Elements of Information Theory
Ch. 6 Differential Entropy and the Gaussian Channel

Bilingual course
(Chinese taught course)
Information and Communication Eng. Dept.
Deng Ke

Introduction

- So far, we have investigated various aspects of information theory, but we only considered discrete random variables.
- How to deal with the continuous random variables?
- concept of differential entropy, which is the entropy of a continuous random variable.
- Differential entropy is similar in many ways to the entropy of a discrete random variable.
- We will also derive the famous formula operational capacity of additional Gaussian white noise channel.

Outline

- Differential Entropy
- Relation to discrete entropy
- Mutual Information
- Properties and Relations
- AEP for Continuous Random Variables
- Gaussian Channel
- Channel Capacity of Gaussian Channel

Differential Entropy

- need to define entropy, mutual information between CONTINUOUS random variables
- Definition: The *differential entropy* $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

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

- where S is the support set of the random variable.
- Support set of X is the set where $f(x)>0$

Differential Entropy

- Differences with entroy
  - Some times the density function does not exist for a random variables or the above integral does not exist.
  - Differential entropy can be negative.
- Example 8.1.1 (Uniform distribution) Consider a random variable distributed uniformlyfrom 0 to a so that its density is $1/a$ from 0 to a and 0 elsewhere.Then its differential entropy is

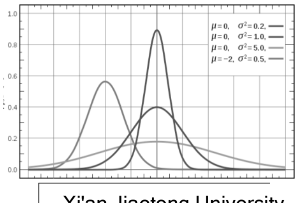
$$h(X) = -\int_0^a \frac{1}{a}\log\frac{1}{a}dx = \log a$$

- Note:For $a< 1$, $\log a< 0$, and the differential entropy is negative.

Differential Entropy

- Example 8.1.2 (Normal distribution)

$$X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-x^2}{2\sigma^2}}$$



$$\begin{aligned}
h(\phi) &= -\int \phi \ln \phi \\
&= -\int \phi(x)\left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right] \\
&= \frac{EX^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 \\
&= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 \\
&= \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2 \\
&= \frac{1}{2}\ln 2\pi e\sigma^2 \qquad \text{nats.}
\end{aligned}$$

Differential Entropy

- Relation of differential entropy and discrete entropy
  - Consider a random variable X with density $f(x)$.
  - Divide the range of X into bins of length Δ. There exists a value within each bin such that
    $f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$
  - Quantize random variable
    $X^{\Delta} = x_i \quad if \ i\Delta \le X < (i+1)\Delta$
  - Then the probability that
    $p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$

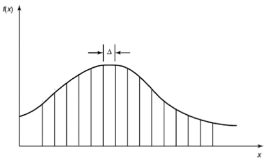


FIGURE 8.1. Quantization of a continuous random variable.

Differential Entropy

- Quantized entropy

$$
\begin{aligned}
H(X^{\Delta}) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\
&= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\
&= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\
&= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta ,
\end{aligned}
$$

When $\Delta \to 0$, the first term is $-\int f(x)\log f(x)dx$

Differential Entropy

- Theorem 8.3.1 If the density $f(x)$ of the random variable X is Riemann integrable, then
  $H(X^{\Delta}) + \log\Delta \to h(f) = h(X), \quad as \ \Delta \to 0$
- Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X)+n$.
- Conclusions of differential entropy
- Good ones
  - $h_1(X)$-$h_2(X)$ does compare the uncertainly of two continuous r.v. (quantized to the same precision)
  - Mutual information still works

Differential Entropy

- Bad ones and ugly
  - $h(X)$ does not give the amount information in X
  - $h(X)$ is not necessarily positive
  - $h(X)$ changes with a change of coordinate system
- Theorem 8.6.4

$$h(aX)=h(X)+\log|a|$$

  - Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$

$$h(aX) = -\int f_Y(y)\log f_Y(y)dy = -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right)\log\left[\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right]dy$$

$$= -\int f_X(x)\log f_x(x)dx + \log|a| = h(X) + \log|a|$$

AEP for continuous random variables

- Consider a multiple channel
- AEP for continuous random variables
- Theorem 8.2.1 *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots X_n) \to E[-\log f(X)] = h(X)$$ In probability

AEP for continuous random variables

- *Definition* For $\varepsilon > 0$ and any n, we define the *typical set* $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) \in S^n : \left|-\frac{1}{n}\log f(x_1, x_2, \ldots, x_n) - h(X)\right| \le \epsilon\right\}$$

- where $f(x_1, x_2, \ldots x_n) = \Pi_{i=1}^{n} f(x_i)$
- the volume of the typical set for continuous random variables is the analog of the cardinality of the typical set for the discrete case
- *Definition* The *volume* Vol(A) of a set $A \subset R^n$ is defined as $\mathrm{Vol}(A) = \int_A dx_1 dx_2 \ldots dx_n$

AEP for continuous random variables

- Theorem8.2.2 *The typical set* $A_\epsilon^{(n)}$ *has the following properties:*
  - $\Pr(A_\epsilon^{(n)}) > 1-\varepsilon$ *for n sufficiently large*
  - $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\varepsilon)}$ *for all n*
  - $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\varepsilon) 2^{n(h(X)-\varepsilon)}$ *for n sufficiently large*

$$
\begin{aligned}
1 &= \int_{S^n} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \\
&\ge \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \\
&\ge \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}\, dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)}\, \mathrm{Vol}\left(A_\epsilon^{(n)}\right).
\end{aligned}
$$

Joint Differential Entropy

- *Definition* The *differential entropy* of a set $X_1, X_2, \ldots, X_n$ of random variables with density $f(x_1, x_2, \ldots x_n)$ is defined as $h(X_1, X_2, \ldots, X_n) = -\int f(x^n) \log f(x^n) dx^n$
- *Definition* If X,Y have a joint density function $f(x,y)$, we can define the conditional differential entropy $h(X|Y)$ as $h(X|Y) = -\int f(x,y) \log f(x|y) dxdy$
- And we have $h(X|Y)=h(X, Y)- h(Y)$
- Theorem8.4.1 (*Entropy of a multivariate normal distribution*) *Let* $X_1, X_2, \ldots, X_n$ *have a multivariate normal distribution with mean* μ *and covariance matrix K. Then*

Joint Differential Entropy

$$h(X_1, X_2, \ldots, X_n) = \tfrac{1}{2}\log(2\pi e)^n |K|$$

- *where* $|K|$ *denotes the determinant of K.*
- Proof $f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}.$

$$
\begin{aligned}
h(f) &= -\int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1} (\mathbf{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}} \right] d\mathbf{x} \\
&= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i)\left(K^{-1}\right)_{ij} (X_j - \mu_j) \right] + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i)(X_j - \mu_j)\left(K^{-1}\right)_{ij} \right] + \frac{1}{2}\ln(2\pi)^n |K|
\end{aligned}
$$

Joint Differential Entropy

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i,j} E[(X_j - \mu_j)(X_i - \mu_i)]\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\sum_j \sum_i K_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\sum_j \left(K K^{-1}\right)_{jj} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\sum_i I_{jj} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\ln(2\pi e)^n |K| \\
&= \frac{1}{2}\log(2\pi e)^n |K|
\end{aligned}
$$

Mutual Information

- *Definition* The *mutual information* $I(X;Y)$ between two random variables with joint density $f(x,y)$ is defined as

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

- ~~Corollary $I(X;Y) \geqslant 0$ *with equality iff X and Y are independent*.~~
- ~~Corollary $h(X|Y) \leqslant h(X)$ *with equality iff X and Y are independent*.~~

Gaussian Distribution

- When we have the constraints that $EX=0$ and $EX^2=\sigma^2$, the Gaussian (normal) distribution have the maximum differential entropy.
- Proof: Let $p(x) \sim N(0,\sigma^2)$, we will show that the differential entropy of another distribution $q(x)$ will not greater than that of $p(x)$

$$
\begin{aligned}
-\int q(x) \log p(x) dx &= -\int q(x) \log\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2} \right] \right\} dx \\
&= -\int q(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx + \log e \int q(x) \frac{x^2}{2\sigma^2} dx
\end{aligned}
$$

Gaussian Distribution

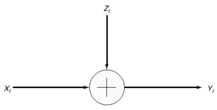
$$= \tfrac{1}{2}\log 2\pi\sigma^2 + \log e\frac{\sigma^2}{2\sigma^2} = \tfrac{1}{2}\log(2\pi e\sigma^2)$$

$$h(X, q(x)) - \int q(x)\log\frac{1}{p(x)}dx = \int q(x)\log\frac{p(x)}{q(x)}dx$$

$$\le \log\int q(x)\frac{p(x)}{q(x)}dx = \log 1 = 0$$

$$\therefore h(x, q(x)) \le \tfrac{1}{2}\log(2\pi e\sigma^2)$$

- Additive Gaussian white noise channel
  - Output $Y_i = X_i + Z_i$
  - The noise $Z_i \sim N(0,N)$
  - Z_i is independent of the signal X_i



AWGN Channel

$$P_{Y|X}(y \mid x) = P_{Y|X}(x + z \mid x) = P_Z(z) = P_Z(y - x)$$

- There is an average energy or power constraint on the input, what is the capacity?
- *Definition*: The information capacity of the Gaussian channel with power constraint P is

$$C = \max_{EX^2 \le P} I(X;Y)$$

$$I(X;Y) = h(Y) - h(Y \mid X) = h(Y) - h(X + Z \mid X)$$
$$= h(Y) - h(Z \mid X) = h(Y) - h(Z)$$

AWGN Channel

$$h(Z) = \tfrac{1}{2}\log(2\pi eN)$$

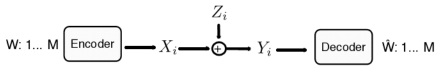
$$EY^2 = E(X+Z)^2 = EX^2 + 2EXEZ + EZ^2 = P + N$$

$$I(X;Y) = h(Y) - h(Z)$$
$$\le \tfrac{1}{2}\log 2\pi e(P+N) - \tfrac{1}{2}\log 2\pi eN$$
$$= \tfrac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

- Hence $C = \max_{EX^2 \le P} I(X;Y) = \tfrac{1}{2}\log\left(1 + \frac{P}{N}\right)$
- The optimum input is Gaussian and the worst noise is Gaussian

The capacity of AWGN channel

- Operational capacity



Definition: An (M, n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, ..., M\}$.
2. An encoding function $x : \{1, 2, ..., M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), ..., x^n(M)$, satisfying the power constraint P; that is for every codeword
$$\sum_{i=1}^{n} x_i^2(w) \le nP, w = 1, 2, ..., M.$$
3. A decoding function
$$g : \mathcal{Y}^n \to \{1, 2, ..., M\}.$$

The capacity of AWGN channel

Definition: A rate R is said to be *achievable* with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates.

Theorem: The capacity of a Gaussian channel with power constraint P and noise variance N is

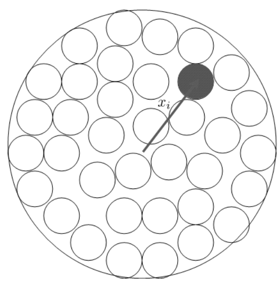
$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \quad \text{bits per transmission.}$$

Conversely, the rates $R > C$ are not achievable.

Intuition about why it works - sphere packing

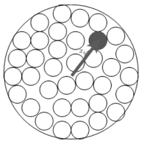
Each transmitted x_i is received as a probabilistic cloud y_i

The capacity of AWGN channel



The capacity of AWGN channel

- Volume of the 'cloud' $[2\pi eN]^{n/2}$



- Volume of the whole sphere

$$[2\pi e(P+N)]^{n/2}$$

- Max number of non-overlapping clouds

$$\frac{[2\pi e(P+N)]^{n/2}}{[2\pi eN]^{n/2}} = 2^{n\frac{1}{2}\log(1+\frac{P}{N})}$$

Bandlimited Gaussian Channels

- Nyquist Theorem
- Sampling a bandlimited signal at a sampling rate 1/2W is sufficient to reconstruct the signal from the samples.
- White noise with double-sided psd $\frac{1}{2}N_0$
- Capacity

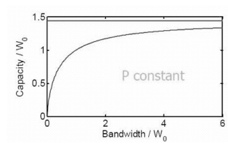
$$\begin{aligned} C &= \frac{1}{2}\log(1+\frac{1}{2}P/W(\frac{1}{2}N_0)^{-1})2W \\ &= W\log(1+\frac{P}{WN_0}) \text{ bits/second} \end{aligned}$$

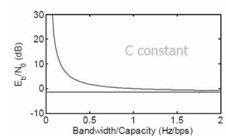
Shannon Equation

- This is also called Shannon equation
- It show the relations among the channel capacity, SNR, and bandwidth.
- Capacity will increase with the SNR
- But will have a limit with the increase of the bandwidth

$$W \to \infty \qquad C \to \frac{P}{N_0}\log_2 e$$

Shannon Equation





$$W \to \infty,\ \frac{E_b}{N_0} = \frac{W_0}{C} \to \ln 2 = -1.6\text{dB}$$

Example: telephone channel

- Telephone signals bandlimited to 3300Hz.
- SNR is 33dB.
- What is capacity of a telephone line?
- For further reading, please refer to
- David Tse, Fundamentals of Wireless Communication