

ECE 771

Lecture 9 – Differential entropy

Objective: Differential entropy is entropy defined for distributions with a continuous random variable. We will learn about differential entropy.

Reading:

1. Read chapter 9.

Differential entropy

A **continuous random variable** (for the purpose of this class) is one in which the distribution function $F(x)$ is continuous: there are no jumps (discrete outputs).

Definition 1 The **differential entropy** $h(X)$ of a continuous random variable X with density $f(x)$ and support S is

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

□

Example 1 Let $X \sim \mathcal{U}(0, a)$. (Uniform). Then

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

Note that the differential entropy can be negative (if $a < 1$). This is why we refer to this as *differential* entropy, since entropy should always be positive. □

Example 2 Normal: $X \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} = \phi(x)$. (We have zero mean, but that will make no difference on the entropy.) We will compute the entropy in nats.

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

□

Having defined the differential entropy, we can now go through and define all the sorts of things we did before.

AEP

Theorem 1 Let X_1, \dots, X_n be a sequence of random variables drawn *i.i.d.* according to $f(x)$. Then

$$-\frac{1}{n} \log f(X_1, \dots, X_n) \rightarrow E[-\log f(X)] = h(X),$$

convergence in probability.

Proof This is just the WLLN. \square

Definition 2 Typical sets For any $\epsilon > 0$ and any n , the **typical set** $A_\epsilon^{(n)}$ with respect to $f(x)$ is

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

\square

That is, it is the set for which the empirical differential entropy is close to the differential entropy.

For discrete random variables, we talked about the number of elements in the typical set. For continuous random variables, the analogous concept is the **volume** of the typical set.

Definition 3 The **volume** of a set $A \in \mathbb{R}^n$ is

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

\square

Theorem 2 *The typical set has the following properties:*

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large. (Typical sets occur most of the time.)
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. Just WLLN again.
2. Note that

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

3. If n is sufficiently large that property 1 is true, then

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

\square

Discretization

When a r.v. X with continuous distribution is broken into a range of bins, then (mean value theorem) there is a value x_i in each range s.t.

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

Let X^Δ be the quantized r.v. defined by

$$X^\Delta = x_i \text{ with probability } p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

The entropy of the quantized r.v. is

$$H(X^\Delta) = - \sum f(x_i)\Delta \log(f(x_i)\Delta) = - \sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

In the limit as $\Delta \rightarrow 0$ then

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

Thus the entropy of a n -bit quantization of a continuous r.v. increases with n .

Joint, conditional, and relative differential entropy

More basic definitions:

$$h(X_1, \dots, X_n) = - \int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) dx_1 \cdots x_n$$

is the **joint differential entropy**.

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

is the **conditional differential entropy**.

An important special case is the following:

Theorem 3 Let X_1, \dots, X_n have a multivariate normal distribution with mean μ and covariance K . Then

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

Proof For convenience, assume $\mu = 0$; it won't make any difference.

$$\begin{aligned} h(f) &= - \int f(\mathbf{x}) \left[-\frac{1}{2} \mathbf{x}^T K^{-1} \mathbf{x} - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}} \right] d\mathbf{x} \\ &= \frac{1}{2} E \left[\sum_{i,j} x_i (K^{-1})_{ij} x_j \right] + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_{i,j} E[x_j x_i] (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_j \sum_i K_{ji} (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_j (K K^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} n + \frac{1}{2} \ln(2\pi)^n |K| \\ &= \frac{1}{2} \ln(2\pi e)^n |K| \text{ nats} \\ &= \frac{1}{2} \log(2\pi e)^n |K| \text{ bits.} \end{aligned}$$

□

The **relative** entropy is

$$D(f\|g) = \int f \log \frac{f}{g}.$$

The **mutual information** is

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

Some properties:

1. $D(f\|g) \geq 0$.
2. $I(X; Y) \geq 0$.
3. $h(X|Y) \leq h(X)$.
4. Chain rule:

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

5. $h(X + c) = h(X)$. (Shifting does not affect the entropy.)
6. $h(aX) = h(X) + \log |a|$. To prove this, let $Y = aX$, then $f_Y(y) = 1/|a|f_X(y/a)$.
7. $h(\mathbf{A}\mathbf{X}) = h(\mathbf{X}) + \log |\mathbf{A}|$.

An important result is the following:

Theorem 4 Let the random vector \mathbf{X} have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^T$. Then $h(\mathbf{X}) \leq \frac{1}{2} \log(2\pi e)^n |K|$, with equality iff $X \sim \mathcal{N}(0, K)$.

That is, for a given covariance, the normal (Gaussian) distribution has the one which maximizes the entropy.

Proof Let $g(\mathbf{X})$ be a distribution with the same covariance, and let ϕ denote the Gaussian density.

$$\begin{aligned} 0 &\leq D(g\|\phi) \\ &= \int g \log(g/\phi) \\ &= -h(g) - \int g \log \phi \\ &= -h(g) - \int \phi \log \phi \\ &= -h(g) + h(\phi). \end{aligned}$$

The key step is observing that

$$\int g \log \phi = a \int g(\mathbf{x}^T K^{-1} \mathbf{x})$$

and that both g and ϕ have the same second moments. □