Joint Entropy and Conditional Entropy

Consider a vector valued variable (X,Y), The <u>joint</u> entropy $H(X,Y)$ is

$$H(X,Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x,y)$$

$$H(X,Y) = -E\{\log p(X,Y)\}$$

The conditional entropy is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x) \end{aligned}$$

Chain rule :

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

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x,y) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x)p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

Corollary $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$ - show!

Example

Y \ X	1	2	3	4	P(Y)
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
P(X)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$H(X) = \frac{7}{4}$ Sh bits

$H(Y) = 2$ Sh bits

$$H(X|Y) = \sum_{i=1}^{4} P(Y=i)\, H(X|Y=i)$$

$$H(X|Y) = \sum_{x \in Y} \sum_{y \in Y} p(x,y) \log p(x|y)$$

$$= \frac{11}{8} \text{ bits}$$

$$P(X=1|Y=1)\, P(Y=1) = P(X=1, Y=1)$$

$$P(X=1|Y=1) = \frac{P(x=1Y=1)}{P(Y=1)} = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$$

Mutual Information

Consider two random variables X and Y with a joint probability distribution $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$

The mutual information $I(X;Y)$ is relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

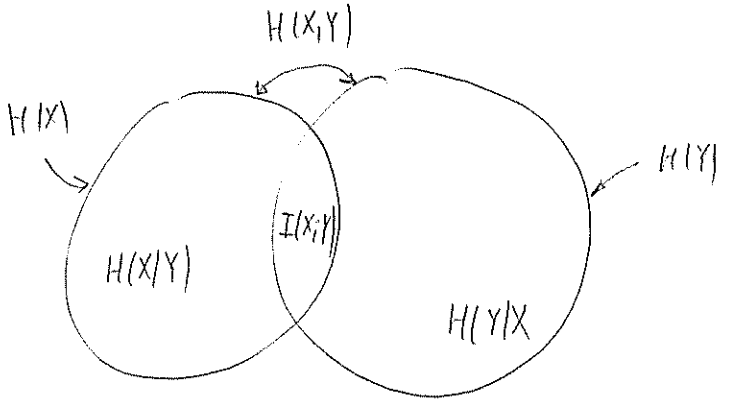
We can show that $I(X;Y) = H(X) - H(X|Y)$

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y) \\ &= -\sum_{x \in X} p(x) \log p(x) - \left(-\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y) \right) \\ &= H(X) - H(X|Y) \end{aligned}$$

- The mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.
- It follows

$$I(X;Y) = H(Y) - H(Y|X)$$

$I(X;X) = H(X)$ — $H(X)$ is referred to as self information



$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Some interesting properties.

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_n|X_1, X_2, \ldots X_{n-1})$$

$$I(X_1, X_2, \ldots, X_n ; Y) = I(X_1;Y) + I(X_2;Y|X_1) + I(X_3;Y|X_1, X_2) + \ldots I(X_n;Y|X_1, \ldots X_{n-1})$$

$$I(X;Y) \geq 0$$

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

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

Data processing inequality

If X,Y,Z form a Markov chain ($p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$) then

$$I(X;Y) \geq I(X;Z)$$

$$\overset{\geq}{} I(X;Y,Z) = I(X;Z) + \overbrace{I(X;Y|Z)}^{\geq 0}$$

or $= I(X;Y) + I(X;Z|Y)$ — since X and Z are conditionally independent given Y $I(X;Z|Y) = 0$

$I(X;Y) \geq I(X;Z)$ ← - - — since $I(X;Y|Z) \geq 0$

Jensen's Inequality

Def: — A real-valued function f is concave on an interval I if

$$f\left(\frac{x+y}{2}\right) \geq \frac{f(x)+f(y)}{2}$$

for all $x, y \in I$. It is strictly concave if $f\left(\frac{x+y}{2}\right) > \frac{f(x)+f(y)}{2}$

Thm — Suppose f is a continuous and strictly concave on I, and let $a_i > 0$, $1 \leq i \leq n$, such that $\sum_{i=1}^{n} a_i = 1$. Then

$$\sum_{i=1}^{n} a_i f(x_i) \leq f\left(\sum_{i=1}^{n} a_i x_i\right)$$

where $x_i \in I$. Equality occurs iff $x_1 = x_2 = \dots = x_n$

- Ex: Concave functions: $\log x$, $\sqrt{x}$
  Convex: x^2, e^x, $|x|$, $x \log x$

Proof by induction

$$a_1 f(x_1) + a_2 f(x_2) \leq f(a_1 x_1 + a_2 x_2)$$ — true by def. of concave function

$$\sum_{i=1}^{n} a_i f(x_i) = a_n f(x_n) + \sum_{i=1}^{n-1} a_i f(x_i)$$

$$= a_n f(x_n) + (1-a_n) \cdot \sum_{i=1}^{n-1} \frac{a_i}{1-a_n} f(x_i)$$

$$\leq a_n f(x_n) + (1-a_n) \cdot f\left(\sum_{i=1}^{n} \frac{a_i}{1-a_n} \cdot x_i\right)$$

$$\leq f\left(a_n x_n + (1-a_n) \sum_{i=1}^{n} \frac{a_i}{1-a_n} x_i\right)$$

$$= f\left(\sum_{i=1}^{n} a_i x_i\right)$$

Thm: $H(X,Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

$$H(X)+H(Y) = -\sum_x p(x) \log p(x) - \sum_y p(y) \log p(y)$$

$$= -\sum_x \sum_y p(x,y) \log p(x) - \sum_y \sum_x p(x,y) \log p(y)$$

$$= -\sum_x \sum_y p(x,y) \log p(x) p(y)$$

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

$$H(X,Y) - H(X) - H(Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)}$$

$$+ \sum_x \sum_y p(x,y) \log p(x) p(y)$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y)}{p(x,y)}$$ log – strictly convex

$$\leq \log \Big(\sum_x \sum_y \cancel{p(x,y)} \frac{p(x)p(y)}{\cancel{p(x,y)}}$$ – Jensen's inequality

$$= \log 1$$

equality for $\frac{p(x)p(y)}{p(x,y)} = c$

$$I_{X;Y}(x_i;y_j) = \log \frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)}$$

$$I_{X;Y}(x_i;y_j) = \log \frac{P_{X|Y}(x_i|y_j)\;P(y_i)}{P_X(x_i)\;P(y_i)}$$

$$= \log \frac{P_{XY}(x_i,y_i)}{P_X(x_i)\,P_Y(y_j)} = \log \frac{P_{Y|X}(y_j|x_i)\,\cancel{P(x_i)}}{\cancel{P_X(x_i)}\,P_Y(y_j)}$$

$$= \log \frac{P_{Y|X}(y_j|x_i)}{P(y_j)}$$

$$= I_{Y;X}(y_j;x_i)$$

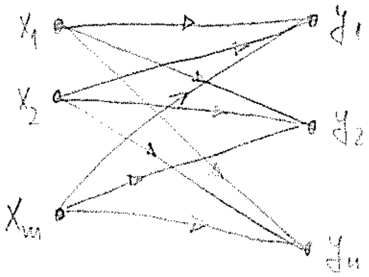
$$I(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} P_{XY}(x_i,y_j)\,\log\frac{P_{XY}(x_i,y_i)}{P_X(x_i)\,P_Y(y_j)}$$

$$I(u_1;u_2|u_3) = \log\frac{P(u_1|u_2,u_3)}{P(u_1|u_3)} = I(u_1|u_3) - I(u_1|u_2,u_3)$$

$$I(U_1;U_2|U_3) = \sum_{u_1}\sum_{u_2}\sum_{u_3} P(u_1,u_2,u_3)\cdot\log\frac{P(U_1|U_2,U_3)}{P(u_1|u_3)}$$

$$= H(U_1|u_3) - H(U_1|U_2U_3)$$

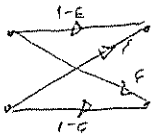
Average Mutual Information

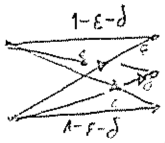


Discrete
Memoryless
Channel

Channel Models

Binary Symmetric Channel





Binary Erasure Channel

$P(x_i) \quad 1 \le i \le m$

$P(y_j) \quad 1 \le j \le m$

$P(y_j | x_i)$

$$H(X) = \sum_{i=1}^{m} P(x_i) \log \frac{1}{P(x_i)}$$

$$H(Y) = \sum_{j=1}^{n} P(y_j) \log \frac{1}{P(y_j)}$$

mutual information

$$I(x_i ; y_j) = \log \frac{P(x_i | y_j)}{P(x_i)}$$

– information provided about x_i by y_j

$$I(U_1 ; U_2 | U_3) = \sum_{u_1} \sum_{u_2} \sum_{u_3} P(u_1, u_2, u_3) \log \frac{P(u_1 | u_2, u_3)}{P(u_1 | u_3)}$$

$$I(x_i ; y_j) = \log \frac{P(x_i, y_i)}{P(x_i) P(y_j)}$$

amount of average uncertainty remaining in X after observation of Y

$$H(X|Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) \log \frac{1}{P(x_i | y_j)}$$

$$I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j)\, I(x_i, y_j)$$

average mutual information

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(Y) - H(Y|X)$$

$$I(X,Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i) P(y_j | x_i) \log \frac{P(y_j | x_i)}{P(y_j)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) \log P(y_j | x_i) - \sum_{j=1}^{n} \log P(y_j) \underbrace{\sum_{i=1}^{m} P(x_i) P(y_j | x_i)}_{P(y_j)}$$

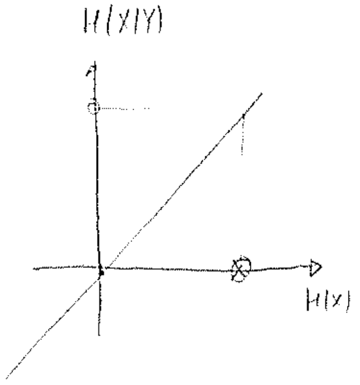
$$= -H(Y|X) + H(Y)$$

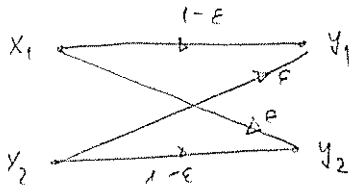
$$I(X;Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} P(x_i)P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} P(x_i)P(y_j|x_i) \log P(y_j|x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} P(x_i)P(y_j|x_i) \log P(y_j)$$

$$= \sum_{i=1}^{m} P(x_i) \sum_{j=1}^{n} P(y_j|x_i) \log P(y_j|x_i) + \sum_{j=1}^{n} \log P(y_j) \underbrace{\sum_{i=1}^{m} P(x_i)P(y_j|x_i)}_{P(y_j)}$$

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





$$I(X,Y) = \sum_{i=1}$$

$$I(X,Y) = P(x_1)(1-\varepsilon) ,$$

$$I(X,Y) = H(Y) - H(Y|X)$$

$$H(Y|X) = \sum_{i=1}^{m}\sum_{j=1}^{m} P(x_i, y_i) \log \frac{1}{P(y_j|x_i)}$$

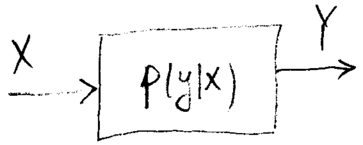
$$(1-p)(1-\varepsilon) \log \frac{1}{1-\varepsilon} + (1-p)\varepsilon \log \frac{1}{\varepsilon}$$

$$p(1-\varepsilon) \log \frac{1}{1-\varepsilon} + p\varepsilon \log \frac{1}{\varepsilon}$$

$$(1-\varepsilon) \log \frac{1}{1-\varepsilon} + \varepsilon \log \frac{1}{\varepsilon}$$

Arimoto- Blahut Algorithm for computing channel capacity

- Discrete memoryless channel



- The goal is to solve the following optimization problem

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- Rewrite the expression for C
using $p(x,y) = p(y)\, p(x|y) = p(x) \cdot p(y|x)$.

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \sum_x \sum_y p(x)\, p(y|x) \log \frac{\cancel{p(y)}\, p(x|y)}{p(x)\, \cancel{p(y)}}$$

$$= \max_{p(x)} \sum_x \sum_y p(x)\, p(y|x) \log \frac{p(x|y)}{p(x)}$$

Arimoto - Blahut : - treat $p(x)$ and $p(x|y)$ as independent variables - the algorithm monotonically converges to $p(x)$.

$p(x) \rightarrow q(x)$

$p(x|y) \rightarrow Q(x|y)$

- if $p(x)$ is nonunique the error can be bounded by $\log|X|/n$. does not depend on number of steps. channel
- $p(x)$ - unique error exponentially decreases

$$C = \max_{q(x),\, Q(x|y)} \sum_x \sum_y q(x) \cdot p(y|x) \cdot \log \frac{Q(x|y)}{q(x)}$$

1) fix $q(x)$ $\rightarrow$ $Q^*(x|y) = \dfrac{q(x) \cdot p(y|x)}{\sum_{x'} q(x')\, p(y|x')}$

2) fix $Q(x|y)$ $\rightarrow$ $q^*(x) = \dfrac{\prod_y Q(x|y)^{p(y|x)}}{\sum_{x'} \prod_y Q(x'|y)^{p(y|x')}} \;(=)\; \dfrac{e^{\sum_y p(y|x) \log Q(x|y)}}{\sum_{x'} e^{\sum_y p(y|x') \log Q(x'|y)}}$

1),2) ⇒ 3) $q^{(1)}(x) \rightarrow Q^{(1)}(x|y) \rightarrow q^{(2)}(x) \rightarrow Q^{(2)}(x|y) \rightarrow \ldots$

Information Measures for Continuous Random Variables

$$I(X;Y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x)p(y|x) \log \frac{p(y|x)p(x)}{p(x)p(y)} \, dx\,dy$$

- Mutual Information is thus analogous extention of discrete random variable mutual inf
- The concept of self information does not carries over to continuous rv because

$$H(X) = -\int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

can be infinite

$$H(X|Y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y) \log p(x,y) \, dx\,dy$$

$$I(X;Y) = H(X) - H(X|Y)$$

In some cases X is discrete while Y is continuous

$$p(y) = \sum_{i=1}^{n} P(x_i) p(y|x_i)$$

$$I(x_i;Y) = \log \frac{p(y|x_i)P(x_i)}{p(y)P(x_i)} = \log \frac{p(y|x_i)}{p(y)}$$

$$I(X;Y) = \sum_{i=1}^{n} \int_{-\infty}^{+\infty} p(y|x_i)P(x_i) \log \frac{p(y|x_i)}{p(y)} dy$$

Example: $p(y|\pm A) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y\mp A)^2}{2\sigma^2}}$

Asymptotic Equipartition Property (EAP)

This property in information theory is analogous to the weak law of large numbers in probability

Recall that Bernoulli proved that if the probability of an event A in a given experiment equals p, and the number of occurence of A in n trials equals k, then

$$\Pr\left\{ \left| \frac{k}{n} - p \right| < \varepsilon \right\} \to 1 \qquad \text{as} \quad n \to \infty$$

(Borel gave a strong law of large numbers showing that the convergence is not only in probability)

In other words the weak law of large numbers states that for independent identically distributed (i.i.d) random variables $X_1, X_2, .. X_n$,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx E\{X\} \qquad \text{for large } n$$

The EAP states that

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

Thus the probability of observing the sequence $X_1, X_2, .. , X_n$

$$p(X_1, X_2, .., X_n) \approx 2^{-nH(X)}$$

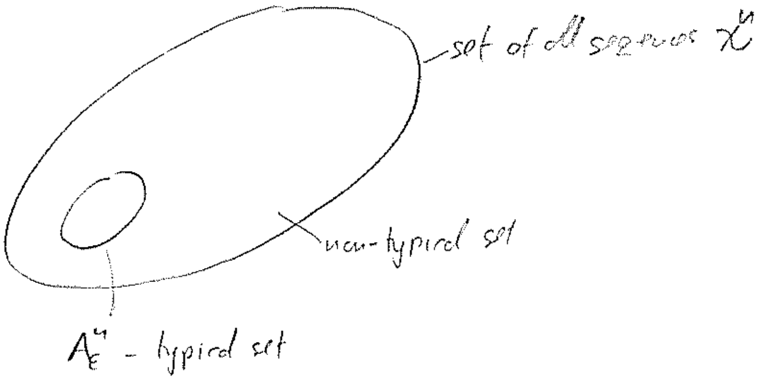
or more rigorously

$$\Pr\left\{ (X_1, X_2, .., X_n) \;\middle|\; p(X_1, X_2, ... X_n) = 2^{-n(H(X) \pm \varepsilon)} \right\} \approx 1$$

"Almost all events are almost equally surprising."

The typical set $A_\varepsilon^{(n)}$ with respect to $p(x)$ is set of sequences $(x_1, x_2, .. x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\varepsilon)} \le p(x_1, x_2, .., x_n) \le 2^{-n(H(X)-\varepsilon)}$$


set of all sequences $\mathcal{X}^n$
non-typical set
A_ε^n - typical set

Properties of typical sets

$$(x_1, x_2, .. x_n) \in A_\varepsilon^{(n)} \Rightarrow p(x_1, x_2, .., x_n) = 2^{-n(H \pm \varepsilon)}$$

$$\Pr\{A_\varepsilon^{(n)}\} > 1-\varepsilon \qquad n \to \infty$$

$$|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$$

Consequence : - Not much information is lost by not transmitting the non-typical sequences

- To transmit typical sequences we need at most $n(H(X)+\varepsilon)+1$ bits