J. KSIAM Vol.4, No. 1, 133-142, 2000

SOME PROPERTIES OF MUTUAL INFORMATION AND TYPICAL SET

YOUNG SOO LEE

Abstract

In this note we define typical set and differential entropy for continuous random variables. Using Markov chain, we show that the various properties of the mutual information and entropies (theorems 3.2 and 3.4) and show the properties of typical set in continuous random variables(lemma 4.2 and theorem 4.3.)

1. Introduction.

The subjects of information theory and coding theory began in 1948 with a famous paper by Claude Shannon, of Bell Labs, entitled A Mathematical Theory of Communication.

Indeed, elementary information theory is a beautiful application of discrete probability theory to the problem of encoding for efficiency and elementary coding theory is a beautiful application of alphabet and combinatorics to the problem of error detection and correction. The study of information theory consists of three parts :

one is related to cipher theory and signal transition and the rest is concerned with communication theory. ([1], [4], [10])

The concepts of information theory is too wide to define to constant one way. But there

are a few key ideas and techniques that when mastered, make the subject appear simple and provide great inituition on new questions. Quantities like entropy and mutual information arise as the answers to fundamental questions. The concept of entropy in information theory is closely connected with the concept of entropy in statistical mechanics.

The purpose of this paper is to find some properties of information theory with various entropies.

Many properties of entropies and typical set is found in discrete random variables, but is not proved in continuous random variables. We wish to do this.

In section 2, we describe some important terminologies and notations and basic definitions which are needed to prove theorem3.2, theorem 3.4 and lemm4.2, theorem4.3 in sections 3 and 4.

In section 3, we discuss some kinds of information inequality and thermodynamics and prove theorem3.2 and theorem3.4 using the propositions 3.1

Key words and phrases : Typical set, Differential entropy

This research was supported by Woosuk University Research Grant.

and 3.3.

In section 4, we construct the typical sequences and typical set and discuss the differential entropy for continuous random variables. Here we prove lemma4.2 and theorem4.3 using proposition4.1.

2. Terminologies and Basic definitions.

Let X be a discrete random variable with alphabet æ and probability mass function $p(x)$. Then the entropy $H(X)$ of a discrete random variable X is defined by $H(X)=-\sum_{x\in X}p(x)\log p(x)$.([10])

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

The differential entropy $h(X)$ of a continuous random variable X with a probability density function $f(x)$ is defined as $h(X)=-\int_S f(x)\log f(x)dx$, where S is the support set of the random variable.that is, $S=\{x:f(x)>0\}$.

We often denote $h(f)$ rather than $h(X)$

If X, Y have a joint probability density function $f(x,y)$, we can define the conditional differential entropy $h(X\mid Y)$ as $h(X\mid Y)=-\int f(x,y)\log f(x\mid y)dxdy$

Since in general $f(x\mid y)=f(x,y)\mid f(y)$,

$$h(X\mid Y)=h(X,Y)-h(Y)$$

The relative entropy (or Kullback Leibler distance) $D(f\|g)$ between two densities f and g is defined by $D(f\|g)=\int f\log\frac{f}{g}$

The mutual information $I(X:Y)$ between two random variables X and Y with joint probability density function $f(x,y)$ is defined as $I(X:Y)=\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$.

The properties of $D(f\|g)$ and $I(X:Y)$ are the same as in the discrete case.

The following properties are well-known([2],[6],[10])

(i) $H(X,Y)=H(X)+H(Y\mid X)$ (chain rule)

(ii) $I(X:Y)=H(X)-H(X\mid Y)$ (mutual information and entropies)

(iii) $H(X_1,X_2,\cdots,X_n)=\sum_{i=1}^{n}H(X_i\mid X_{i-1},\cdots,X_1)$ (Chain rule for entropy)

(iv) $I(X_1,X_2,\cdots,X_n,Y)=\sum_{i=1}^{n}I(X_i:Y\mid X_{i-1},X_{i-2},\cdots,X_1)$

(Chain rule for information)

A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2\in(a,b)$ and $0\le x\le 1$,

$$f(\lambda x_1+(1-\lambda)x_2)\le\lambda f(x_1)+(1-\lambda)f(x_2)$$

If $-f(x)$ is convex, then $f(x)$ is said to be concave.

A function is convex if it always lies below any chord and is concave if it alway lies above any chord.

If the function f has a second derivative $f''(x)$ which is non-negative everywhere, then the function is convex.

proposition 2.1 Let f be a convex function and X be a random variable, then

$$Ef(X) \geq f(EX)$$

Moreover, if f is strictly convex, then equality implies that $X = EX$ with probability 1. i.e., X is a constant. This is called Jensen's inequality.([10],[11]).

3. Mutual information inequality.

Let $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_n$ be non-negative numbers. Then by the concavity of logarithm

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \Big(\sum_{i=1}^{n} a_i\Big) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff $\frac{a_i}{b_i}$ = constant.

By convention we use $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$.

Indeed,

Let $a_i > 0$ and $b_i > 0$. Take the function $f(t) = t \log t$. Then by Jensen's inequality, we have

$$\sum a_i f(t_i) \geq f(\sum a_i t_i)$$

for $a_i \geq 0$, $\sum_i a_i = 1$. Since f(t) is strictly convex, setting $a_i = b_i / \sum_{j=1}^{n} b_j$ and $t_i = a_i / b_i$, we obtain

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_i}$$

which is the log sum inequality ‡

Proposition 3.1. we can show the following two inequality.

1) Let $p(x)$, $q(x)$, $x \in X$ be two probability mass functions. Then $D(p \| q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x. (Information inequality)

2) For any two random variables x, y, $I(x;y) \geq 0$ with equality if and only if x and y are independent. (Non-negative of mutual information).

proof. 1) Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$-D(p \| q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)}$$
$$= \log \sum_{x \in A} q(x) = \log 1 = 0,$$

Hence we have $D(p \| q) = 0$ if and only if $p(x) = q(x)$ for all x.

we also can show this by the log sum inequality.

$$D(p \| q) = \sum p(x) \log \frac{p(x)}{q(x)} \geq \Big(\sum p(x)\Big) \log \frac{\big(\sum p(x)\big)}{\big(\sum q(x)\big)} = 1 \log \frac{1}{1} = 0.$$

with equality if and only if $p(x)/q(x) = c$ ‡

(2) Similarly, $I(x,y) = D(p(x,y) \| p(x)p(y)) \geq 0$, with equality if and only if $p(x,y) = p(x)p(y)$,

i.e., x and y are independent.

Let $x \rightarrow y \rightarrow z$ be Markov chain. Then $I(x,y) \geq I(x,z)$.

Indeed, by the chain rule

$$I(x,y,z)=I(x,y)+I(x,y\mid z)=I(x;y)+I(x,z\mid y)$$

Since x and z are conditionally independent for given y, we have $I(x,z\mid y)=0$. We have $I(x,y)\geq I(x,z)$ because of $I(x;y\mid z)\geq 0$.

We equality if and only if I(x:y | z) = 0, i.e., $x\rightarrow z\rightarrow y$ forms a Markov chain.

Theorem 3.2. Let $x\rightarrow y\rightarrow z$ be Markov chain. Then

I) $I(y;z)\geq I(x;z)$. 2) $I(x;y\mid z)\leq I(x,y)$.

proof. 1) we can expand mutual information in two different ways.

By the chain rule, Since if $x\rightarrow y\rightarrow z$, *then* $z\rightarrow y\rightarrow x$

$$I(z;y,x)=I(z;y)+I(z;x\mid y)$$
$$=I(z;x)+I(z;y\mid x)$$

Since x and z are conditionally independent given y, we have $I(z;x\mid y)=0$.

Since $I(z;x\mid y)\geq 0$, we have $I(z;y)\geq I(z;x)$.

Therefore $I(y;z)\geq I(x;z)$ ‡

2) We can prove similar to 1).

By the chain rule,

$I(x;y,z)=I(x;z)+I(x;y\mid z)=I(x;y)+I(x;z\mid y)$.

Since $I(x;z\mid y)=0$ by Markovity, $I(x;z)\geq 0$.

Hence we have

$$I(x;y\mid z)\leq I(x;y).$$

The second law of thermodynamics states that the relative entropy always decreases and the entropy of an isolated system is non-decreasing. ([4], [8], [10])

Let u_n and u'_n be true probability distributions on the state of a Markov chain at time n and let u_{n+1} and u'_{n+1} be the corresponding distributions at time n+1. Let the corresponding mass functions be denoted by p and q.

Thus $p(x_n,x_{n+1})=p(x_n)r(x_{n+1}\mid x_n)$ and $q(x_n,x_{n+1})=q(x_n)r(x_{n-1}\mid x_n)$, where $r(\cdot\mid\cdot)$ is the probability transition function for the Markov chain. Then by the chain rule for relative entropy, we have two expansions:

$$D(p(x_n,x_{n+1})\parallel q(x_n,x_{n-1}))$$
$$=D(p(x_n)\parallel q(x_n))+D(p(x_{n+1}\mid x_n)\parallel q(x_{n+1}\mid x_n))$$
$$=D(p(x_{n+1})\parallel q(x_{n+1}))+D(p(x_n\mid x_{n+1})\parallel q(x_n\mid x_{n+1})).$$

Since both p and q are derived from the Markov chain, the conditional probability mass functions $p(x_{n+1}\mid x_n)$ and $q(x_{n+1}\mid x_n)$ are equal to $r(x_{n+1}\mid x_n)$ and hence $D(p(x_{n-1}\mid x_n)\parallel q(x_n\mid x_{n+1}))=0$,

Now using the non-negativity of $D(p(x_n\mid x_{n+1})\parallel q(x_{n+1}))$, Hence we have $D(p(x_n)\parallel q(x_n)\geq D(p(x_{n+1})\parallel q(x_{n+1}))$

or

$$D(\mu_n\parallel\mu'_n)\geq D(\mu_{n+1}\parallel\mu'_{n+1})$$

Here μ'_n is any distribution on the states at time n. if we let μ'_n be any stationary distribution μ, then μ'_{n-1} is the same stationary distribution . Hence

$$D(\mu_n\parallel\mu)\geq D(\mu_{n+1}\parallel\mu),$$

$D(\mu_n\parallel\mu)$ between a distribution μ_n on the states at time n and a stationary distribution μ decreases with n. If the stationary distribution is the uniform distribution, then we can express the relative entropy as

$$D(\mu_n \| \mu) = \log \ | X | \ - H(x_n).$$

In this case the monotonic decrease in the relative entropy implies a monotonic increase
in entropy. This is the explanation that ties in most closely with statistical thermodynamics ‡
Let $| X |$ be the number of elements in the range of x. Then $H(x) \leq \log | X |$, with equality if and only if x has a uniform distribution over X.
Let u(x) be the uniform probability mass function over x. Then

$$D(p \| \mu) = \sum p(x) \log \frac{p(x)}{\mu(x)} = \log | X | - H(x).$$

Hence by the non-negativity of relative entropy,

$$0 \leq D(p \| \mu) = \log | X | - H(x) \ddagger$$

Proposition 3.3. The conditional entropy $H(x_n | x_1)$ increases with n for a stationary Markov process .
proof. If the Markov process is stationary, then $H(x_n)$ is constant.
So the entropy is non-increasing. we shall prove to two different ways.
First, By the properties of entropy, we have

$$\begin{aligned} H(x_n | x_1) &\geq H(x_n | x_1, x_2) && \text{(conditioning reduces entropy)} \\ &= H(x_n | x_2) && \text{(by Markovity)} \\ &= H(x_{n-1} | x_1) && \text{(by stationarity)} \end{aligned}$$

Thus $H(x_n | x_1)$ increases with n.
Alternatively, by an application of the data processing inequality to the markov chain
$x_1 \to x_{n-1} \to x_n$, we have

$$I(x_1; x_{n-1}) \geq I(x_1; x_n).$$

By the mutual informations in terms of entropies ,

$$H(x_{n-1}) - H(x_{n-1} | x_1) \geq H(x_n) - H(x_n | x_1)$$

By stationarity,

$$H(x_{n-1}) = H(x_n).$$

Hence we have

$$H(x_{n-1} | x_1) \leq H(x_n | x_1). \ddagger$$

Theorem 3.4. Let $x_0 \to x_{n-1} \to x_n$ be Markov chain with n,
$H(x_0 | x_1)$ is non-decreasing.
proof. If $x_0 \to x_{n-1} \to x_n$ is Markov chain, then $x_n \to x_{n-1} \to x_0$ is a markov chain. That is,
$x_0 \leftrightarrow x_{n-1} \leftrightarrow x_n$ is a Markov chain. Then

$$I(x_0 | x_{n-1}) \geq I(x_0, x_n) \quad \text{by markovity.}$$
$$H(x_0) - H(x_0 | x_{n-1}) \quad \text{by mutual information.}$$

So $H(x_0 | x_n) \geq H(x_0 | x_{n-1})$.

Alternativly , the sequence $x_0 \leftrightarrow x_{n-1} \leftrightarrow x_n$ is markov chain.

$H(x_0 | x_n) \geq H(x_0 | x_{n-1}, x_n)$ (conditioning reduces entropy).
By markov chain,
Hence $H(x_0 | x_{n-1}, x_n) = H(x_0 | x_{n-1})$.
$H(x_0 | x_n) = H(x_0 | x_{n-1})$.

This is an entropy of initial conditions ‡

4. Typical sets in continuous random variables.

The asymptotic equipartition property(AEP) shows which most sequence are typical in that they have a sampled entropy close to H.

Let $X_1, X_2, \cdots, X_n$ be independent identically distributed (i.i.d) random variables and $p(X_1, X_2, \cdots, X_n)$ be the probability of 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.

This set has approximately 2^{nH} typical sequence of length n and the probability of each set is approximately 2^{nH}.

We can therefore represent the typical sequences of length n using approximately $nH(æ)$

This enables us to divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the non-typical set, which contains the other sequences.

We can show that the typical set has a probability close to 1.

Let X be a random variable with cumulative distribution function $F(x)=P_r\ (X\leq r)$.

Let $f(x)=F'(x)$ when the derirative is defined.

The set where $f(x)>0$ is called the support set of X.

Proposition 4.1 Let $X_1, X_2, \cdots, X_n$ be a sequence of random variable drawn i.i.d. according to the density $f(x)$. Then

$$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n)\rightarrow h(X) \text{ in probability.}$$

Proof. Functions of independent random variables are also independent random variables. Thus,since the x_i are i.i.d, so are $\log f(X_i)$.

By the weak law of lange numbers,

$$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n)=-\frac{1}{n}\sum_i \log f(X_i)=h(X)$$

For $\varepsilon>0$ and any n, the set $A_\varepsilon^{(n)}$ with respect to f(x) is called the typical set as follows.

$$A_\varepsilon^{(n)}=\left\{(x_1, x_2, \cdots, x_n)\in S^n:\left|-\frac{1}{n}\log f(x_1, x_2, \cdots, x_n)-h(x)\right|\leq\varepsilon\right\}$$

where $f(x_1, x_2, \cdots, x_n)=\pi_{i=1}^n f(x_i)$

We define the volume of a set $A\in R^n$,denoted by as follows.

$$Vol(A)=\int_A dx_1 dx_2\cdots dx_n$$

We must remark that the analog of the cardinality of the typical set for the discrete random variable s is the volume of the typical set in the continuous random variables. Then by the definition of the typical set $A_\varepsilon^{(n)}$, We can show that if

$(x_1, x_2, \cdots, x_n) \in A_\varepsilon^{(n)}$, then

$$h(X) - \varepsilon \leq -\frac{1}{n} \log f(x_1, x_2, \cdots, x_n) \leq h(X) + \varepsilon$$

, where $h(X)$ denotes the differential entropy .
As a consequence of the asymptotic equipartition property, we can show that the set $A_\varepsilon^{(n)}$ has the following properties:

Lemma 4.2
1. $P_r(A_\varepsilon^{(n)}) > 1-\varepsilon$ for n sufficiently large
2. $Vol(A_\varepsilon^{(n)}) \leq 2^{n(h(X)+\varepsilon)}$ for all n
3. $Vol(A_\varepsilon^{(n)}) \leq (1-\varepsilon)2^{n(h(X)-\varepsilon)}$ for n sufficiently large

Proof 1. By the definition of the AEP and typical set $A_\varepsilon^{(n)}$, taking the log with base 2 to both sides, $-n(h(x)+\varepsilon) \leq \log_2 f(x_1, x_2, \cdots, x_n) \leq -n(h(x)-\varepsilon)$

Therefore $h(x) - \varepsilon \leq -\frac{1}{n} \log_2 f(x_1, x_2, \cdots, x_n) \leq h(x) + \varepsilon$

since the probability of the event $(x_1, x_2, \cdots, x_n) \in A_\varepsilon^{(n)}$ tends to 1 as $n \to \infty$.
For any $\delta > 0$, there exist an n_o, such that for all $n \geq n_o$,

$$P_r\left\{\left| -\frac{1}{n} \log f(x_1, x_2, \cdots, x_n) - h(x) \right| < \varepsilon\right\} > 1-\delta$$

Proof 2. $1 = \int_{s^n} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \geq \int_{A_\varepsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n$

$$\geq \int_{A_\varepsilon^{(n)}} 2^{-n(h(x)+\varepsilon)} dx_1 dx_2 \cdots dx_n = 2^{-n(h(x)+\varepsilon)} \int_{A_\varepsilon^{(n)}} dx_1 dx_2 \cdots dx_n = 2^{-n(h(x)+\varepsilon)} Vol(A_\varepsilon^{(n)})$$

hence $Vol(A_\varepsilon^{(n))} \leq 2^{n(h(x)+\varepsilon)}$ for all n.

Proof 3. Finally, for sufficiently lange n, $P_r\{A_\varepsilon^{(n)}\} > 1-\varepsilon$.
So

$$1-\varepsilon \leq \int_{A_\varepsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \leq \int_{A_\varepsilon^{(n)}} 2^{-n(h(x)-\varepsilon)} dx_1 dx_2 \cdots dx_n$$

$$= 2^{-n(h(x)-\varepsilon)} \int_{A_\varepsilon^{(n)}} dx_1 dx_2 \cdots dx_n = 2^{-n(h(x)-\varepsilon)} Vol(A_\varepsilon^{(n)})$$

is establishing property 3.
We argue further that the volume of the typical set is at least this lange
Hence we have

$$(1-\varepsilon)2^{n(h(x)-\varepsilon)} \leq Vol(A_\varepsilon^{(n)} \leq 2^{n(h(x)+\varepsilon)}$$

which completes the proof of the properties of $A_\varepsilon^{(n)}$
We divide all sequence in $æ^n$ into two sets:
One is the typical set $A_\varepsilon^{(n)}$ and the other is complement $(A_\varepsilon^{(n)})^c$. We order all elements in each set according to some order.
Then we can reprensent each sequence of $A_\varepsilon^{(n)}$ by giving the index of the sequence in the set.
Since there exists $\leq 2^{n(h+\varepsilon)}$ sequence in $A_\varepsilon^{(n)}$ the indexing sequence n_0 more than $n(h+\varepsilon)' + 1$ bits.
we prefix all these sequence by a 0, giving a total length of $\leq n(h+\varepsilon)+2$ fits

to represent each sequence in $A_\varepsilon^{(n)}$

Similarly, we can index each sequence not in $A_\varepsilon^{(n)}$ by using not more than $n\log|\text{æ}|-1$ bits.

prefixing these indices by 1. we have a code for all the sequence in æ^n.U

Note that the typical sequence have short description of length $=nh$.

Indeed, let the notation x^n to denote a sequence $x_1, x_2, \cdots, x_n$ and $l(x^n)$ be the length of the code word corresponding to x^n.

If n is sufficiently large so that $f\{A_\varepsilon^{(n)}\}\geq 1-\varepsilon$, then the expected length of the codeword is

$$\begin{aligned}E(l(x^n)) &= \int_{x^n\in S} f(x^n)l(x^n) = \int_{x^n\in A_\varepsilon^{(n)}} f(x^n)l(x^n) + \int_{x^n\in A_\varepsilon^{(n)c}} f(x^n)l(x^n)\\ &\leq \int_{x^n\in A_\varepsilon^{(n)}} f(x^n)[n(h+\varepsilon)+2] + \int_{x^n\in A_\varepsilon^{(n)c}} f(x^n)(n\log|\text{æ}|+2)\\ &= f\{A_\varepsilon^{(n)}\}[n(h+\varepsilon)+2] + f\{A_\varepsilon^{(n)c}\}(n\log|\text{æ}|+2) \leq n(h+\varepsilon)+\varepsilon n(\log|\text{æ}|)+2\\ &= n(h+\varepsilon^1)\end{aligned}$$

where $\varepsilon^1=\varepsilon+\varepsilon\log|\text{æ}|+\frac{2}{n}$

From the definition of $A_\varepsilon^{(n)}$, the typical set $A_\varepsilon^{(n)}$ is a small set that contains most of probability.

But it is not clear whether it is the smallest such set. Now we shall show that $A_\varepsilon^{(n)}$ has essentially the same number of the elements as the smallest set, to first order in the exponent.

we define $B_\delta^{(n)}$ as follows:

Let $B_\delta^{(n)}\subset \text{æ}^n$ be any set with $f\{B_\delta^{(n)}\}\geq 1-\delta$ for each $n=1, 2, \cdots$

We argue that $B_\delta^{(n)}$ must have significant intersetion with $A_\varepsilon^{(n)}$ and therefore must have about as many elements.

Theorem 4.3 The typical set $A_\varepsilon^{(n)}$ is the smallest volume set with probability $f\{B_\delta^{(n)}\}\geq 1-\varepsilon$ to first order in the exponent. This theorem shows that the volume of the smallest set that contains most of the probability is approximately 2^{nh}.

This is an n-dimensional volume, so the corresponding side length is $(2^{nh})^{\frac{1}{n}}=2^h$

Proof. Let any two sets A, B as $f(A)>1-\varepsilon_1$ and $f(B)>1-\varepsilon_2$. Since $X_1, X_2, \cdots, X_n$ are i.i.d with probability density function $f(x)$, if we fix $\varepsilon<\frac{1}{2}$,

$f(A\cap B)=f(A)\cdot f(B) > (1-\varepsilon_1)(1-\varepsilon_2) = 1-\varepsilon_1-\varepsilon_2$.

Accordingly, $f(A_\varepsilon^{(n)}\cap B_\delta^{(n)})=f(A_\varepsilon^{(n)})\cdot f(B_\delta^{(n)})\geq(1-\varepsilon)(1-\delta)=1-\varepsilon-\delta$ by proposition 4.1

Next by the chain rule of inequality,

$$1-\varepsilon-\delta < f(A_\varepsilon^{(n)}\cap B_\delta^{(n)}) = \int_{A_\varepsilon^{(n)}\cap B_\delta^{(n)}} f(x^n) \leq \int_{A_\varepsilon^{(n)}\cap B_\delta^{(n)}} 2^{-n(h-\varepsilon)}$$

$$Vol(A_\varepsilon^{(n)}\cap B_\delta^{(n)})2^{-n(h-\varepsilon)} \leq Vol\{B_\delta^{(n)}\}2^{-n(h-\varepsilon)}$$

$$Vol\{B_\delta^{(n)}\}\geq(1-\varepsilon-\delta)2^{n(h-\varepsilon)}.$$

Taking the logarithm with have 2 to both sides,

$$\log_2 Vol\{B_\delta^{(n)}\} \geq \log_2(1-\varepsilon-\delta)+n(h+\varepsilon)$$

$$\frac{1}{n}\log_2 Vol(B_\delta^{(n)}) > \frac{1}{n}\log_2(1-\varepsilon-\delta)+(h-\varepsilon)$$

For n sufficiently large, we obtain

$$\frac{1}{n}\log Vol(B_\delta^{(n)}) > h-\delta^1$$

Let us define the notation $a \doteq b$ as follows.

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

Then we obtain

$$Vol(B_\delta^{(n)}) \doteq Vol(A_\varepsilon^{(n)}) \doteq 2^{nh}$$

REFERENCES

[1] R. L. Adler, D. Coppersmith, and M. Hassner. Algorithms for sliding block codes-an application of symbolic dynamics to information theory. *IEEE Trans. Inform. Theory*, IT-29: 5-22, 1983.

[2] P. Algoet and T. M. Cover. Asymptotic optimality and asymptotic equipartition property of log-optimal investment. *Annals of Probability*, 16: 876-898, 1988.

[3] R. M. Fano. Class notes for Transmission of Information, Course 6.574. MIT, Cambridge, MA, 1952.

[4] R. G. Gallager. A simple derivation of the coding theorem and some applications. *IEEE Trans. Inform. Theory*, IT-11: 3-18, 1965.

[5] E. W. Gilbert and E. F. Moore. Variable length binary encodings. *Bell Sys. Tech. Journal*, 38: 933-967, 1959.

[6] J. Karush. A. simple proof of an inequality of McMillan. *IRE Trans. inform. Theory*, IT-7: 118, 1961.

[7] L. Lovasz. On the Shannon capacity of a graph. *IEEE Trans. Inform. Theory*, IT-25: 1-7, 1979.

[8] B. McMillan. The basic theorems of information theory. *Ann. Math. Stat.*, 24: 196-219, 1953

[9] B. McMillan. Two inequalities implied by unique decipherability. *IEEE Trans. Inform. Theory*, IT-2: 115-116, 1956.

[10] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.

[11] C. E. Shannon. The zero-error capacity of a noisy channel. *IRE Trans. Inform.*

Theory, IT-2: 8-19, 1956.

DEPARTMENT OF MATHEMATICS,
WOOSUK UNIVERSITY,
WANJU-GUN 565-701 CHONBUK,
KOREA.
yslee@core.woosuk.ac.kr