Information Theory with Applications, Math6397
Lecture Notes from October 28, 2014

taken by Nicole Leonhard

Recall from last time we discussed:

- Continuous Entropy
- Differential Entropy
- Quantization/Rounding
- Gaussian Maximizers of Differential Entropy

Warm up:

5.1.6 Question. What number of bits is needed to store the decay line T of an atom with half life $\lambda = 80$ years at an accuracy $\triangle = 10^{-3}$(years)?

5.1.7 Answer. Suppose you have an experiment with half life of 80 years. Now, suppose you want to store decays at specific times. How large does your hard drive need to be to store the data? For a discrete random variable, say $T^{\triangle}$, the rounded one, we know the expected number of bits needed to store the outcomes is $H_2\left(T^{\triangle}\right)$. Note, we are using base 2. Last time we compared the rounded entropy to the differential entropy for base 2 (binary) entropies. Repeating the argument from last time gives us $H_2\left(T^{\triangle}\right) \approx h_2\left(T\right) - \log_2 \triangle$.

Here, $\triangle \approx \frac{1}{2^{10}}$, so $\log_2 \triangle = -10$. Now, compute the differential entropy. For decay processes, we have $p\left(x\right) = \gamma e^{-\gamma t}$ where $\gamma =$ decay constant and $p\left(x\right)$ must integrate to one.

When $\lambda = 80$ years $\int_0^{80} \gamma e^{-\gamma t} dt = 0.5$. So, $\gamma = 0.0086$. Thus,

$$H_2\left(T^{\triangle}\right) \approx -\int_0^{80} \gamma e^{-\gamma t} \log_2\left(\gamma e^{-\gamma t}\right) dt - -10 = \log_2\left(\frac{e}{\gamma}\right).$$

Then, plugging in γ, $H_2\left(T^{\triangle}\right) = 18.29$ bits.

Since delta is about 8 hours not much space needed. To make this relevant we would need a sequence of Ts and look at the asymptotic behavior.

5.2 Properties of Differential Entropy

Now, we will look at properties of differential entropy. We adopt the notation of the discrete case, but will use lower case letters. Also, when we talk about a joint distribution inducing a measure with density associated with two, or more, random variable we are looking at the density with respect to a product measure. Since we know the joint density is Lebesgue integrable with respect to that product measure by looking at one variable and ignoring the other variables we see that the densities of each random variable, marginals, also exist. However, since we don't have guaranteed integrability of $w(x,y)$ even with the natural log function we will need to add "whenever the integral exists" to our statements.

5.2.8 Definition. Given two random variables X and Y with a continuous joint distribution with density $p_{X,Y}$ on $\mathbb{R}^2$ then the conditional differential entropy is given by

$$h(X|Y) = \int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln w(x|y)\, dxdy$$

where $w(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$, whenever the integral exists.

5.2.9 Lemma. *(Conditioning Decreases Entropy): Let X and Y be random variables inducing a measure with a density $p_{X,Y}$ on $\mathbb{R}^2$, then $h(X|Y) \leq h(X)$ and equality holds if and only if X and Y are independent.*

Proof. Since the density of the joint distribution, is assumed to be Lebesgue integrable $p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x,y)\, dx$ exists Lebesgue almost everywhere $y \in \mathbb{R}^2$ and in that case, $w(x|y) = \frac{p_{X,Y}(x,y)}{\int_{\mathbb{R}} p_Y(x\prime,y)dx\prime}$, is a probability density with $p_Y(y)\, w(x|y) = p_{X,Y}(x,y)$.

Next,

$$\begin{aligned}
h(X) - h(X|Y) &= -\int_{\mathbb{R}} \underbrace{p_X(x)}_{\text{By inserting}= \underbrace{p_{X,Y}(x,y)}_{=p_Y(y)w(x|y)a.e}} \ln p_X(x)\, dx + \int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln w(x|y)\, dxdy \\
&= -\int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln p_X(x)\, dx + \int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln w(x|y)\, dxdy \\
&= -\int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln \left(\frac{p_X(x)}{w(x|y)} \right) dxdy \quad \text{(Combining the integrals.)} \\
&= -\int_{\mathbb{R}} p_Y(y) \int_{\mathbb{R}} w(x|y) \ln \left(\frac{p_X(x)\, p_Y(y)}{p_{X,Y}(x,y)} \right) dxdy \quad \text{(Using the product of the marginal.)} \\
&= -\int_{\mathbb{R}^2} p_{X,Y}(x,y) \ln \left(\frac{p_X(x)\, p_Y(y)}{p_{X,Y}(x,y)} \right) dxdy \quad \text{(Replace with the joint density)} \\
&\geq -ln \left(\int_{\mathbb{R}^2} p_{X,Y}(x,y) \frac{p_X(x)\, p_Y(y)}{p_{X,Y}(x,y)} dxdy \right) \quad \text{(Jensen's Inequalitly, since -ln is convex.)} \\
&= -\ln 1 = 0
\end{aligned}$$

In Jensen's inequality equality holds if and only if the inside of the function is constant on the support of the function. Therefore, there is a $c \in \mathbb{R}^2$ such that $\frac{p_X(x)p_Y(y)}{p_{X,Y}(x,y)} = c$ almost everywhere on the support of $p_{X,Y}$. Then, by normalization we see that $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$ almost everywhere. So, X and Y are independent. □

5.2.10 Lemma. *(Additivity): Let $X_1, ..., X_n$ be random variables with an induced measure on $\mathbb{R}^n$ that has density $p_{X_1,...,X_n}$, then*

$$h(X_1, ..., X_n) = \sum_{j=1}^{n} h(X_j | X_1, ..., X_{j-1})$$

and

$$h(X_1, ..., X_n) \leq \sum_{j=1}^{n} h(X_j),$$

whenever the right hand side exists.

We will only address the case of two random variables, the n variable case follows by induction.

Proof. Let X_1 and X_2 be random variables. Then,

$$\begin{aligned} h(X_1, X_2) &= -\int_{\mathbb{R}_2} p_{X_1,X_2}(x_1, x_2) \ln \underbrace{p_{X_1,X_2}(x_1, x_2)}_{=p_{X_1}(x_1)w(x_2|x_1)} dx_1 dx_2 \\ &= \underbrace{-\int_{\mathbb{R}^2} p_{X_1,X_2}(x_1, x_2) \ln(p_{X_1}(x_1))\, dx_1 dx_2 - \int_{\mathbb{R}_2} p_{X_1,X_2}(x_1, x_2) \ln(w(x_2|x_1))\, dx_1 dx_2}_{\text{These can be thought of iterative integrals.}} \\ &= h(X_1) + h(X_2|X_1) \end{aligned}$$

The inequality follows from the fact that conditioning decreases entropy. So, we have

$$h(X_1, X_2) = h(X_1) + h(X_2|X_1) \leq h(X_1) + h(X_2).$$

□

So far we have looked at the properties that have been the same for the discrete and continuous case, as long as the integrals exist. There are some properties differential entropy that differ from the discrete case. For Differential Entropy you can get an idea of what should occur by looking at the rounded variable. However, this is not always the case. In the discrete case, if you create a one to one map to the values change but the probabilities don't. For differential entropy this can't be done because the densities change. That is if $f : \mathbb{A} \to \mathbb{B}$ is one to one and $\mathbb{A}$ and $\mathbb{B}$ are discrete alphabets then $H(f(x)) = H(x)$. Now, we will examine an example of how differential entropy acts differently.

5.2.11 Lemma. *Let $a \neq 0$ and $f : \mathbb{R} \to \mathbb{R}$, $x \to ax$ then for random variable X that induces measure with density p $h(f(x)) = h(x) + \ln|a|$.*

Proof. The first thing we have to address is the density of $f(x)$. The image measure under f has density p_f so that $\int_{\mathbb{B}} p_f(y)\,dy = \int_{f^{-1}(\mathbb{B})} p(x)\,dx$. Assume that $a > 0$ and change variables by defining $x = \frac{y}{a}$. So, we have

$$\int_{\mathbb{B}} p_f(y)\,dy = \int_{\underbrace{\frac{y}{a} \in f^{-1}(\mathbb{B})}_{y \in \mathbb{B}}} p\left(\frac{y}{a}\right) \underbrace{d\left(\frac{y}{a}\right)}_{\frac{1}{a}dy}$$

so, $p_f(y) = \frac{1}{a} p\left(\frac{y}{a}\right)$ almost everywhere with respect to y. Thus,

$$\begin{aligned}
h(f(x)) &= -\int_{\mathbb{R}} p_f(y) \ln p_f(y)\,dy \\
\text{(By plugging in our defintion.)} &= -\int_{\mathbb{R}} \frac{1}{a} p\left(\frac{y}{a}\right) \ln \frac{1}{a} p\left(\frac{y}{a}\right) dy \\
\left(\text{Change variables using } x = \frac{y}{a}\right) &= -\int_{\mathbb{R}} p(x) \ln \frac{p(x)}{a} dx \\
&= -\int_{\mathbb{R}} p(x) \ln \frac{p(x)}{a} dx - \ln\left(\frac{1}{a}\right) \int_{\mathbb{R}} p(x)\,dx \\
&= h(x) + \ln(a).
\end{aligned}$$

In the case of $a < 0$, the proof follows similarly by using $x = \frac{1}{|a|}$ □

Notice, small values of a, which compresses, will have a negative contribution, thus yield smaller entropy.

5.2.12 Corollary. *If $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear and invertible, and random variables$X_1, ..., X_n$ which induce a density on $\mathbb{R}^n$, then $h(AX) = h(X) + \ln|A|$, where X is a vector of random variables $X_1, ..., X_n$ and $|A|$ is the Jacobian. On the other hand, if we transform by $f(X) = X + C$ then $h(X + C) = h(X)$.*

5.2.13 Corollary. *If X and Y are random variables which induce a measure with density on $\mathbb{R}^n$ then $h(X|Y) = h(X + Y|Y)$.*

Proof. Let X and Y be random variables. Given a matrix in block form $A = \begin{pmatrix} I & I \\ 0 & I \end{pmatrix}$, where I is the identity matrix, $A\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X + Y \\ Y \end{pmatrix}$. Thus

$$\begin{aligned}
h(X + Y|Y) &= -h(Y) + h(X + Y|Y) \\
&= -h(Y) + h\left(A\begin{pmatrix} X \\ Y \end{pmatrix}\right) \\
\text{(By Cor. 5.2.12)} &= -h(Y) + h(X, Y) + \ln \underbrace{|A|}_{detA=1} \\
\text{(By additivity.)} &= h(X|Y)
\end{aligned}$$

□

This map A can be thought of like a shearing operation, which is a measure preserving change of coordinates.

5.3 Shannon-McMillan Asymptotic Equipartitioning for Continuous Sources

This will be similar to the Discrete Memoryless Source version covered previously. However, in the Shannon-McMillan-Breiman Theorem for a Discrete Memoryless Source we had a bound on the size of typical sets. With a Continuous Memoryless Source we don't have size. In this case since everything is absolutely continuous with respect to the Lebesgue measure the size of a typical set will be replaced with volume.

5.3.14 Definition. A continuous memoryless source (CMS) $\{X_j\}_{j=1}^{\infty}$ is a sequence of identical independent distribution random variables with values in $\mathbb{R}$ and X_1 induces a measure which is absolutely continuous with respect to the Lebesgue measure.

Note, that the induced measure will have a probability density on $\mathbb{R}$.