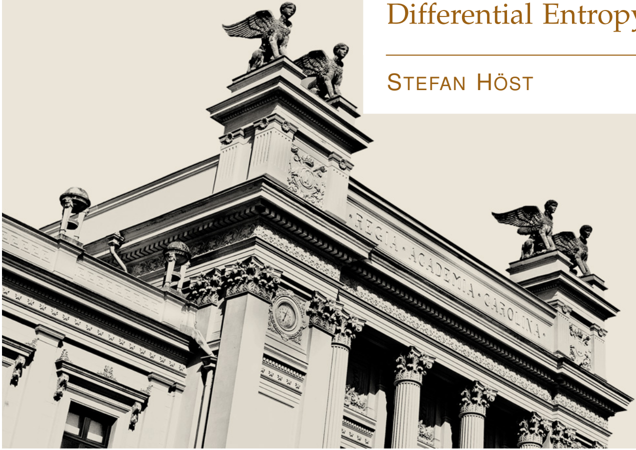
LUND UNIVERSITY

Information Theory

Lecture 10
Differential Entropy

STEFAN HÖST



Differential Entropy

Definition

Let X be a real valued continuous random variable with probability density function $f(x)$. The differential entropy is

$$H(X) = E\big[-\log f(X)\big] = -\int_{\mathbb{R}} f(x)\log f(x)dx$$

where it is used that $0\log 0 = 0$.

Sometimes the notation $H(f)$ is also used in the literature.

Interpretation

Differential entropy can not be interpreted as *uncertainty*

Differential entropy

Example

For a uniform distribution, $f(x) = \frac{1}{a}$, $0 \leq x \leq a$, the differential entropy is

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

Note that $H(X) < 0$, $0 < a < 1$.

Example

For a Gaussian (Normal) distribution, $\mathcal{N}(\mu, \sigma)$ the differential entropy is

$$H(X) = -\int_{\mathbb{R}} f(x) \log f(x) dx = \frac{1}{2} \log(2\pi e \sigma^2)$$

Translation and scaling

Theorem

- *Let* $Y = X + c$*, then* $f_Y(y) = f_X(y - c)$ *and*

$$H(Y) = H(X)$$

- *Let* $Y = \alpha X$*, then* $f_Y(y) = \frac{1}{\alpha} f_X(\frac{y}{\alpha})$ *and*

$$H(Y) = H(X) + \log \alpha$$

Gaussian distribution

Lemma

Let $g(x)$ be the distribution function of $X \sim \mathcal{N}(\mu, \sigma)$. Let $f(x)$ be an arbitrary distribution function with the same mean, μ, and variance, σ^2. Then

$$\int_{\mathbb{R}} f(x) \log g(x) dx = \int_{\mathbb{R}} g(x) \log g(x) dx$$

Differential entropy

Definition

The joint differential entropy is the entropy for a 2-dimensional random variables (X, Y) with the joint density function $f(x, y)$,

$$H(X, Y) = E\big[-\log f(X, Y)\big] = -\int_{\mathbb{R}^2} f(x, y) \log f(x, y) dxdy$$

Definition

The multi-dimensional joint differential entropy is the entropy for an n-dimensional random vector $(X_1, \ldots, X_n)$ with the joint density function $f(x_1, \ldots, x_n)$,

$$H(X_1, \ldots, X_n) = -\int_{\mathbb{R}^n} f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1, \ldots, dx_n$$

Mutual information

Definition

The mutual information for a pair of continuous random variables with joint probability density function $f(x, y)$ is

$$I(X;Y) = E\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right] = \int_{\mathbb{R}^2} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$$

The mutual information can be derived as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned}$$

where

$$H(X|Y) = -\int_{\mathbb{R}^2} f(x,y) \log f(x|y) dxdy$$

Relative entropy

Definition

The relative entropy for a pair of continuous random variables with probability density functions $f(x)$ and $g(x)$ is

$$D(f||g) = E_f\left[\log \frac{f(X)}{g(X)}\right] = \int_{\mathbb{R}} f(x) \log \frac{f(x)}{g(x)} dx$$

Theorem

The relative entropy is non-negative,

$$D(f||g) \geq 0$$

with equality if and only if $f(x) = g(x)$, for all x.

Mutual information

Corollary

The mutual information is non-negative, i.e.

$$I(X;Y) = H(X) - H(X|Y) \geq 0$$

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

Corollary

The entropy will not increase by considering side information, i.e.

$$H(X|Y) \leq H(X)$$

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

Chain rule

Theorem

The chain rule for differential entropy states that

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

Corollary

From the chain rule it follows

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Gaussian distribution

Theorem

The Gaussian distribution maximises the differential entropy over all distributions with mean μ and variance σ^2.

Continuous vs Discrete

Theorem

Let X be continuous r.v. with density $f_X(x)$. Construct a discrete version X^Δ, where $p(x_k^\Delta) = \int_{k\Delta}^{(k+1)\Delta} f_X(x)dx = \Delta f_X(x_k)$

Then, in general, $\lim_{\Delta\to 0} H(X^\Delta)$ does not exist.

Theorem

Let X and Y be continuous r.v. with density $f_X(x)$ and $f_Y(y)$. Construct discrete versions X^Δ and Y^δ, where $p(x_k^\Delta) = \Delta f(x_k)$, $f(x_k) = \sum_\ell \delta f(x_k, y_\ell)$, and $p(y_\ell^\delta) = \delta f(y_\ell)$, $f(y_\ell) = \sum_k \Delta f(x_k, y_\ell)$. Then

$$\lim_{\Delta,\delta\to 0} I(X^\Delta, Y^\delta) = I(X;Y)$$