Differential Entropy

2009 2-2 Course

- Information Theory -

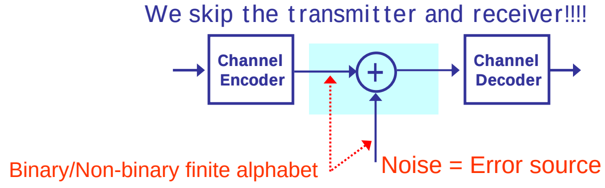
Tetsuo Asano and
Tad matsumoto
Email: {t-asano, matumoto}@jaist.ac.jp

Japan Advanced Institute of Science and Technology
Asahidai 1-1, Nomi, Ishikawa 923-1292, Japan
http://www.jaist.ac.jp

In the last Chapters, we learned .....

We assumed that the channel input and output both take discrete values:

We skip the transmitter and receiver!!!!



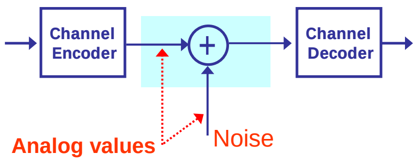
We derived Channel Coding Theorem:

(1) There exists a $(2^{nR}, n)$ rate R code such that the maximum error probability $\lambda^{(n)}$ can be made arbitrarily small, if the code rate is lower than the capacity $R<C$.
(2) Conversely, any $(2^{nR}, n)$ rate R code that can achieve arbitrarily small $\lambda^{(n)}$ must satisfy $R<C$.

In this and the next Chapters, we derive .....

Channel Coding Theorem for the channel where input and output of the channel both take analog values.

For this, we eliminate the finite alphabet assumption.



However, still we ignore the transmitter and receiver. The impact of having to use physical transmitter and receiver is investigated in another course, “Channel Coding”.

Outline

1. Differential Entropy
   - Definition
   - Some Examples
   - Asymptotic Property
2. Relation of Differential and Discrete Entropies
3. Joint and Conditional Differential Entropy
4. Mutual Information
   - Some Important Properties

Differential Entropy (1)

Definition 9.1.1: Differential Entropy

For continuous random variable X with a probability density function $f(x)$, the differential entropy is defined as:

$$h(X) = -\int_S f(x)\log f(x)dx$$

where S is the set where X is defined.

Example: Uniform Distribution

If a random variable X is distributed over a uniform distribution $f(x) = 1/a, \quad 0 \le x \le a$ the differential entropy of X is:

$$h(X) = -\int_0^a \frac{1}{a}\log\frac{1}{a}dx = \log a$$

Property: Value Range

The differential entropy can take negative values, as seen in the above example.

Differential Entropy (2)

Example: Normal Distribution

If a random variable X is distributed over a Normal distribution $N(0,\sigma^2)$ the differential entropy of X is:

$$h(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx \quad \text{with} \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

After minor mathematical manipulations, we have:

$$h(f) = \frac{E(x^2)}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 = \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 = \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2$$

$$= \frac{1}{2}\ln 2\pi e\sigma^2 \quad [nats] \quad = \frac{1}{2}\log 2\pi e\sigma^2 \quad [bits]$$

Theorem 9.1.2: Asymptotic Property

Let X_1, X_2, ... , X_n be a sequence of i.i.d. random variable following a density function $f(x)$. Then, the following holds:

$$-\frac{1}{n}\log f(x_1, x_2, \cdots, x_n) \underset{n\to\infty}{\rightarrow} E(-\log f(x)) = h(X)$$

Proof: Obvious

Relation of Differential and Discrete Entropies (1)

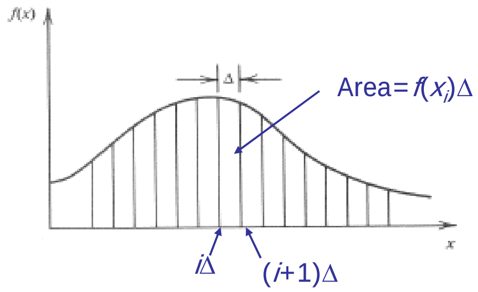
Property:

There exists a value x_i such that the probability that a random variable X takes a value between $i\Delta$ and $(i+1)\Delta$ is given by:

$$\Pr(i\Delta \le X \le (i+1)\Delta) = f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$$

Where $f(x)$ is the density function of X.

Proof: Obvious



Relation of Differential and Discrete Entropies (2)

Definition 9.2.1: X^Δ

Define a discrete random variable, $X^\Delta = x_i, \quad if \quad i\Delta \le x \le (i+1)\Delta$

Entropy of this random variable X^Δ is given by:

$$H(X^\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i, \quad with \quad p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

which leads to:

$$H(X^\Delta) = -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) = -\sum_{-\infty}^{\infty} \Delta f(x_i)\log f(x_i) - \underbrace{\sum_{-\infty}^{\infty} \Delta f(x_i)}_{=1}\log\Delta$$

$$= -\sum_{-\infty}^{\infty} \Delta f(x_i)\log f(x_i) - \log\Delta$$

Relation of Differential and Discrete Entropies (3)

Theorem 9.2.1: Discrete → Continuous

$$H(X^{\Delta}) + \log\Delta = -\sum_{-\infty}^{\infty} \Delta f(x_i)\log f(x_i) \to h(f) = h(X), \quad as \quad \Delta \to 0$$

Proof: Obvious

Example:

As we saw previously in an example, differential entropy $h(X)$ of a random variable uniformly distributed over [0, a] is log a, and with a=1, $h(X)$=0. If we "quantize" this random variable following distribution [0, 1] with n-bit A/D converter,

$$h(X) = H\left(X^{\Delta}\right) + \log\Delta = H\left(X^{\Delta}\right) + \log 2^{-n} = H\left(X^{\Delta}\right) - n = 0$$

Therefore,

$$H\left(X^{\Delta}\right) = n$$

which means that n bits are enough to express the *contiguous* random variable X while keeping n-bit accuracy.

Joint and Conditional Differential Entropy (1)

Definition 9.3.1: Joint Differential Entropy

The differential entropy of a set $X_1, X_2, \ldots, X_n$ of random variables following the density function $f(x_1, x_2, \ldots, x_n)$ is defined as:

$$h(X_1, X_2, \cdots, X_n) = -\int f(x_1, x_2, \cdots, x_n)\log f(x_1, x_2, \cdots, x_n)dx_1 dx_2 \cdots dx_n$$

Definition 9.3.2: Conditional Differential Entropy

The conditional differential entropy of random variables X and Y is defined as:

$$h(X|Y) = -\int f(x,y)\log f(x|y)dxdy = -\int f(x,y)\log\frac{f(x,y)}{f(y)}dxdy = h(X,Y) - h(Y)$$

where $f(X,Y)$ is the joint density of the random variable X and Y.

Joint and Conditional Differential Entropy (2)

Theorem 9.3.1: Normal Distribution

The differential entropy of random variables $X_1, X_2, \ldots, X_n$ following the multivariate normal density function

$$f(x_1, x_2, \cdots, x_n) = f(\boldsymbol{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\boldsymbol{K}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

is given by

$$h(X_1, X_2, \cdots, X_n) = h(N_n(\boldsymbol{\mu}, \boldsymbol{K})) = \frac{1}{2}\log(2\pi e)^n |\boldsymbol{K}| \quad bits$$

where $\boldsymbol{\mu} = E(\boldsymbol{x})$ and

$$\boldsymbol{K} = E\left[(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\right] = E\left[\begin{pmatrix}(x_1-\mu_1)\\(x_2-\mu_2)\\ \vdots \\(x_n-\mu_n)\end{pmatrix}\begin{pmatrix}(x_1-\mu_1) & (x_2-\mu_2) & \cdots & (x_n-\mu_n)\end{pmatrix}\right]$$

Proof:

$$h(f) = -\int f(\boldsymbol{x})\log f(\boldsymbol{x})d\boldsymbol{x} = -\int f(\boldsymbol{x})\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - \ln(\sqrt{2\pi})^n |\boldsymbol{K}|^{1/2}\right]d\boldsymbol{x}$$

Joint and Conditional Differential Entropy (3)

Proof *Continued*:

$$= E\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] + \ln(\sqrt{2\pi})^n |\boldsymbol{K}|^{1/2}$$

$$= \frac{n}{2} + \ln(\sqrt{2\pi})^n |\mathbf{K}|^{1/2}$$

$$= \frac{1}{2}\ln e^n + \ln(\sqrt{2\pi})^n |\mathbf{K}|^{1/2} = \frac{1}{2}\ln(2\pi e)^n |\mathbf{K}| \quad nats = \frac{1}{2}\log(2\pi e)^n |\mathbf{K}| \quad bits$$

Exercise: Provide more detailed and more concrete proof for this mathematical manipulation.

Mutual Information (1)

Definition 9.4.1: Kullback Leibler Distance

Kullback Leibler distance $D(f||g)$ between two density functions is defined as:

$$D(f\|g)=\int_S f\log\frac{f}{g}$$

with $0\,log\,(0/0)=0$ and S being the region where the random variables are defined.

Definition 9.4.2: Mutual Information

The mutual information $I(X,Y)$ between the two continuous random variables X and Y with the joint density function $f(x,y)$ is defined as

$$I(X;Y)=\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

Property 9.4.1:

$$I(X;Y)=h(X)-h(X|Y)=h(Y)-h(Y|X)=D\big(f(x,y)\|f(x)f(y)\big)$$

Proof: Obvious.

Mutual Information (2)

Theorem 9.4.1: Non Negativity of Kullback Leibler Distance

$D(f\|g)\ge 0$, and equality holds if and only if $f=g$.

Proof: $$-D(f\|g)=\int_S f\log\frac{g}{f}\underbrace{\le}_{due\ \ to\ \ Jensen's\ \ unequality}\log\int_S f\frac{g}{f}=\log\int_S g\le\log 1=0$$

Jensen's inequality for concave functions states that the equality holds if and only of $f=g$.

Theorem 9.4.2: Non Negativity of Mutual Information

$I(X;Y)\ge 0$, and equality holds if and only X and Y are independent.

Proof: Obvious

Theorem 9.4.3: Knowledge Decreases Uncertainty

$h(X|Y)\le h(X)$, and equality holds if and only X and Y are independent.

Proof: Obvious

Mutual Information (3)

Theorem 9.4.4: Chain Rule

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

Proof: Obvious

Theorem 9.4.5:

$$h(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} h(X_i)$$

and equality holds if and only if X and Y are independent.

Proof: Obvious

Theorem 9.4.6:

$$h(X + c) = h(X)$$

i.e., translation does not change the differential entropy.

Proof: Obvious

Mutual Information (4)

Theorem 9.4.7: Normal Distribution Maximize Entropy

The Normal distribution maximizes the entropy among those distributions having the same variance.

Proof: Let X be a zero mean random variable with variance σ^2, following the density function $g(x)$. Also, let $\phi(X)$ be a zero mean Normal random variable with variance σ^2. Then,

$$0 \le D(g\|\phi) = \int g \ln \frac{g}{\phi} dx = -h(g) - \int g \ln \phi dx = -h(g) - \int \left[-\ln \sqrt{2\pi}\sigma - \frac{x^2}{2\sigma^2} \right] g dx$$

$$= -h(g) + \ln \sqrt{2\pi}\sigma + \frac{1}{2\sigma^2} \int x^2 g dx$$

However, since g and f have the same variance, $\int x^2 g dx = \int x^2 \phi dx$

Therefore,

$$0 \le D(g\|\phi) = -h(g) + \ln \sqrt{2\pi}\sigma + \frac{1}{2\sigma^2} \int x^2 g dx = -h(g) + \ln \sqrt{2\pi}\sigma + \frac{1}{2\sigma^2} \int x^2 \phi dx$$

$$= -h(g) - \int \phi \ln \phi dx = -h(g) + h(\phi)$$

The equality holds if $g = \phi$

Mutual Information (5)

Theorem 9.4.8: Normal Distribution Maximize Entropy

- Multi-variable's Case -

The n-dimensional zero-mean Normal distribution maximizes the entropy among other n-dimensional zero-mean distributions having the same covariance matrix $\boldsymbol{K}=\{K_{ij}\}$.

Proof: **Exercise**

Summary

We have visited.....

1. Differential Entropy
   - Definition
   - Some Examples
   - Asymptotic Property
2. Relation of Differential and Discrete Entropies
3. Joint and Conditional Differential Entropy
4. Mutual Information
   - Some Important Properties