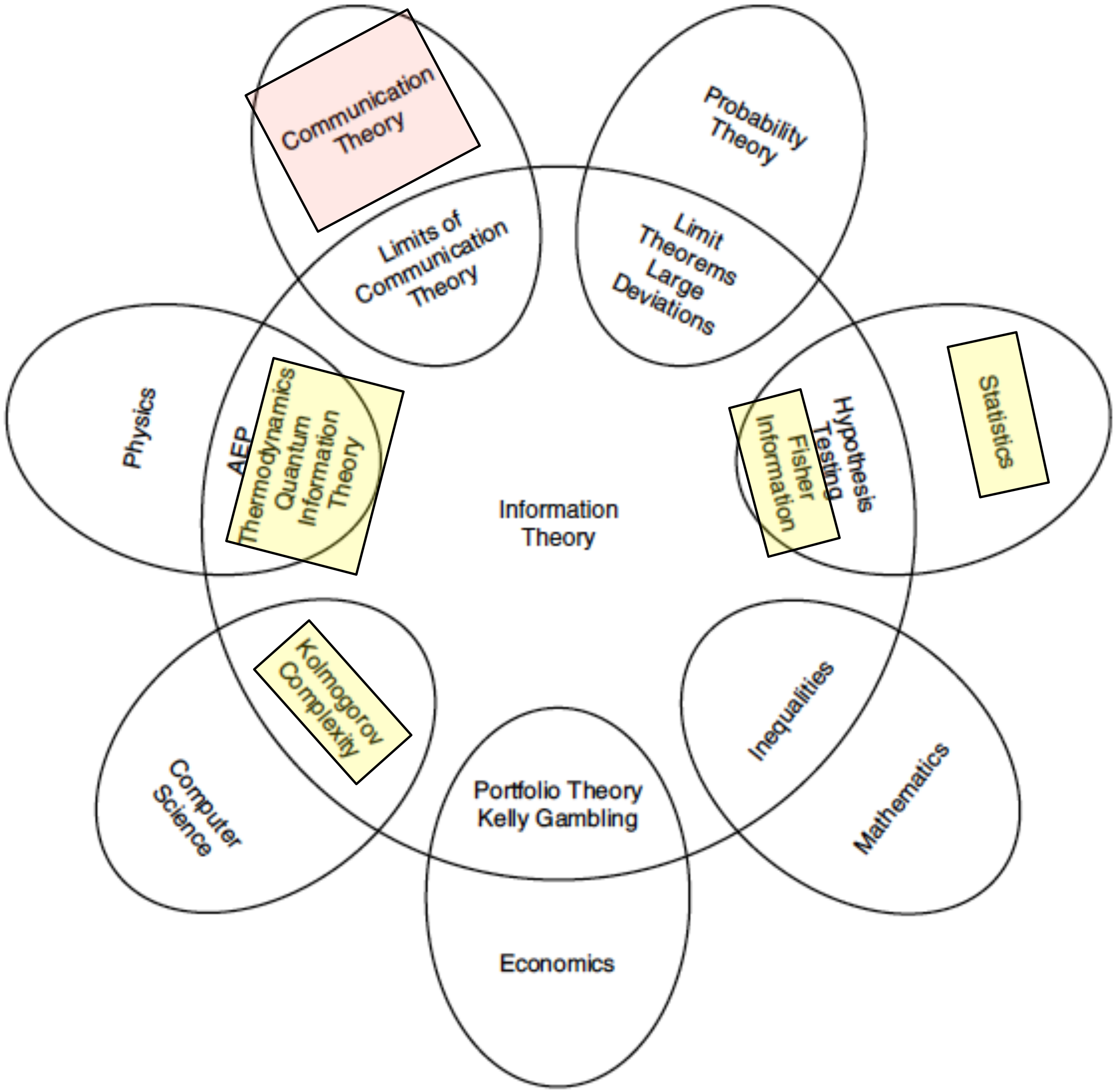
Elements of Information Theory



Relationship of information theory to other fields.

Materials from the book of T. M. Cover and J. M. Thomas, “Element of information theory”, Wiley.

Measure of Information (of an event)

- Given a probability mass function (pmf) p(x) of a random variable X.
- **The information, associated to an event with probability p(x), is defined as**

$$I(x) = -\log\left[p(x)\right]$$

Units: bit

- Less frequent event ➔➔➔ A LOT OF information.
- More frequent event ➔➔➔ SMALL information.
- Base of the Log is 2 (we do not lose generality).

Auto-información

una medida de información debe cumplir las siguientes condiciones:

1. El contenido de información de un suceso, que denotamos $I_X(x_i)$ y se denomina *auto-información*, debe depender de la probabilidad del suceso y no del propio suceso
$$I_X(x_i) = f(p_X(x_i)).$$
2. Debe ser una función decreciente de la probabilidad.
$$p_X(x_i) > p_X(x_j) \rightarrow I_X(x_i) < I_X(x_j).$$
3. Debe ser una función continua de la probabilidad.
4. Si $p_{X,Y}(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j)$, entonces
$$I_{X,Y}(x_i, y_j) = I_X(x_i) + I_Y(y_j).$$

Se puede demostrar que la única función que cumple estas propiedades es la función logarítmica. Por tanto, la auto-información se define como

$$I_X(x_i) = -\log(p_X(x_i)).$$

La base del logaritmo no es importante. Lo único que implica son las unidades en que se expresa la información. Si la base es 2, las unidades son *bits*, y si se usa el logaritmo natural o neperiano, las unidades son *nats*.

Discrete Entropy

- Expected value of the information

$$H(X) = H_X = -\sum_{i=1}^{N} p(x=i)\log\left[p(x=i)\right]$$

- IT IS A SCALAR VALUE.
- It can be considered as a DISPERSION MEASURE of the pmf p(x).
- The notation H(X) means that is related to the r.v. X.
- H(X) represents the UNCERTAINTY over the values that the random variable X can take.

Discrete Entropy

The entropy of a random variable X with a probability mass function $p(x)$ is defined by

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \tag{1.1}$$

We use logarithms to base 2. The entropy will then be measured in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on average required to describe the random variable.

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

→ The log is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit. We will use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $x \log x \to 0$ as $x \to 0$. Adding terms of zero probability does not change the entropy.

Example Consider a random variable that has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values. Thus, 5-bit strings suffice as labels.

The entropy of this random variable is

$$H(X) = -\sum_{i=1}^{32} p(i) \log p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits},$$

which agrees with the number of bits needed to describe X. In this case, all the outcomes have representations of the same length.

Example Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. We can calculate the entropy of the horse race as

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\frac{1}{64}\log\frac{1}{64} \\ &= 2 \text{ bits.} \end{aligned}$$

If the base of the logarithm is b, we denote the entropy as $H_b(X)$. If the base of the logarithm is e, the entropy is measured in *nats*. Unless otherwise specified, we will take all logarithms to base 2, and hence all the entropies will be measured in bits. Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.

Lemma $H(X) \geq 0.$

Proof: $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0.$ □

Lemma $H_b(X) = (\log_b a) H_a(X).$

Proof: $\log_b p = \log_b a \log_a p.$ □

Example Let

$$X = \begin{cases} a & \text{with probability } \frac{1}{2}, \\ b & \text{with probability } \frac{1}{4}, \\ c & \text{with probability } \frac{1}{8}, \\ d & \text{with probability } \frac{1}{8}. \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits.}$$

The entropy does not depend on the values that the r.v. X can take.

(in the example above they can be considered generic math-variables or simply “letters”….)

IMPORTANT:

The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable.

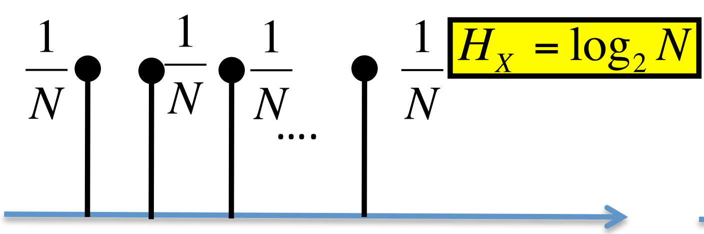
H(X)=0 if the probability is of type
0,0,0,1,0,....0

H(X)=log N (i.e, its maximum value) if the probability is of type
1/N,1/N, 1/N, 1/N, 1/N,.... 1/N

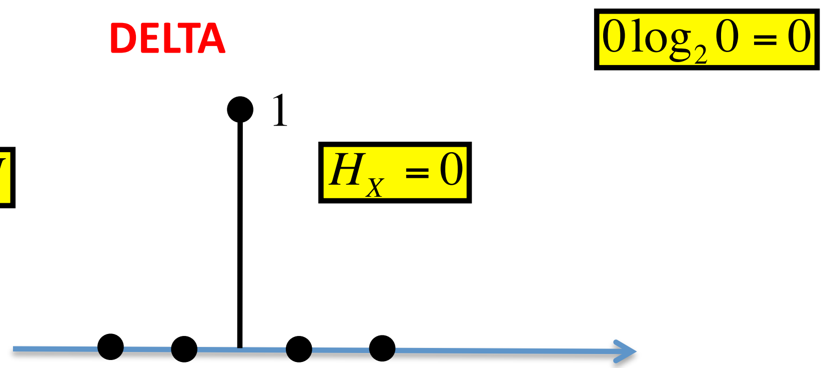
Entropy: measure of dispersion

- H(X) is a measure of DISPERSION (UNCERTAINTY):

MAX DISCRETE ENTROPY: UNIFORM PMF



MIN DISCRETE ENTROPY: DELTA



- we do not consider the continuous scenario: *Differential entropy (continuous case) is max when p(x) is a Gaussian density.*

Relationship with the variance

- Another dispersion measure is the variance. BUT the variance depends on the support of the r.v. X (i.e., the values than X can take).

In this two pmfs: the entropy is the same!!! But the variance no!

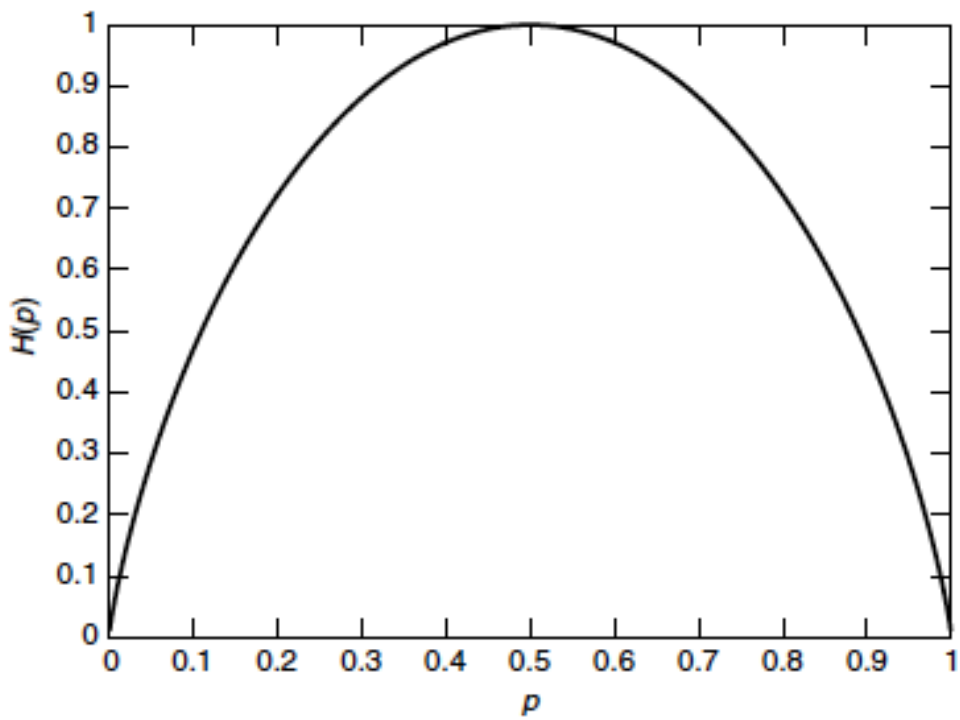
- For instance, we can permute the positions of the deltas and the entropy does not change.

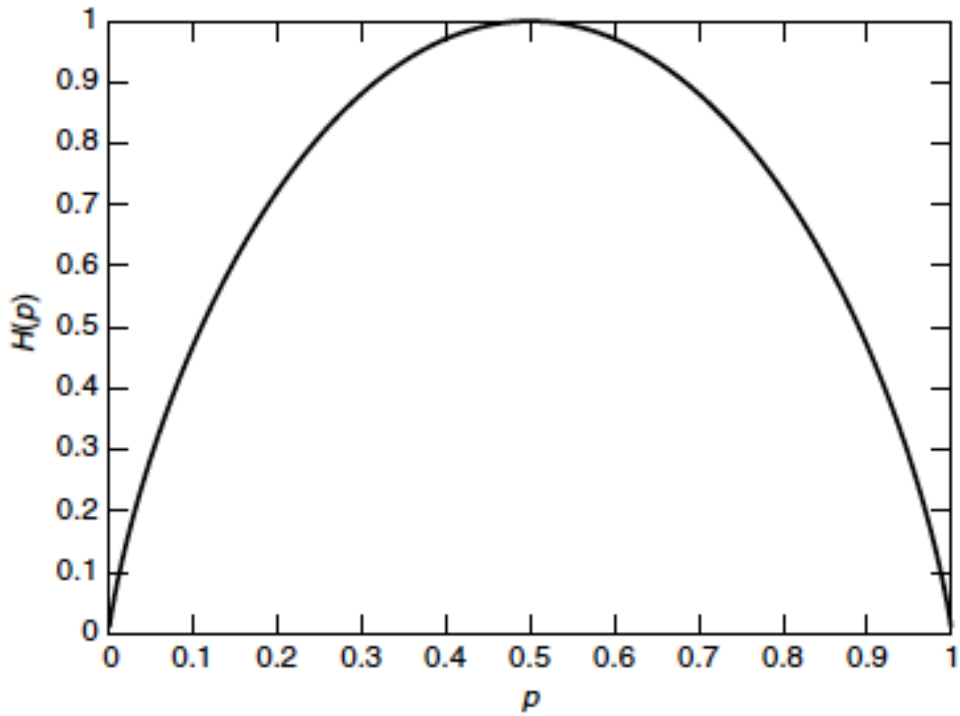
Example Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Then

$$H(X) = -p \log p - (1-p) \log(1-p) \stackrel{\text{def}}{=\!=} H(p).$$





-What is the more “informative” system?
The first one, 2 events with probability of 0.9 and 0.1
The second one, 2 events with probability of 0.5 and 0.5

We have more “questions” in the second case….

We denote expectation by E. Thus, if $X \sim p(x)$, the expected value of the random variable $g(X)$ is written

$$E_p[g(X)] = \sum_{x \in \mathcal{X}} g(x) p(x),$$

Remark The entropy of X can also be interpreted as the expected value of the random variable $\log \frac{1}{p(X)}$, where X is drawn according to probability mass function $p(x)$. Thus,

$$H(X) = E_p \left[\log \frac{1}{p(X)} \right]$$

Joint Entropy of two r.v.'s X, Y

Definition The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

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

$$H(X, Y) = -E[\log p(X, Y)].$$

Conditional Entropy - Y|X

Definition If $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y|X)$ is defined as

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x) H(Y|X = x)$$

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

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

$$= -E[\log p(Y|X)]$$

This guy does not need presentation… it is a standard entropy!

Relationship among entropy, joint entropy and conditional entropy

Theorem (*Chain rule*)

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

Proof

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

Equivalently, we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X)$$

and take the expectation of both sides of the equation to obtain the theorem. □

Corollary

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

Remark Note that $H(Y|X) \neq H(X|Y)$. However, $H(X) - H(X|Y) = H(Y) - H(Y|X)$, a property that we exploit later.

En general, aplicando la regla de la cadena, se tiene la relación

$$H(\boldsymbol{X}) = H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X_1, X_2, \cdots, X_{N-1}).$$

Cuando $(X_1, X_2, \cdots, X_N)$ son variables aleatorias independientes,

$$H(\boldsymbol{X}) = \sum_{i=1}^{N} H(X_i).$$

Example
"Joint"

Let (X, Y) have the following joint distribution:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

This is the joint pmf…

Are you able to find marginal and conditional pmfs?

→ The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and hence $H(X) = \frac{7}{4}$ bits and $H(Y) = 2$ bits. Also,

$$\begin{aligned} H(X|Y) &= \sum_{i=1}^{4} p(Y=i)H(X|Y=i) \\ &= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\ &\quad + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4}H(1, 0, 0, 0) \\ &= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 \\ &= \frac{11}{8} \text{ bits.} \end{aligned}$$

Similarly, $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.

Relative entropy – KL divergence

Definition The *relative entropy* or *Kullback–Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

$$= E_p \left[\log \frac{p(X)}{q(X)} \right].$$

In the above definition, we use the convention that $0 \log \frac{0}{0} = 0$ and the convention (based on continuity arguments) that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Thus, if there is any symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

We will soon show that relative entropy is always nonnegative and is zero if and only if $p = q$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often useful to think of relative entropy as a "distance" between distributions.

Relative entropy – KL divergence

The *relative entropy* is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.

Note that again it **is not symmetric**, and it is quite useful for the the **causality** (where important concept, especially in biomedical applications).

Theorem 2.6.3 (*Information inequality*) *Let* $p(x), q(x), x \in \mathcal{X}$, *be two probability mass functions. Then*

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ *for all* x.

Example Let $\mathcal{X} = \{0, 1\}$ and consider two distributions p and q on $\mathcal{X}$. Let $p(0) = 1 - r$, $p(1) = r$, and let $q(0) = 1 - s$, $q(1) = s$. Then

$$D(p||q) = (1 - r)\log\frac{1 - r}{1 - s} + r\log\frac{r}{s}$$

and

$$D(q||p) = (1 - s)\log\frac{1 - s}{1 - r} + s\log\frac{s}{r}.$$

If $r = s$, then $D(p||q) = D(q||p) = 0$. If $r = \frac{1}{2}$, $s = \frac{1}{4}$, we can calculate

$$D(p||q) = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 1 - \frac{1}{2}\log 3 = 0.2075 \text{ bit},$$

whereas

$$D(q||p) = \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{3}{4}\log 3 - 1 = 0.1887 \text{ bit}.$$

Note that $D(p||q) \neq D(q||p)$ in general.

MUTUAL INFORMATION

Definition Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) \\ &= E_{p(x,y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \end{aligned}$$

It is symmetric.

MUTUAL INFORMATION

Corollary *(Nonnegativity of mutual information) For any two random variables, X, Y,*

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

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

IMPORTANT: WE CAN STUDY DEPENDENCY/INDEPENDENCY BETWEEN RANDOM VARIABLES (different from the correlation coefficient…).

---

1. $I(X,Y) = I(Y,X) \geq 0$.
   La igualdad se cumple en el caso de que X e Y sean independientes.

2. $I(X,Y) \leq \text{mín}(H(X), H(Y))$.
   La información mutua nunca puede ser mayor de la que tiene cada una de las variables

Relationship between ENTROPY and MUTUAL INFORMATION

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

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right)$$

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

Thus, the mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.

By symmetry, it also follows that

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

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

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

Finally, we note that

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

Thus, the mutual information of a random variable with itself is the entropy of the random variable. This is the reason that entropy is sometimes referred to as *self-information.*

Example For the joint distribution of Example **Ex-Joint** is easy to calculate the mutual information $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = 0.375$ bit.

“information can’t hurt”

Theorem *(Conditioning reduces entropy)(Information can’t hurt)*

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

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

Proof: $0 \le I(X;Y) = H(X) - H(X|Y)$. □

Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average. Specifically, $H(X|Y = y)$ may be greater than or less than or equal to $H(X)$, but on the average $H(X|Y) = \sum_y p(y)H(X|Y = y) \le H(X)$. For example, in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.

Example Let (X, Y) have the following joint distribution:

$Y \backslash X$	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Then $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$ bit, $H(X|Y = 1) = 0$ bits, and $H(X|Y = 2) = 1$ bit. We calculate $H(X|Y) = \frac{3}{4} H(X|Y = 1) + \frac{1}{4} H(X|Y = 2) = 0.25$ bit. Thus, the uncertainty in X is increased if $Y = 2$ is observed and decreased if $Y = 1$ is observed, but uncertainty decreases on the average.

SUMMARY

- Recall the definitions:

$$H(X,Y) = H_{XY} = -\sum_{j=1}^{L}\sum_{i=1}^{N} p(x=i,y=j)\log\left[p(x=i,y=j)\right]$$

$$H(X|Y) = H_{X|Y} = -\sum_{j=1}^{L}\sum_{i=1}^{N} p(x=i,y=j)\log\left[p(x=i|y=j)\right]$$

$$H(Y|X) = H_{Y|X} = -\sum_{j=1}^{L}\sum_{i=1}^{N} p(x=i,y=j)\log\left[p(y=j|x=i)\right]$$

$$I(X;Y) = I_{XY} = -\sum_{j=1}^{L}\sum_{i=1}^{N} p(x=i,y=j)\log\left[\frac{p(x=i)p(y=j)}{p(x=i,y=j)}\right]$$

Recall that:

$$p(x,y) = p(y|x)p(x)$$

$$p(x,y) = p(x|y)p(y)$$

SUMMARY - RELATIONSHIPS

Theorem *(Mutual information and entropy)*

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

---

1. $I(X,Y) = I(Y,X) \geq 0$.
   La igualdad se cumple en el caso de que X e Y sean independientes.
2. $I(X,Y) \leq \text{mín}(H(X), H(Y))$.
   La información mutua nunca puede ser mayor de la que tiene cada una de las variables

SUMMARY - RELATIONSHIPS

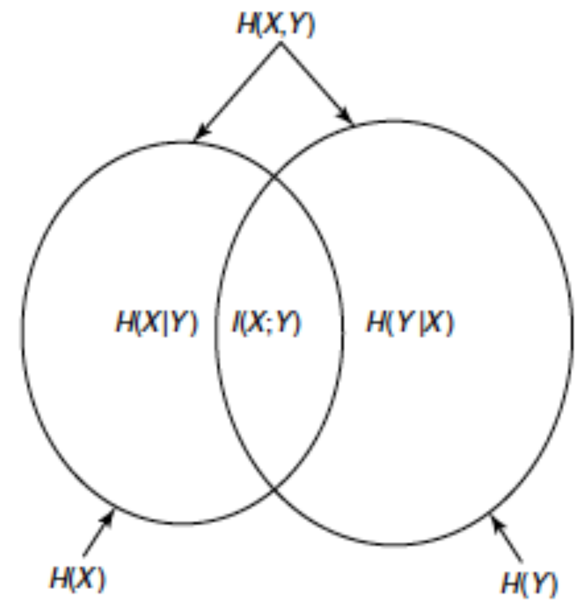
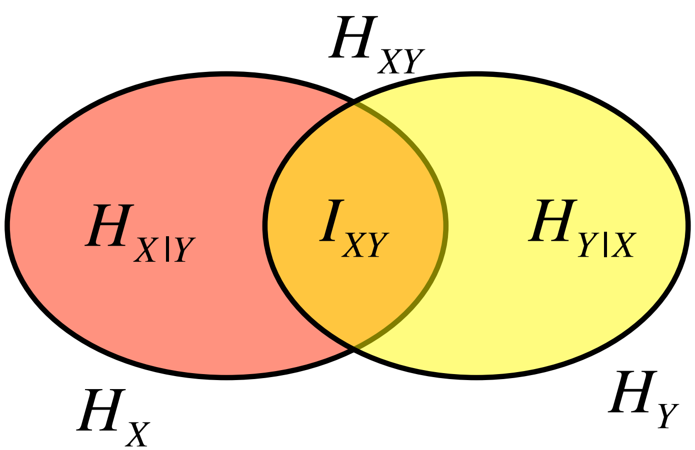


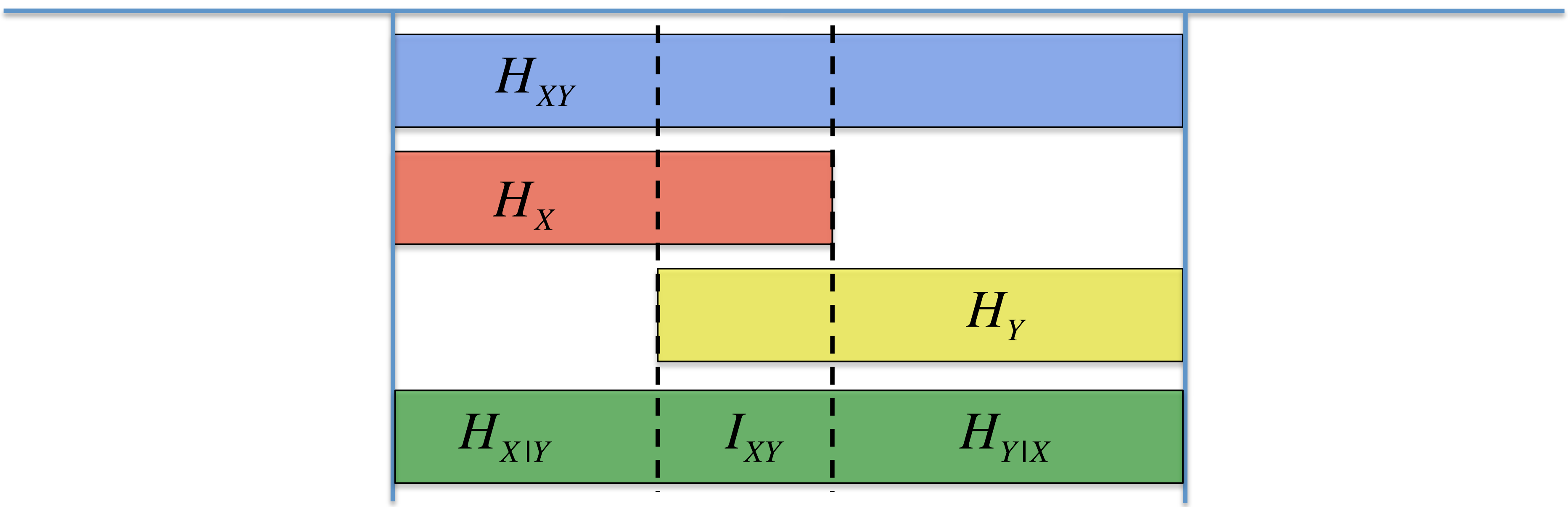
FIGURE Relationship between entropy and mutual information.

The relationship between $H(X)$, $H(Y)$, $H(X, Y)$, $H(X|Y)$, $H(Y|X)$, and $I(X; Y)$ is expressed in a Venn diagram → Notice that the mutual information $I(X; Y)$ corresponds to the intersection of the information in X with the information in Y.

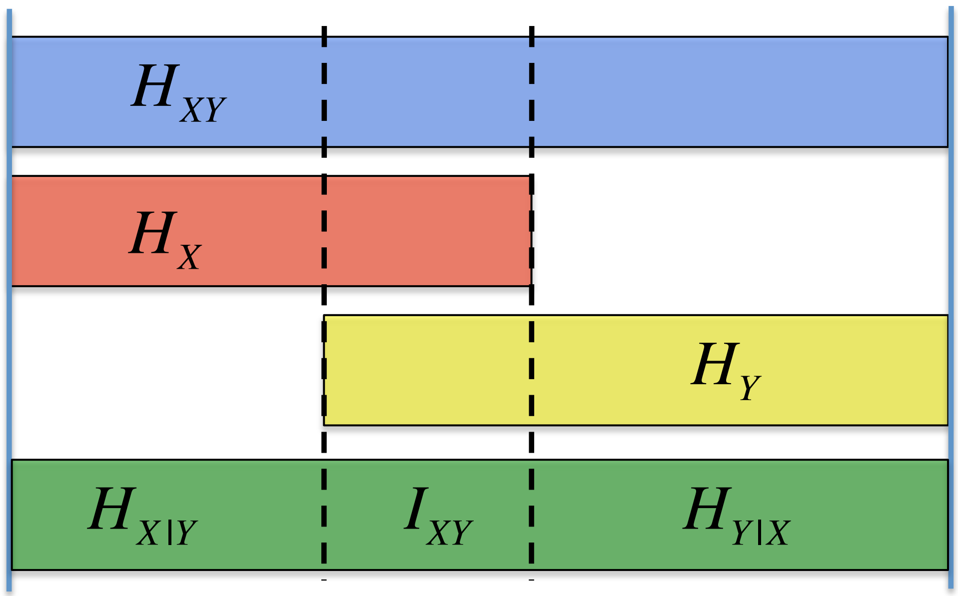
RELATIONSHIPS



Red: Hx
Yellow: Hy
Red+Yellow=Hxy (joint)



RELATIONSHIPS



We can obtain the inequalities:

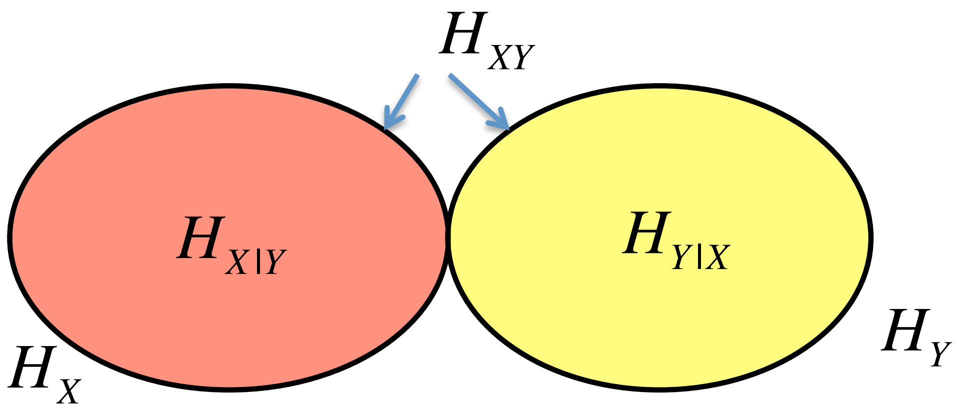
$$H_{XY} \le H_X + H_Y$$
$$H_{XY} = H_X + H_Y - I_{XY}$$
$$H_{XY} = H_{X|Y} + H_{Y|X} + I_{XY}$$
$$H_{XY} = H_X + H_{Y|X}$$
$$H_{XY} = H_Y + H_{X|Y}$$

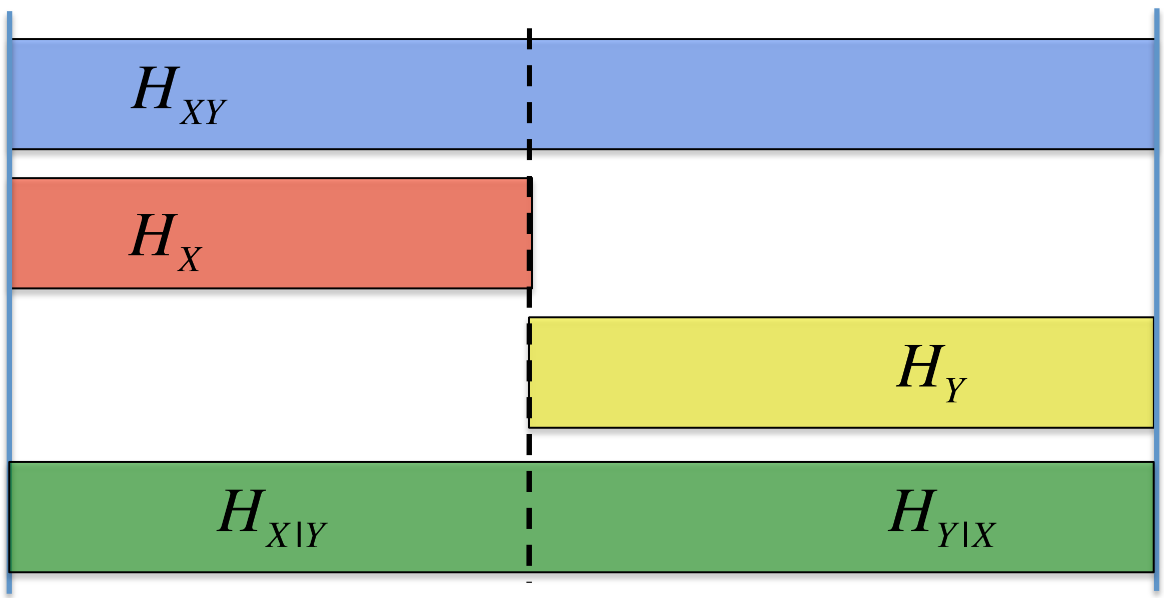
$$H_X = H_{X|Y} + I_{XY}$$
$$H_Y = H_{Y|X} + I_{XY}$$

$$H_X \le H_{XY} \le H_X + H_Y$$
$$H_Y \le H_{XY} \le H_X + H_Y$$

$$I_{XY} = H_X - H_{X|Y}$$
$$I_{XY} = H_Y - H_{Y|X}$$
$$I_{XY} = H_X + H_Y - H_{XY}$$
$$I_{XY} = I_{YX}$$

Independent Variables





$$I_{XY} = 0$$

$$H_X = H_{X|Y}$$
$$H_Y = H_{Y|X}$$
$$H_{XY} = H_X + H_Y$$

The joint entropy is max, and I(X,Y) is min

Case X=Y (totally dependent)

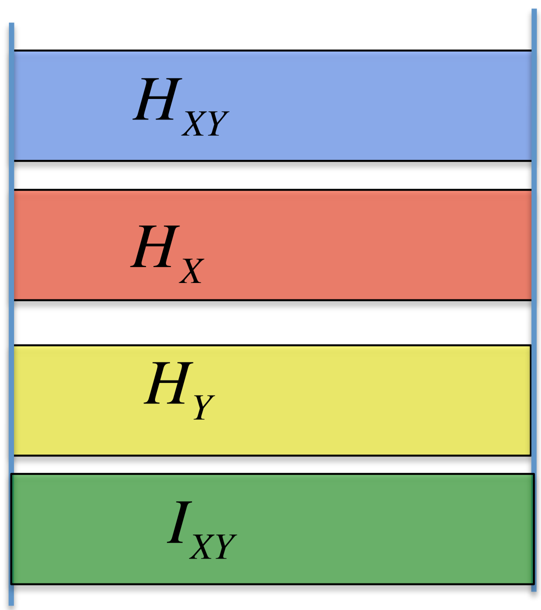
$$H_{XY} = H_X = H_Y = I_{XY}$$

$$I_{XY} = H_X = H_Y$$

$$I_{XY} = H_X = H_Y = H_{XY}$$

$$H_{X|Y} = 0$$

$$H_{Y|X} = 0$$



Important formulas

- Recall:

p(x) delta → $0 \le H_X \le \log_2 M$ ← p(x) uniform

$0 \le H_Y \le \log_2 L$

X=Y → $(H_Y =) H_X \le H_{XY} \le H_X + H_Y$ ← Independent variables

Independent variables → $0 \le I_{XY} \le H_X (= H_Y)$ ← X=Y

X=Y → $0 \le H_{X|Y} \le H_X$ ← Independent variables

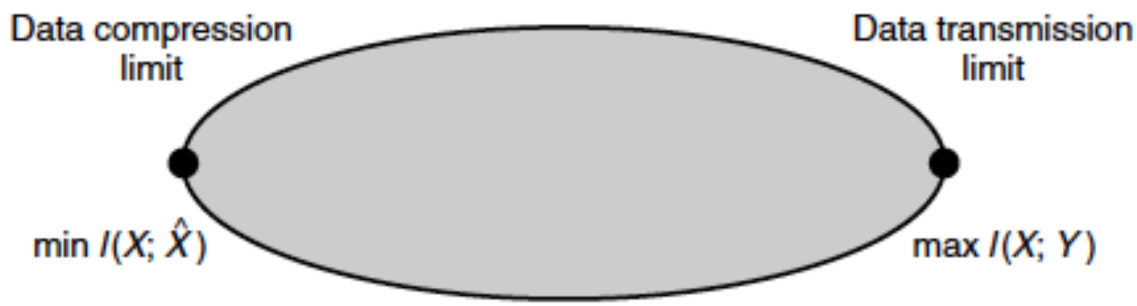
X=Y → $0 \le H_{Y|X} \le H_Y$ ← Independent variables

Data-processing inequalities

Theorem (*Data-processing inequality*) *If* $X \to Y \to Z$*, then* $I(X;Y) \geq I(X;Z)$.

Thus, the dependence of X and Y is decreased (or remains unchanged) by the observation of a "downstream" random variable Z.

More processing on the data, more loss of information….



Information theory as the extreme points of communication theory.

Some Material is from the book of T. M. Cover and J. M. Thomas, “Element of information theory”, Wiley.