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

taken by Kedar Grama

1.1.3 Conditional Entropy

1.1.5 Definition. The conditional entropy of Y given X is obtained by averaging $H(Y|a)$ over all $a \in \mathbb{A}$, with respect to the probabilities $\mathbb{P}_X(a)$:

$$\begin{aligned} H(Y|X) &\equiv \sum_{a\in\mathbb{A}} \mathbb{P}_X(a) H(Y|a) \\ &= \sum_{a\in\mathbb{A}} \mathbb{P}_X(a) \left(-\sum_{b\in\mathbb{B}} \mathbb{W}(b|a) \ \ln \mathbb{W}(b|a) \right) \\ H(Y|X) &= -\sum_{a\in\mathbb{A}, b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b) \ \ln \mathbb{W}(b|a) \end{aligned}$$

A few properties of conditional entropy are listed below:

- $H(Y|X) = 0$ implies that either $\mathbb{P}_{X,Y}(a,b) = 0$ or $\ln \mathbb{W}(b|a) = 0$. So, if $\mathbb{P}_{X,Y}(a,b) \neq 0$ then $\mathbb{W}(b|a) = 1$. Thus there is a map $f : \mathbb{A} \to \mathbb{B}$ such that $Y = f(X)$ with probability one. So, Y almost surely depends on X in a deterministic fashion.
- $H(Y|X) \leq H(Y)$ This property implies that knowing X decreases the entropy of Y. We would like to show this property and we prepare this with lemma below.

Before we begin a reminder of two properties:
Independence: If we have X, Y with alphabets $\mathbb{A}, \mathbb{B}$ and joint probability measure $\mathbb{P}_{X,Y}$ then X and Y are independent if and only if $\mathbb{P}_{X,Y}(a,b) = \mathbb{P}_(a)\mathbb{P}_Y(b)$ for all $(a,b) \in (\mathbb{A} \times \mathbb{B})$
Convexity of $x \mapsto x \ln x$: From basics of calculus we know that a function $f(x) \in C^2(\mathbb{R})$ is convex if its second derivative is non-negative. For $f(x) = x \ln x$, this implies $f'(x) = 1 + \ln x$ and $f''(x) = \frac{1}{x} > 0,$ for all $x > 0$ hence the function is convex in $x > 0$.

1.1.6 Lemma (Log-Sum Inequality)**.** *For any non-negative $a_1, a_2, \ldots, a_n$ and strictly positive $b_1, b_2, \ldots, b_n$ we have*

$$\sum_{j=1}^{n} a_j \ln \frac{a_j}{b_j} \geq \left(\sum_{j=1}^{n} a_j \right) \ \ln \left(\frac{\left(\sum_{j=1}^{n} a_j \right)}{\left(\sum_{j=1}^{n} b_j \right)} \right)$$

and equality holds if and only if for all $1 \leq j \leq n$, $\frac{a_j}{b_j} = \frac{a_1}{b_1}$

Proof. Assume $a'_j \geq 0$, $\sum_{j=1}^{n} a'_j = 1$ and f is strictly convex, then, Jensen's inequality gives $\sum_{j=1}^{n} a'_j f(x_j) \geq f\left(\sum_{j=1}^{n} a'_j x_j\right)$ and equality holds if and only if $x'_j s$ are constant for each j where $a'_j > 0$.

The log-sum inequality follows from choosing $a'_j = \frac{b_j}{\sum_{l=i}^{n} b_l}$, $x_j = \frac{a_j}{b_j}$ and $f(x) = x \ln x$, because then

$$\sum_{j=1}^{n} \frac{b_j}{\sum_{l=1}^{n} b_l} \frac{a_j}{b_j} \; \ln \frac{a_j}{b_j} \geq \sum_{j=1}^{n} \frac{a_j}{\sum_{l=1}^{n} b_l} \; \ln \left(\sum_{k=1}^{n} \frac{a_k}{\sum_{l=1}^{n} b_l} \right)$$

$$\sum_{j=1}^{n} a_j \; \ln \frac{a_j}{b_j} \geq \sum_{j=1}^{n} a_j \; \ln \left(\sum_{k=1}^{n} \frac{a_k}{\sum_{l=1}^{n} b_l} \right)$$

Hence we get the asserted inequality. □

We use this lemma to show that knowing X helps reduce uncertainity about Y.

1.1.7 Proposition. *Given two discrete random variables $X : \Omega \to \mathbb{A}$ and $Y : \Omega \to \mathbb{B}$, then $H(X|Y) \leq H(Y)$ and equality holds if and only if X and Y are indpendent.*

Proof. Using the definition of entropy and conditional entropy:

$$\begin{aligned} H(Y) - H(Y|X) &= \underbrace{\sum_{b \in \mathbb{B}} \mathbb{P}_Y(b) \; \ln \frac{1}{\mathbb{P}_Y(b)}}_{\sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b) \; \ln \frac{1}{\mathbb{P}_Y(b)}} + \sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b) \; \ln \mathbb{W}(b|a) \\ &= \sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b) \; \ln \frac{\mathbb{W}(b|c)}{\mathbb{P}_Y(b)} \\ &= \sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b) \; \ln \frac{\mathbb{P}_{X,Y}(a,b)}{\mathbb{P}_X(a)\mathbb{P}_Y(b)} \\ (\text{Using LogSum inequality}) \quad &\geq \left(\sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b) \right) \; \underbrace{\ln \frac{\left(\sum_{(a,b)\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_{X,Y}(a,b)\right)}{\left(\sum_{(a',b')\in(\mathbb{A}\times\mathbb{B})} \mathbb{P}_X(a)\mathbb{P}_Y(b)\right)}}_{=\ln 1} \quad = 0 \end{aligned}$$

Because the probabilities $\mathbb{P}_{X,Y}(a,b)$ are non-negative, equality holds if and only if $\frac{\mathbb{P}_{X,Y}(a,b)}{\mathbb{P}_X(a)\mathbb{P}_Y(b)} = 1$ when $\mathbb{P}_{X,Y}(a,b) \neq 0$. This means $\mathbb{P}_{X,Y}(a,b) = \mathbb{P}_X(a)\mathbb{P}_Y(b)$ for such (a,b). Using the fact that the probabilities of all the outcomes have to sum to one, we see that this inequality holds for all $(a,b) \in \mathbb{A} \times \mathbb{B}$, which means that the random variables are independent. □

1.1.8 Question. We have the inequality for the conditional entropy $H(Y|X) \leq H(Y)$, but what about the entropy of the conditional probability measure of Y given that $X = a$, $H(Y|a) = -\sum_{b\in\mathbb{B}} \mathbb{W}(b|a) \ln \mathbb{W}(b|a)$?

In general we cannot compare $H(Y)$ and $H(Y|a)$. For example, consider the pair of binary random variables X,Y.

$\mathbb{P}_{X,Y}$	Y=0	Y=1
$X=0$	0.8	0
$X=1$	0.1	0.1

We compute:

$$\begin{aligned} H(Y) &= -0.9\ln 0.9 - 0.1\ln 0.1 \approx 0.33 \\ H(Y|X=1) &= -\ln 0.5 \approx 0.69 \\ H(Y|X) &= 0.2 \times (-\ln 0.5) + 0.8 \times 0 \approx 0.14 \end{aligned}$$

We see that knowing the value $X = 1$ occurred does not necessarily decrease the entropy resulting for the distribution of Y.

1.2 Additivity of Entropy

1.2.9 Proposition. *Let $X : \Omega \to \mathbb{A}$, $Y : \Omega \to \mathbb{B}$ as above, then*

$$H(X,Y) = H(X) + H(Y|X)$$

Proof. Using the definition of entropy of joint distributions:

$$\begin{aligned} H(X,Y) &= -\sum_{a\in\mathbb{A},b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln \mathbb{P}_{X,Y}(a,b) \\ &= -\sum_{a\in\mathbb{A},b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\left(\ln\frac{\mathbb{P}_{X,Y}(a,b)}{\mathbb{P}_X(a)} + \ln \mathbb{P}_X(a)\right) \\ &= -\sum_{a\in\mathbb{A}} \underbrace{\sum_{b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b)}_{\mathbb{P}_X(a)} \ln \mathbb{P}_X(a) - \sum_{a\in\mathbb{A},b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\frac{\mathbb{P}_{X,Y}(a,b)}{\mathbb{P}_X(a)} \\ &= -\sum_{a\in\mathbb{A}} \mathbb{P}_X(a)\ln \mathbb{P}_X(a) - \sum_{a\in\mathbb{A},b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln \mathbb{W}(b|a) \\ &= H(X) + H(Y|X) \end{aligned}$$

□

1.2.10 Corollary. *Let $\{X_j\}_{j=1}^n$ be random variables with discrete alphabets, then*

$$H(X_1,\dots,X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1,\dots,X_n)$$

The proof of the corollary is done by induction over the number of random variables.

1.3 Concavity of Entropy

1.3.11 Proposition. *Given discrete random variables X, Y then*

i $H(\mathbb{P}_X)$ *is concave in* $\mathbb{P}_X$.

ii $H(\mathbb{P}_{X,Y})$ *is concave in the probability measure of* X*, that is, for* $\lambda \in [0,1]$

$$H(\lambda \mathbb{P}_{X_1,Y} + (1-\lambda)\mathbb{P}_{X_2,Y}) \geq \lambda H(P_{X_1,Y}) + (1-\lambda) H(\mathbb{P}_{X_2,Y}) \,.$$

iii $H(Y|X)$ *is concave with respect to* $\mathbb{W}(b|a)$.

Proof. .

i It suffices to show that for $\lambda \in [0,1]$ and the random Variables X_1, X_2 with probability measures $\mathbb{P}_{X_1}$ and $\mathbb{P}_{X_2}$, that $H(\lambda \mathbb{P}_{X_1} + (1-\lambda)\mathbb{P}_{X_2} \geq \lambda H(\mathbb{P}_{X_1}) + (1-\lambda) H(\mathbb{P}_{X_2})$

With Jensen's inequality and since $x \ln x$ is strictly convex we have:

$$\lambda \mathbb{P}_{X_1} \;\ln \mathbb{P}_{X_1} + (1-\lambda)\mathbb{P}_{X_2} \;\ln \mathbb{P}_{X_2} \leq (\lambda \mathbb{P}_{X_1} + (1-\lambda)\mathbb{P}_{X_2}) \;\ln(\lambda \mathbb{P}_{X_1} + (1-\lambda)\mathbb{P}_{X_2})$$

Multiplying the above by -1 and substituting the definition of entropy, we get the required inequality.

ii This is similar to above and can be obtained by replacing $\mathbb{P}_{X_1}$ by $\mathbb{P}_{X_1,Y}$ and $\mathbb{P}_{X_2}$ by $\mathbb{P}_{X_2,Y}$.

iii Again, here we use the definition of conditional entropy:

$$H(Y|X) = \sum_{a \in \mathbb{A}} \mathbb{P}_X(a) \sum_{b \in \mathbb{B}} (-\mathbb{W}(b|a) \;\ln \mathbb{W}(b|a))$$

and use the same program for the sum over the outcomes $b \in \mathbb{B}$ as in i to show the concavity.

□