for every shift l and for all $x_1, x_2, \cdots \in X$. Then a stochastic process is called to be stationary.

Let random variables $X_1, X_2, \cdots$ be a discrete stochastic process. If for $n = 1, 2, \cdots$
$Pr(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_1 = x_1)$
$= Pr(X_{n+1} = x_{n+1}|X_n = x_n)$for all $x_1, x_2, \cdots, x_n, x_{n+1} \in$ æ.

Then a discrete stochastic process $X_1, X_2, \cdots$ is said to be a Markov chain or a Markov process. ([6],[7])

Definition 4.1. The Markov chain is said to be time invariant if the conditional probability $P(X_{n+1}|X_n)$ does not depend on n, i.e. for $n = 1, 2, \cdots$

$$Pr\{X_{n+1} = p|X_n = q\} = Pr\{X_2 = p|X_1 = q\}, \text{ for all } p, q \in \text{æ}.$$

Let $\{X_i\}$ be a Markov chain. Then X_n is said the state at time n. The Markov chain is called to be irreducible if it is possible to go with positive probability from any state of the Markov chain to any other state in a finite numer of steps.

Definition4.2. The entropy rate of a stochastic process $\{X_i\}$ is defined as follow.

$$H(X) = \lim_{n\to\infty}\frac{1}{n}H(x_1, x_2, \cdots, x_n), \text{ when the limit exists.}$$

Another definition of entropy rate is, when the limit exists,
$H'(X) = \lim_{n\to\infty} H(x_n|x_{n-1}, x_{n-2}, \cdots, x_2, x_1)$.

Proposition 4.3. *For a stationary stochastic process, their entropy rate $H(æ)$ and $H'(æ)$ are equal, i.e.*

$$H(æ) = H'(æ).$$

Proof. Since $H(X_{n+1}|X_1, X_2, \cdots, X_n) \leq H(X_{n+1}|X_n, \cdots, X_2)$ $\leq H(X_n|X_{n-1}, \cdots, X_1)$, $H(X_n|X_{n-1}, \cdots, X_1)$ is a decresing sequence of nonnegative numbers.

Hence it has a limit, $H'(æ)$.([6]) $\cdots\cdots\cdots (*_1)$ In general, we can prove that if $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=0}^{\infty} a_i$, then $b_n \to a$.([6]) $\cdots\cdots\cdots (*_2)$

Since by the chain rule the entropy rate is the time average of the conditional entropies,

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

Also the conditional entropies tend to a limit $H'(æ)$. By $(*_2)$, their running average has a limit equal to the limit $H'(æ)$ of the terms. By $(*_1)$,

$$H(æ) = \lim \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim H(X_n|X_{n-1}, \cdots, X_1) = H'(æ).$$

For a stationary Markov chain, the entropy rate is

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

Let μ be stationary distribution and P be transition matrix. Then since

$$H(æ) = H(X_2|X_1) = \sum\nolimits_i \mu_i (\sum\nolimits_j - P_{ij} \log P_{ij}),$$

$$H(æ) = -\sum\nolimits_{ij} \mu_{ij} P_{ij} \log P_{ij}.$$

For example, consider a two-state Markov chain with probability transition matrix

$$P = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}$$

Put μ_1 and μ_2 be the stationary probability of state a and b respectively. Then we obtain $\mu_1 p = \mu_2 \cdot 1$, since $\mu_1 + \mu_2 = 1$, the stationary distribution is $\mu_1 = \dfrac{1}{1+p}, \mu_2 = \dfrac{p}{1+p}$

Accordingly the entropy of the state X_n at time n is

$$H(X_n) = H(\frac{1}{1+p}, \frac{p}{1+p}).$$

The entropy rate is $H(æ) = H(x_2|x_1) = \dfrac{1}{1+p}H(p) + \dfrac{p}{1+p}H(1)$.

Theorem 4.4. *(i) Let* $\{X_i\}_{i=-\infty}^{\infty}$ *be a stationary stochastic process, then*

$$H(x_0|x_{-1}, x_{-2}, \cdots, x_{-n}) = H(x_0|x_1, x_2, \cdots, x_n).$$

(ii) Let $X_1, X_2, \cdots, X_n$ *be a stationary stochastic process. Then*

$$\frac{H(X_1, X_2, \cdots, X_n)}{n} \leq \frac{H(X_1, X_2, \cdots, X_{n-1})}{n-1}.$$

Proof. (i) For a stationary stochastic process, the probability of any sequence of state is the same forward or backward. i.e. time-reversible.

$Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$
$= Pr(X_n = x_1, X_{n-1} = x_2, \cdots, X_1 = x_n)$.
$H(X_0|X_{-1}, X_{-2}, \cdots, X_{-n})$
$= H(X_0|X_{-n}, X_{-n+1}, \cdots, X_{-n+(n-2)}, X_{-n+(n-1)})$.
Replacing $n = -1$,
$H(X_0|X_{-1}, X_{-2}, \cdots, X_{-n}) = H(X_0|X_1, X_2, \cdots, X_{n-1}, X_n)$.

This means that the present has a conditional entropy given the past equal to the conditional entropy given the future.

(ii) $0 \leq p(x) \leq 1$ implies $H(x) = \sum p(x) \log \frac{1}{p(x)} \geq 0$. So $H(X_1, X_2, \cdots, X_n) \geq 0$. By the chain rule, $H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$.

By the proposition 4.3, the conditional probability has a limit. Since the running average $\dfrac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$ has a limit equal to the limit $H(X)$ of the terms.

$\dfrac{H(X_1, \cdots, X_n)}{n} = \dfrac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$
$= H(X_n|X_{n-1}, \cdots, X_1)$ as $n \to \infty$.
Therefore

$$\begin{aligned}\frac{H(X_1, \cdots, X_n)}{n} &= H(X_n|X_{n-1}, X_{n-2}, \cdots, X_2, X_1)\\ &\leq H(X_n|X_{n-1}, X_{n-2}, \cdots, X_2)\\ &= H(X_{n-1}|X_{n-2}, \cdots, X_1)\\ &= \frac{H(x_1, x_2, \cdots, x_{n-1})}{n-1} \quad \text{for large } n \text{ enough.}\end{aligned}$$

We wish to compute the entropy of a random tree. From this we can find that the expected number of necessary to describe the random tree T_n grows linearly with n.

The following method of generating random trees yields the same probability distribution on trees with n terminal nodes. Choose an integer N_1 uniformly distributed on $\{1, 2, \cdots, n-1\}$. Then we have the picture.

Choose an integer N_2uniformly distributed over $\{1, 2, \cdots, N_1 - 1\}$ and independently choose an other integer N_3 uniformly over $\{1, 2, \cdots, (n - N_1) - 1\}$.

We continue the process until no further subdivision can be made. Then we can make n-terminal nodes tree.

Let T_n denote a random n-node tree generated as above, then the entropy $H(T_2) = 0, H(T_3) = \log 2$. For $n = 4$, we have five possible trees, with probbilities $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$.

Let $N_1(T_n)$ denote the number of terminal nodes of T_n in the right half of the tree.

Theorem 4.5.

$$(n-1)H_n = nH(n-1) + (n-1)log(n-1) - (n-2)log(n-2)$$

or $\dfrac{H_n}{n} = \dfrac{H_{n-1}}{n-1} + C_n$ *for appropriately defined* C_n.

Proof. By the definition of entropy and the construction of random tree,

$$\begin{aligned}
H(T_n) &= H(N_1, T_n) = H(N_1) + H(T_n|N_1) \cdots\cdots < \text{by chain rule for entropy} > \\
&= \log\ (n-1) + H(T_n|N_1) \cdots\ < \text{by conditional disribution} > \\
&= \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}[H(T_k) + H(T_{n-k})] \cdots\cdots < \text{by definition of tree} > \\
&=\ \log\ (n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k) \cdots\ < \text{by restriction} > \\
&=\ \log\ (n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k.
\end{aligned}$$

Let $H(T_n)$ be H_n. Then $H(T_{n-1}) = H_{n-1}, H(T_{n-2}) = H_{n-2}, \cdots,$
$H_{n-1} =\ \log\ (n-2) + \dfrac{2}{n-2}\sum_{k=1}^{n-2} H(T_n)$. Accordingly,

$$\begin{aligned}
(n-1)H_n &= (n-1)\log(n-1) + 2\sum_{k=1}^{n-2} H_k + 2H_{n-1} \\
&= (n-1)\log(n-1) + (n-2)\{\log(n-2) \\
&\quad + \frac{2}{n-2}\sum_{k=1}^{n-2} H_k\} - (n-2)\log(n-2) + 2H_{n-1} \\
&= (n-1)\ \log\ (n-1) + nH_{n-1} - (n-2)\ \log\ (n-2) + 2H_{n-1} \\
&= (n-1)\ \log\ (n-1) + nH_{n-1} - (n-2)\ \log\ (n-2).
\end{aligned}$$

By dividing both sides as $n(n-1)$,

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + C_n$$

where $C_n = \frac{1}{n}\ \log\ (n-1) - (1 - \frac{2}{n})(\frac{1}{n-1})\ \log\ (n-2)$. Since $\sum C_n = C < \infty$, you have proved that $\frac{1}{n}H(T_n)$ converges to a constant.

References

1. R.L.Alder, D.Coopersmith, M.Hassner, *Algorithms for slidingBlock Codes- an application of symbolic dynamics to information theory*, IEEE Trans, Inform, theory, IT 5-22,, 1983.
2. P.Algoet, T.M.Cover, *Asymptotic optimality and asymptotic equipartition property of log-optimal investment*, Annals of Probability, 16 : 876-898, 1988.
3. Thomas M.Cover, Joy A. Thomas, *Elements of information theory*, A Wiley-Inter Science Publication, 1991.
4. R.M.Fano, *Class notes for transmission of information*, Course 6.574 MIT Cambridge, MA, 1952.
5. E.T.Janes, *Papers on probability*, Statistics and Statistical Physics. Reidel, Dordrecht, 1982.
6. Douglas Lind, Brian Marcus, *An introduction to symbolic dynamics and coding*, 1995.
7. B.Macmillan, *The basic theorems of information theory*, Ann.Math.Stat.24 : 196-219, 1953.
8. D.S.Ornstein, *Bernoulli shifts with the same entropy are isomorphic*, Advances in Math, 4 : 337-352, 1970.
9. Steven Roman, *Introduction to coding and information theory*, Springer-Verlag, 1997.
10. C.E.Shannon, *A mathematical theory of model*, Ann.Stat.6 : 461-464, 1978.

Department of mathematics
Woosuk University
Wanju-gun Chonbuk 565-701, Korea