

INFORMATION THEORY & CODING

Week 10 : Differential Entropy

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern University of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

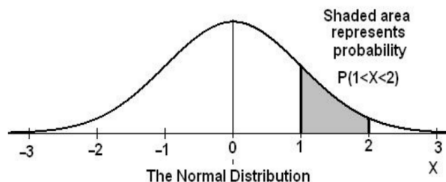
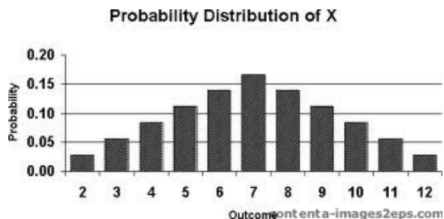
November 16, 2020



Differential Entropy - 1

- Definitions
- AEP for Continuous Random Variables
- Relation of differential entropy to discrete entropy

From discrete to continuous variables



Differential Entropy

Definition

Let X be a random variable with **cumulative distribution function** (CDF) $F(x) = \Pr(X \leq x)$. If $F(x)$ is continuous, the random variable is continuous. Let $f(x) = F'(X)$ when the derivative is defined. If $\int_{-\infty}^{+\infty} f(x) = 1$, $f(x)$ is called the **probability density function** (pdf) for X . The set of x where $f(x) > 0$ is called the **support set** of the X .

Definition

The **differential entropy** $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

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

where \mathcal{S} is the support set of the random variable.

Example: Uniform distribution

- $f(x) = \frac{1}{a}, x \in [0, a]$
- The differential entropy is:

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

- for $a < 1$, $h(X) = \log a < 0$, differential entropy can be **negative!**
(unlike discrete entropy)

Example: Uniform distribution

- $f(x) = \frac{1}{a}, x \in [0, a]$
- The differential entropy is:

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

- for $a < 1$, $h(X) = \log a < 0$, differential entropy can be **negative!**
(unlike discrete entropy)

Example: Normal distribution

- $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right)$, $x \in \mathbb{R}$
- Differential entropy:

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

Calculation:

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

Example: Normal distribution

- $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right)$, $x \in \mathbb{R}$
- Differential entropy:

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

Calculation:

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

AEP for continuous random variables

- Discrete world: for a sequence of i.i.d. random variables

$$\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X).$$

- Continuous world: for a sequence of i.i.d. random variables

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow \mathbb{E}[-\log f(X)] = h(X) \quad \text{in probability}$$

Proof follows from the weak law of large numbers.

- Discrete world: for a sequence of i.i.d. random variables

$$\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X).$$

- Continuous world: for a sequence of i.i.d. random variables

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow \mathbb{E}[-\log f(X)] = h(X) \quad \text{in probability}$$

Proof follows from the weak law of large numbers.

Typical set

- Discrete case: number of typical sequences

$$|A_\epsilon^{(n)}| \approx 2^{nH(X)}$$

- Continuous case: The **volume** of the typical set

$$\text{Vol}(A) = \int_A dx_1 dx_2 \dots dx_n, \quad A \subset \mathbb{R}^n.$$

Definition

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

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

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

Theorem

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.

Proof. 1.

Similar to the discrete case.

By definition, $-\frac{1}{n} \log f(X^n) = -\frac{1}{n} \sum \log f(X_i) \rightarrow h(X)$ in probability. □

Theorem

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.

Proof. 1.

Similar to the discrete case.

By definition, $-\frac{1}{n} \log f(X^n) = -\frac{1}{n} \sum \log f(X_i) \rightarrow h(X)$ in probability. □

Theorem

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

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

Poof. 2.

$$\begin{aligned} 1 &= \int_{\mathcal{S}^n} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)} dx_1 dx_2 \dots dx_n = 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \dots dx_n \\ &= 2^{-n(h(X)+\epsilon)} \text{Vol}(A_\epsilon^{(n)}). \end{aligned}$$



Theorem

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

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

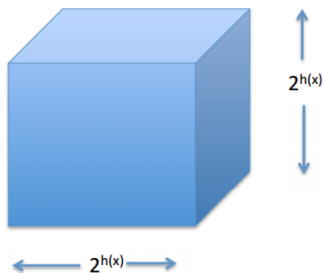
Proof. 3.

$$\begin{aligned} 1 - \epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X) - \epsilon)} dx_1 dx_2 \dots dx_n \\ &= 2^{-n(h(X) - \epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \dots dx_n \\ &= 2^{-n(h(X) - \epsilon)} \text{Vol}(A_\epsilon^{(n)}). \end{aligned}$$



An interpretation

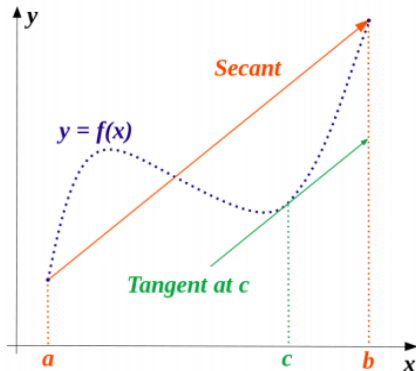
- The volume of the smallest set that contains most of the probability is approximately $2^{nh(X)}$.
- For an n -dim volume, this means that each dim has length $(2^{nh(X)})^{\frac{1}{n}} = 2^{h(X)}$.



Mean value theorem (MVT)

If a function f is continuous on the closed interval $[a, b]$, and differentiable on (a, b) , then there exists a point $c \in (a, b)$ such that

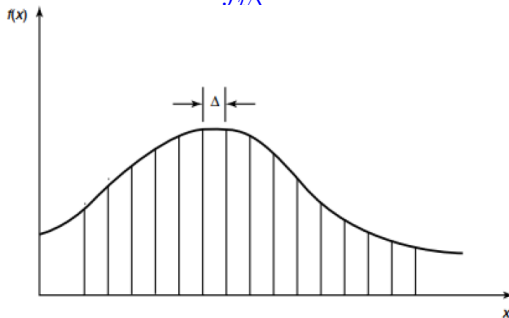
$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



Relation of differential entropy to discrete entropy

- Consider a random variable X with pdf $f(x)$. We divide the range of X into bins of length Δ .
- MVT: there exists a value $x_i \in (i\Delta, (i+1)\Delta)$ within each bin such that

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



Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^\Delta = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ with pmf

$$p_i = \Pr[X^\Delta = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

- The entropy of X^Δ is

$$H(X^\Delta) = - \sum_{-\infty}^{+\infty} p_i \log p_i = - \sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is Riemann integrable, as $\Delta \rightarrow 0$,

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X)$$



Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^\Delta = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ with pmf

$$p_i = \Pr[X^\Delta = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

- The entropy of X^Δ is

$$H(X^\Delta) = - \sum_{-\infty}^{+\infty} p_i \log p_i = - \sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is Riemann integrable, as $\Delta \rightarrow 0$,

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X)$$



Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^\Delta = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ with pmf

$$p_i = \Pr[X^\Delta = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

- The entropy of X^Δ is

$$H(X^\Delta) = - \sum_{-\infty}^{+\infty} p_i \log p_i = - \sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is Riemann integrable, as $\Delta \rightarrow 0$,

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X)$$

Reading & Homework

- **Reading:** Chapter 8: 8.1 - 8.3
- **Homework:** Problems 8.1, 8.5, 8.7