So, if $R > C$ then $P_e^{(n)}$ is bounded away from zero for sufficiently large n, and, hence for all n. The latter comes from the fact that if $P_e^{(n)} \to 0$ for small n, we could construct long codes by concatenating short codes.

X Lecture 7, Oct. 14, 2003

Continuous Sources

A continuous random variable X is characterised by a Cumulative Distribution Function $F_X(x)$,

$$F_X(x) = P_r(X \leq x)$$

the derivative of $F_X(x)$ w.r.t. x is called the probability density function (pdf),

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

when there is no ambiguity, we drop the subscript and denote the pdf by $f(x)$.

It is clear that,

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

therefore

$$\int_S f(x)\,dx = 1$$

where S is the support of the random variable X, i.e., the set of x such that $f(x) > 0$.

Assume that a source generates $X_1, X_2, \dots$ i.i.d. and $\sim f(x)$. Then, for any n-tuple $(x_1, \dots x_n) = x^n$

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Taking logarithm of $f(x_1, \dots x_n)$ we have,

$$\log f(x_1, \dots, x_n) = \sum_{i=1}^{n} \log f(x_i)$$

normalizing by n, we have

$$-\frac{1}{n} \log f(x_1, \dots x_n) = \frac{-1}{n} \sum_{i=1}^{n} \log f(x_i)$$

The right-hand side of the above equation is the sample mean of $\log f(x_i)$ and according to law of large numbers, it tends to the expectation of $\log f(X)$ as $n \to \infty$. So,

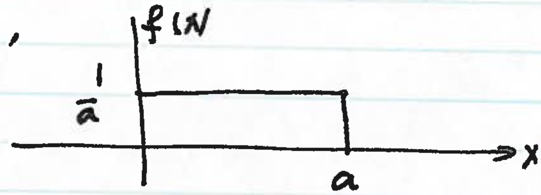
$$-\frac{1}{n} \log f(X_1, \dots X_n) \xrightarrow[n\to\infty]{} -E[\log f(X)] = -\int_S f(x) \log f(x)\, dx$$

We denote this expectation by $h(X)$ or $h(f)$ and call $h(X)$ differential entropy:

$$h(X) = -E[\log f(X)] = -\int_S f(x) \log f(x)\, dx$$

$h(X)$ is similar to $H(X)$ for discrete alphabet sources, however, fails to have some of the properties of $H(X)$ such as positivity which make $H(X)$ a measure of information.

<u>Example</u>: differential entropy for a uniformly distributed source,



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

Note that for $a > 1$ as a increases so does $h(X)$ and this is in agreement with $h(X)$ being a measure of uncertainty as increasing a makes the particular value of X more uncertain and, therefore, more information-beari

However, we also note that for $a < 1$, $h(X)$ is negative

On the other hand, $2^{h(X)} = 2^{\log a} = a$ is the volume of the support set of X which is always non-negative.

<u>Example</u>: $h(X)$ for a Gaussian source.

Here $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$

Then

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \log\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}\right] dx$$

or,

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log\left[\frac{1}{\sqrt{2\pi}\,\sigma}\right] dx$$

$$+ \int_{-\infty}^{\infty} f(x)\left[\frac{x^2}{2\sigma^2}\right] \log_2 e \, dx$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log_2 e \int_{-\infty}^{\infty} f(x)x^2 dx}{2\sigma^2}$$

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

$$= \frac{1}{2}\log(2\pi e\sigma^2) \quad \text{bits}$$

<u>Definition</u>: For any $\varepsilon > 0$ and any n, define the typical set $A_\varepsilon^{(n)}$ w.r.t. $f(x)$ as:

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

<u>Theorem</u> (AEP): The typical set $A_\varepsilon^{(n)}$ has the following properties:

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

where $Vol(A)$ for a set $A \in R^n$ is defined as

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

<u>Proof</u>:

$$-\frac{1}{n} \log f(x_1, x_2, \ldots, x_n) = -\frac{1}{n} \sum_i f(x_i) \rightarrow h(X)$$

So, for any $\varepsilon > 0$ there is some n_0 such that for any $n > n_0$, we have

$$P_r \left\{ \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \geq \varepsilon \right\} \leq \varepsilon$$

or

$$P_r \left((x_1, \ldots, x_n) \notin A_\varepsilon^{(n)} \right\} \leq \varepsilon$$

and, hence,

$$P_r \left((X_1, \ldots x_n) \in A_\varepsilon^{(n)} \right) \geq 1 - \varepsilon$$

This proves part 1.

For part 2:

$$1 = \int_S f(x_1, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

$$\geq \int_{A_\varepsilon^{(n)}} f(x_1, \ldots x_n)\, dx_1 \ldots dx_n$$

$$\geq \int_{A_\varepsilon^{(n)}} 2^{-n(h(X)+\varepsilon)}\, dx_1 \ldots dx_n$$

$$= 2^{-n(h(X)+\varepsilon)}\, Vol(A_\varepsilon^{(n)})$$

QED [property 2]

Property 3:

if n is large enough so that property 1 holds, then:

$$1-\varepsilon \leq \int_{A_\varepsilon^{(n)}} f(x_1, \dots x_n)\, dx_1 \dots dx_n$$

$$\leq \int_{A_\varepsilon^{(n)}} 2^{-n(h(X)-\varepsilon)}\, dx_1 \dots dx_n$$

$$= 2^{-n(h(X)-\varepsilon)}\, Vol\,(A_\varepsilon^{(n)})$$ QED[Property 3]

Joint and Conditional differential entropy:

Definition: The differential entropy of a set $X_1, X_2, \dots X_n$ of random variables is

$$h(X_1, \dots X_n) = -\int_{S^n} f(\underline{x}^n) \log f(\underline{x}^n)\, d\underline{x}^n$$

Definition: If X and Y have joint density function $f(x,y)$, we can define the conditional differential entropy $h(X|Y)$ as,

$$h(X|Y) = -\int f(x,y) \log f(x|y)\, dx\, dy.$$

substituting $f(x|y) = \frac{f(x,y)}{f(y)}$, we get

$$h(X|Y) = h(X,Y) - h(Y)$$

<u>Example</u>: The differential entropy of ~~a~~ the multivariate normal sequence is

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

where $|K|$ is the determinant of ~~a~~ the covariance matrix of $X_1, X_2, \ldots X_n$.

<u>Proof:</u>

$$f(\underline{x}^n) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}^n - \underline{\mu}^n)^T K^{-1} (\underline{x}^n - \underline{\mu}^n)\right]$$

Then,

$$h(\underline{x}^n) = -\int f(\underline{x}^n) \log f(\underline{x}^n)\, d\underline{x}^n$$

$$= \frac{\log e}{2}\int f(\underline{x}^n)\left[(\underline{x}^n - \underline{\mu}^n) K^{-1} (\underline{x}^n - \underline{\mu}^n)\right] d\underline{x}^n$$

$$+ \frac{1}{2}\log(2\pi)^n |K|$$

$$= \frac{\log e}{2} E\left[(\underline{x}^n - \underline{\mu}^n) K^{-1} (\underline{x}^n - \underline{\mu}^n)\right] + \frac{1}{2}\log(2\pi)^n |K|$$

$$= \frac{\log e}{2} E\left[\sum_i \sum_j (x_i - \mu_i) K^{-1}_{ij} (x_j - \mu_j)\right] + \frac{1}{2}\log(2\pi)^n |K|$$

$$= \frac{\log e}{2} \sum_i \sum_j E\left[(x_i - \mu_i)(x_j - \mu_j)\right] K^{-1}_{ij} + \frac{1}{2}\log(2\pi)^n |K|$$

$$= \frac{\log e}{2} \sum_i \sum_j K^{-1}_{ij} K_{ji} + \frac{1}{2}\log(2\pi)^n |K|$$

$$= \frac{\log e}{2} \sum_i (K^{-1}K)_{ii} + \frac{1}{2} \log (2\pi)^n |K|$$

$$= \frac{\log e}{2} \sum_i I_{ii} + \frac{1}{2} \log (2\pi)^n |K|$$

$$= \frac{n}{2} \log e + \frac{1}{n} \log (2\pi)^n |K|$$

$$= \frac{1}{2} \log (2\pi e)^n |K| \qquad \text{QED}$$

Mutual information:

Definition: The mutual information $I(X;Y)$ between X and $Y \sim f(x,y)$ is defined as:

$$I(X;Y) = \int_S f(x,y) \log \frac{f(x,y)}{f(x) f(y)} \, dx \, dy$$

$$= \int_S f(x,y) \log \frac{f(y|x)}{f(y)} \, dx \, dy$$

It is easy to show that

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

Note that

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

The properties of mutual information and diff. entr

1) $I(X;Y) \geq 0$ with equality iff X and Y are independent.

2) $h(X|Y) \leq h(X)$ with equality if X and Y are independent.

3) The chain rule for differential entropy

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

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

with equality if $X_1, \cdots X_n$ are independent.

Hadamard inequality: Let $X^n \sim \mathcal{N}(0, K)$ be a multivariate normal random variable, then

$$|K| \leq \prod_{i=1}^{n} K_{ii}$$

Proof:

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

$$h(X_i) = \frac{1}{2} \log (2\pi e K_{ii})$$

$$\frac{1}{2} \log (2\pi e)^n |K| \leq \frac{1}{2} \sum_{i=1}^{n} \log (2\pi e K_{ii})$$

$$\Rightarrow \log (2\pi e)^n |K| \leq \log \prod_{i=1}^{n} (2\pi e K_{ii})$$

$$\log(2\pi e)^n |K| \leq \log(2\pi e)^n \prod_{i=1}^{n} K_{ii}$$

$$\Rightarrow \quad |K| \leq \prod_{i=1}^{n} K_{ii}$$

an example situation where this inequality is useful is the MIMO channels where correlation between different paths results in reduction in capacity.

<u>Theorem</u> $h(X+c) = h(X)$

Proof: if $X \sim f(x)$ then $X+c \sim f(x+c)$ ~~over~~ with suppor $S+c$

$$h(X+c) = -\int_{S+c} f(X+c)\log f(X+c)\,dx = -\int_{S} f(x)\log f(x)dx$$

$$= h(X).$$

<u>Theorem</u>: $h(aX) = h(X) + \log|a|$

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

$$h(aX) = -\int f_Y(y)\log f_Y(y)\,dy$$

$$= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right)\log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right)dy$$

$$= -\int f_X(x)\log f_X(x)\,dx + \log|a|$$

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

Corollary: for a random vector $\underline{X}$, we have

$$h(\underline{A}\,\underline{X}) = h(\underline{X}) + \log|A|$$

where $|A|$ is the determinant of the matrix A.

The most difficult source to ~~encode~~ Compress:

The following theorem indicates that among all random vectors with zero mean and common covariance matrix $K = E[\underline{X}\,\underline{X}^T]$, the Gaussian vector has the largest differential entropy. This implies that the most difficult source to Compress is a Gaussian source.

Theorem: Let the random vector $X \in \mathbb{R}^n$ have zero mean and Covariance $K = E[XX^T]$, i.e., $K_{ij} = E[X_i X_j]$, $i,j = 1, 2, \ldots, n$. Then

$$h(\underline{X}) \leq \frac{1}{2}\log(2\pi e)^n |K|.$$

Proof: Let $p(\underline{x})$ be the pdf of $\underline{X}$

and $$q(\underline{x}) = \frac{1}{(\sqrt{2\pi}\,)^n |K|^{1/2}} e^{-\frac{1}{2}\underline{x}^T K^{-1}\underline{x}}$$

Then,

$$\int_{\mathbb{R}^n} p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})} d\underline{x} = \int_{\mathbb{R}^n} p(\underline{x}) \ln\left[\frac{p(\underline{x})}{q(\underline{x})}\right] \log e \, d\underline{x}$$

$$\geq \int_{\mathbb{R}^n} p(\underline{x})\left[1 - \frac{q(\underline{x})}{p(\underline{x})}\right] \log e \, d\underline{x} = 0$$

But,

$$\int_{\mathbb{R}^n} p(\underline{x}) \log \frac{p(x)}{q(\underline{x})} dx = \int_{\mathbb{R}^n} p(\underline{x}) \log p(x) dx$$

$$- \int_{\mathbb{R}^n} p(\underline{x}) \log q(\underline{x}) d\underline{x} = -h(x) - \int_{\mathbb{R}^n} p(x) \log \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} dx$$

$$+ \frac{\log e}{2} \int_{\mathbb{R}^n} (\underline{x}^T K^{-1} \underline{x}) \, p(\underline{x}) d\underline{x}$$

$$= -h(\underline{X}) + \frac{1}{2}\log(2\pi)^n |K| + \frac{\log e}{2} \int_{\mathbb{R}^n} \Big(\sum_i \sum_j x_i x_j (K^{-1})_{ij}\Big) p(\underline{x})$$

$$= -h(\underline{X}) + \frac{1}{2}\log(2\pi)^n |K| + \frac{\log e}{2} \sum_i \sum_j \left[\int_{\mathbb{R}^n} x_i x_j \, p(\underline{x}) d\underline{x}\right] (K^{-1})_{ij}$$

$$= -h(\underline{X}) + \frac{1}{2}\log(2\pi)^n |K| + \frac{\log e}{2} \sum_i \sum_j K_{ji} (K^{-1})_{ij}$$

$$= -h(\underline{X}) + \frac{1}{2}\log(2\pi)^n |K| + \frac{\log e}{2} \underbrace{\sum_i (K^{-1}K)_{ii}}_{n}$$

$$= -h(\underline{X}) + \frac{1}{2}\log(2\pi e)^n |K| \geq 0$$

so,

$$h(X) \leq \frac{1}{2}\log(2\pi e)^n|K|$$

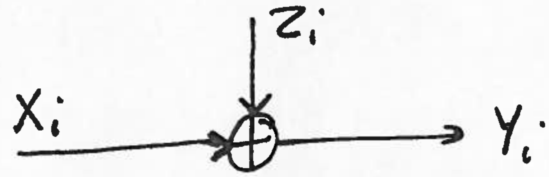
X Lecture 8, Oct. 21, 2003

Gaussian Channel

The channel we consider here is a discrete time channel with inputs $X_1, X_2, \ldots$ and output $Y_1, Y_2, \ldots$ where

$$Y_i = X_i + Z_i$$

where $Z_i \sim N(0, N)$



if the noise power (variance) N is zero or the transmission power is limitless, then it is possible to transmit an infinite number of bits per use. In such a case (unconstrained) capacity of the channel is infinite. In practice, however, there is a limit on the transmi