

# Differential Entropy

吳家麟教授

# Definition

- Let  $X$  be a random variable with cumulative distribution function  $F(x) = P_r(X \leq x)$ . If  $F(x)$  is continuous, the r.v. is said to be continuous. Let  $f(x) = F'(x)$  when the derivative is defined. If  $\int_{-\infty}^{\infty} f(x)dx = 1$ , then  $f(x)$  is called the pdf for  $X$ . The set where  $f(x) > 0$  is called the support set of  $X$ .

# Definition

- The differential entropy  $h(X)$  of a continuous r.v.  $X$  with a density function  $f(x)$  is defined as

$$h(X) = - \int_S f(x) \log f(x) dx \quad (1)$$

where  $S$  is the support of the r.v.

- Since  $h(X)$  depends only on  $f(x)$ , sometimes the differential entropy is written as  $h(f)$  rather than  $h(X)$ .

# EX.1: (Uniform distribution)

$$f(x) = \begin{cases} \frac{1}{a}, & 0 \leq x < a \\ 0, & \text{otherwise} \end{cases}$$

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

Note: For  $a < 1$ ,  $\log a < 0$ , and  $h(X) = \log a < 0$ .

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

# Ex. 2: (Normal distribution)

- Let  $X \sim \Phi = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$ , then

$$\begin{aligned} h(\Phi) &= -\int \Phi \ln \Phi \\ &= -\int \Phi(x) \left[ -\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] \\ &= \frac{1}{2\sigma^2} E[X^2] + \ln(2\pi\sigma^2)^{\frac{1}{2}} \\ &= \frac{1}{2\sigma^2} \sigma^2 + \frac{1}{2} \ln 2\pi\sigma^2 \end{aligned}$$

$$= \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2$$

$$= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2$$

$$= \frac{1}{2} \ln 2\pi e \sigma^2 \quad \text{nats}$$

Changing the base of the logarithm, we have

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

# Theorem 1

- Let  $X_1, X_2, \dots, X_n$  be a sequence of rv's drawn i.i.d. according to the density  $f(x)$ . Then

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow$$

$$E[-\log f(X)] = h(X) \quad \text{in probability}$$

proof: The proof follows directly from the weak law of the large numbers.

- Def: For  $\varepsilon > 0$  and any  $n$ , we define the typical set  $A_\varepsilon^{(n)}$  w.r.t.  $f(x)$  as follows:

$$A_\varepsilon^{(n)} = \left\{ (X_1, X_2, \dots, X_n) \in S^n : \left| -\frac{1}{n} \log f(X_1, X_2, \dots, X_n) - h(X) \right| \leq \varepsilon \right\},$$

where  $f(X_1, X_2, \dots, X_n) = \prod_{i=1}^n f(X_i)$

- Def: The volume  $\text{Vol}(A)$  of a set  $A \in \mathbb{R}^n$  is defined as

$$\text{vol}(A) = \int_A dx_1 dx_2 \cdots dx_n$$

- Thm: The typical set  $A_\epsilon^{(n)}$  has the following properties:

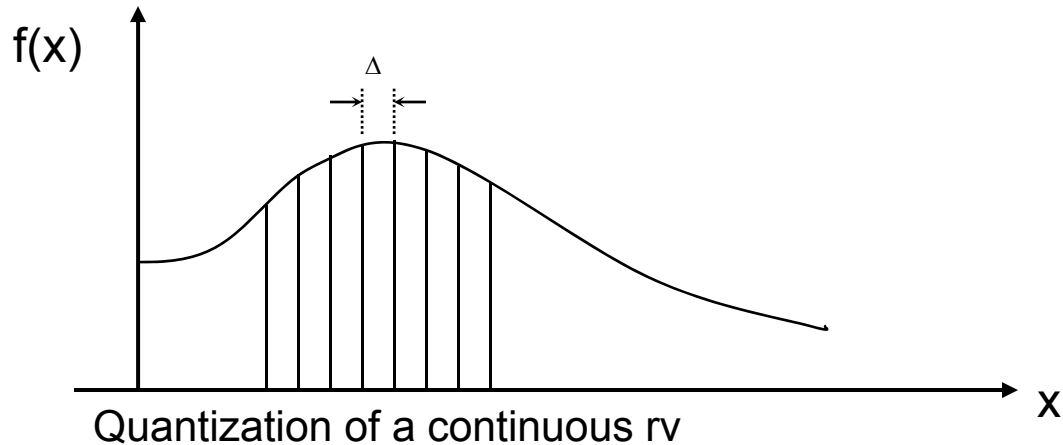
1.  $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$  for  $n$  sufficiently large

2.  $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X+\epsilon))}$  for all  $n$

3.  $\text{Vol}(A_\epsilon^{(n)}) \geq (1 - \epsilon)2^{n(h(X-\epsilon))}$  for  $n$  sufficiently large

- Thm: The Set  $A_{\epsilon}^{(n)}$  is the smallest volume set with probability  $\geq 1-\epsilon$ , to the first order in the exponent.
  - $\Rightarrow$  The volume of the smallest set that contains most of the Prob. Is approximately  $2^{nh}$ . This is an n-D volume, so the corresponding side length is  $(2^{nh})^{1/n}=2^h$ .
  - $\Rightarrow$  The differential entropy is the logarithm of the equivalent side length of the smallest set that contains most of the Prob.
  - $\Rightarrow$  low entropy implies that the rv is confined to a small effective volume and high entropy indicates that the rv is widely dispersed.

# Relation of Differential Entropy to Discrete Entropy



S'pose we divide the range of  $x$  into bins of length  $\Delta$ . Let's assume that the density is continuous within the bins.

By the mean value theorem, there is a value  $x_i$  within each bin such that

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

Consider the quantized rv  $X^\Delta$ , which is defined by

$$X^\Delta = x_i, \quad \text{if } i\Delta \leq X < (i+1)\Delta$$

Then the prob. that  $X^\Delta = x_i$  is

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

The entropy of the quantized version is

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

Since

$$\sum f(x_i) \Delta = \int f(x) = 1$$

If  $f(x)\log f(x)$  is Riemann integrable, then

$$-\sum \Delta f(x_i) \log f(x_i) \rightarrow -\int f(x) \log f(x) dx, \text{ as } \Delta \rightarrow 0$$

This proves the following

Thm: If the density  $f(x)$  of the rv  $X$  is Riemann integrable, then

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \text{ as } \Delta \rightarrow 0$$

Thus the entropy of an  $n$ -bit quantization of a continuous rv  $X$  is approximately  $h(X) + n$

*Since*  $\Delta = 2^{-n}$  for a  $n$ -bit uniform quantizer

## Joint and Conditional Differential Entropy:

- $h(X_1, X_2, \dots, X_n) = -\int f(x^n) \log f(x^n) dx^n$
- $h(X | Y) = -\int f(x, y) \log f(x | y) dx dy$
- $h(X | Y) = h(X, Y) - h(Y)$

## Theorem (Entropy of a multivariate normal distribution)

Let  $X_1, X_2, \dots, X_n$  have a multivariate normal distribution with mean  $\mu$  and covariance matrix  $K$ . Then

$$h(X_1, X_2, \dots, X_n) = h(\mathbf{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| \text{ bits}$$

where  $|K|$  denotes the determinant of  $K$ .

$$\text{pf} : (X_1, X_2, \dots, X_n) \sim N_n(\mu, K) \Rightarrow f(\tilde{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\tilde{x}-\mu)^T K^{-1}(\tilde{x}-\mu)}$$

Then

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

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

## Relative Entropy and Mutual Information

- $D(f // g) = \int f \log \frac{f}{g}$

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

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

- $I(X;Y) = D(f(x, y) // f(x)f(y))$

Remark:

The mutual information between two continuous r.v.'s is the limit of the mutual information between their quantized versions.

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta | Y^\Delta) \\ &\approx h(X) - \log \Delta - (h(X | Y) - \log \Delta) = I(X; Y) \end{aligned}$$

Properties of  $h(x)$ ,  $D(p \parallel q)$ ,  $I(X; Y)$

- $D(f \parallel g) \geq 0$

pf :-  $D(f \parallel g) = \int_s f \log \frac{fg}{f} \leq \log \int_s f \frac{fg}{f}$  (Jensen's inequality)

$= \log \int_s g \leq \log 1 = 0.$

- $I(X; Y) \geq 0$

- $h(X | Y) \leq h(X)$

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

- $h(X_1, X_2, \dots, X_n) = \sum h(X_i)$

Theorems •  $h(X + c) = h(x)$  : translation does not change the differential entropy

- $h(aX) = h(x) + \log|a|$

*pf* : let  $Y = aX$ . Then ,  $f_Y(y) = \frac{1}{|a|} f_x\left(\frac{y}{a}\right)$ , and

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

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

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

**Corollary** :  $h(AX) = h(X) + \log|\det(A)|$

Theorem : The multivariate normal distribution maximizes the entropy over all distributions with the same variance.

Let the random vector  $X \in \mathbb{R}^n$  have zero mean and covariance

$K = E XX^T$  (i.e.,  $K_{ij} = EX_i X_j$ ,  $1 \leq i, j \leq n$ ) , Then

$$h(\mathbf{x}) \leq \frac{1}{2} \log (2\pi e)^n |K|, \text{ with equality iff } X \sim N_n(0, k)$$

**Pf:**

Let  $g(\tilde{x})$  be any density satisfying  $\int g(\tilde{x}) x_i x_j d\tilde{x} = K_{ij}$  for all  $i, j$ .

Let  $\phi_k$  be the density of a  $N(0, k)$  vector.

Note that  $\log \phi_k(\tilde{x})$  is a quadratic form and  $\int x_i x_j \phi_k(\tilde{x}) d\tilde{x} = k_{ij}$ .

$$\begin{aligned} 0 \leq D(g // \phi_k) &= \int g \log\left(\frac{g}{\phi_k}\right) \\ &= -h(g) - \int g \log \phi_k = -h(g) - \int \phi_k \log \phi_k \\ &= -h(g) + h(\phi_k) \end{aligned}$$

where the substitution  $\int g \log \phi_k$  follows from the fact that  $g$  and  $\phi_k$  yield the same moments of the quadratic form  $\log \phi_k(\mathbf{x})$

$\Rightarrow$  the Gaussian distribution maximizes the entropy over all distributions with the same variance.

Let  $X$  be a random variable with differential entropy  $h(x)$

Let  $\hat{X}$  be an estimate of  $X$  and let  $E(X - \hat{X})^2$  be the expected prediction error. Let  $h(x)$  be in nats

Theorem : For any  $r^\sigma$   $X$  and estimator  $\hat{X}$

$$E(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(x)}$$

with equality iff  $X$  is Gaussian and  $\hat{X}$  is the mean of  $X$

pf : Let  $\hat{X}$  be any estimator of  $X$  then

$$E(X - \hat{X})^2 \geq \min_{\hat{x}} E(X - \hat{X})^2 \quad (1)$$

$$= E(X - E(X))^2 \quad [\text{the mean of } X \text{ is the best estimator for } X]$$

$$= \text{var}(X) \geq \frac{1}{2\pi e} e^{2h(x)} \quad (2)$$

[Gaussian distribution has the maximum entropy for a given variance]

$$\text{i.e., } h(x) \leq \frac{1}{2} \ln 2\pi e \sigma^2$$

We have equality, only in (1), only if  $\hat{x}$  is the best estimator (i.e.,  $\hat{x}$  is the mean of  $X$ ) and equality in (2) only if  $X$  is Gaussian.

Corollary: Given side information  $Y$  and estimator  $\hat{X}(Y)$

it follows that

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}$$

→ Fano's inequality