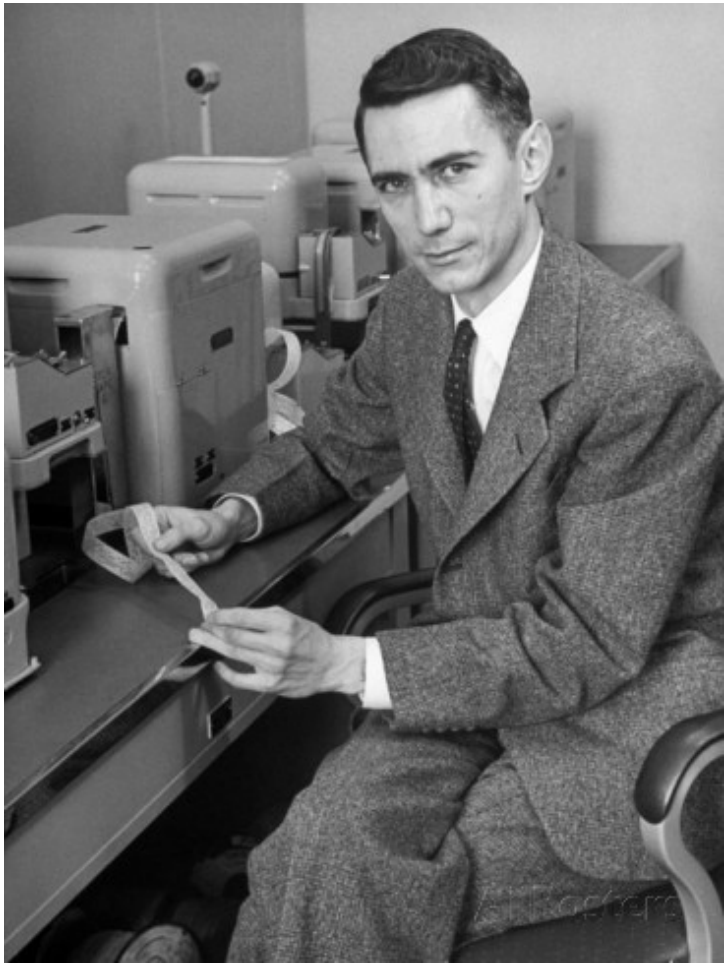
Information Theory
Degree in Data Science and Engineering
Lesson 2: Measures of information

Jordi Quer, Josep Vidal

Mathematics Department, Signal Theory and Communications Department
{jordi.quer, josep.vidal}@upc.edu

2019/20 - Q1

Shannon information theory



Claude E. Shannon (1916–2001)

Electrical engineer (Michigan Univ. and MIT)
and mathematician (Michigan)

Researcher in IAS-Princeton,
the Bell Labs and MIT

Worked in cryptography during WWII, and in 1948 he became the father of *Information Theory* while working in *Bell Labs*

Watch the Youtube video *Claude Shannon - Father of the Information Age*

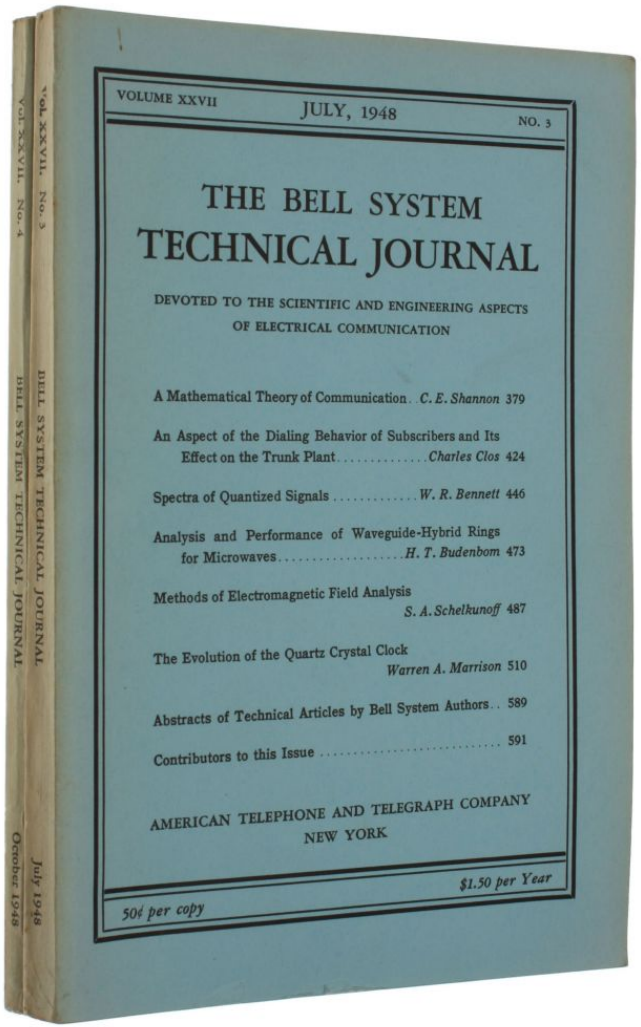
Shannon information theory

Founding articles:

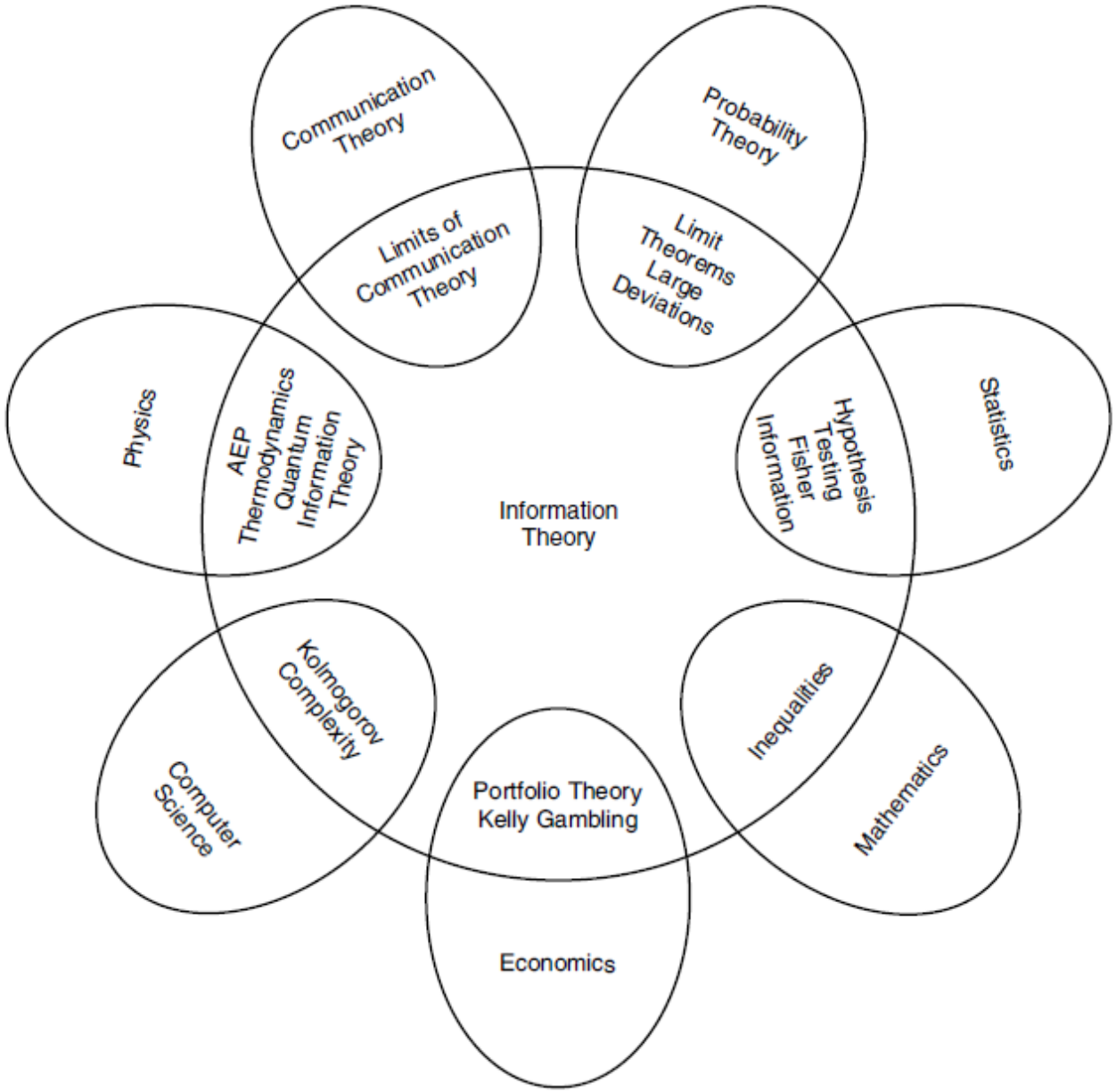
Claude E. Shannon,

A mathematical theory of communication, Bell System Technical Journal, Vol. 27 (1948) issue 3, Pages 379–423 (discrete channel), issue 4, Pages 623–656 (continuous channel)

Communication Theory of Secrecy Systems, Bell System Technical Journal, Vol. 28 (1949) issue 4, Pages 656–715



Fields of influence of information theory



Main applications in data communications

- *Data compression*. Source (or noiseless) coding theorem says that data can be compressed up to a certain minimum value, which is given by its *entropy*. In lectures 3 and 4.
- *Error correction*. Channel (or noisy) coding theorem says that information can be transmitted essentially without errors over a noisy channel, by adding *redundancy*, at a ratio depending on the *channel capacity*. In lectures 5 and 6.
- *Cryptography*. Shannon's perfect secrecy theorem says that absolutely secure systems do exist (*one-time-pad*), but the key must be as long as the message, and can only be used once. In lecture 7.

What is information?

Which information deserves newspaper's first page?

1. Dog bites man.
2. Man bites dog.

1. Volcano eruption in Indonesia.
2. Volcano eruption in Germany.

1. Weather: heavy rain showers over Sahara.
2. Weather: heavy rain showers over Scotland.

Unexpected and non-frequent events contain more information.

Uncertainty of an event depends on its probability. *Information* is a measure of uncertainty.

What is information?

"San Fermín 2019: Protesta de los corredores contra la organización por el uso de cabestros muy veloces y antideslizante en las calles". Predictability raises less interest as less information is conveyed.

What is information?

According to Shannon's theory, information comes from the knowledge of the outcome of an experiment or a data source, equivalently, the value x taken by a random variable X.

The *amount of information* contained in an outcome x of probability $p(X = x) = p(x) \in [0, 1]$ is defined to be:

$$I(x) = -\log_2 p(x) = \log_2 \frac{1}{p(x)} \qquad (\textit{base 2 logarithm}).$$

The unit of measurement is the *bit* (binary information unit), which should not be confused with a binary digit!

Other units may be used taking logarithm in other bases: decits in base 10, nats in natural logarithms of base e, etc. but the standard is the bit: base 2.

From now on, $\log$ will always denote logarithm in base 2.

What is information?

This measure of information has a set of desirable properties:

1. Continuity: smoothly change with $p(x)$.
2. Symmetry: the information associated to a sequence of outcomes does not depend on the order of appearance of those values.
3. Maximum value: the amount of information is maximum if the values observed are equiprobable.
4. Additivity: the information associated with a set of independent outcomes is obtained by adding the information of the individual outcomes.

Information: examples

The outcome (heads or tails) of a coin flip gives

$$I(x) = -\log \frac{1}{2} = 1 \quad \text{bit of information.}$$

The outcome of a dice roll (a number between 1 and 6) gives

$$I(x) = -\log \frac{1}{6} = 2.58496\ldots \quad \text{bits of information.}$$

A gambler plays the lottery with winning probability 10^{-5}.
The two possible outcomes give different amounts of information:

$$I(\text{winning}) = -\log 10^{-5} = 16.60964\ldots \text{ bits,}$$
$$I(\text{not winning}) = -\log(1 - 10^{-5}) = 0.000014427\ldots \text{ bits.}$$

- Low probable outcomes give a lot of information.
- Highly probable outcomes give almost no information.

Entropy

Let X be a random variable taking values in a countable set $\mathcal{X} = \{x_1, \ldots, x_q\}$ with probabilities $p(x_i) = \Pr(X = x_i)$. Shannon defines the entropy of X as the average amount of information contained in a sample of the variable:

$$H(X) = \sum_{x \in \mathcal{X}} p(x) I(x) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}.$$

Lemma

Since $0 \leq p(x) \leq 1$, $H(X) \geq 0$.

Like information, entropy is measured in bits, and it does only depend on the probability distribution $p(x_1), \ldots, p(x_q)$, not on the actual values $x_1, \ldots, x_q$. If $p(x_i) = 0$ the limit value of $p(x_i) \log p(x_i)$ is zero.

Entropy is the expectation of the random variable $I(X)$ who gives the amount of information of the outcomes of X: $I(X)$ takes each value $I(x_i)$ with probability $p(x_i)$.

Example: uniform distribution

Entropy of a coin flip:

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1 \text{ bit}$$

Entropy of a dice roll:

$$H(X) = \sum_{i=1}^{6} \frac{1}{6}\log 6 = 2.58496\ldots \text{ bits}$$

If X has q outcomes of probability $1/q$ (uniform distribution):

$$H(X) = \sum_{i=1}^{q} \frac{1}{q}\log q = \log q \text{ bits}$$

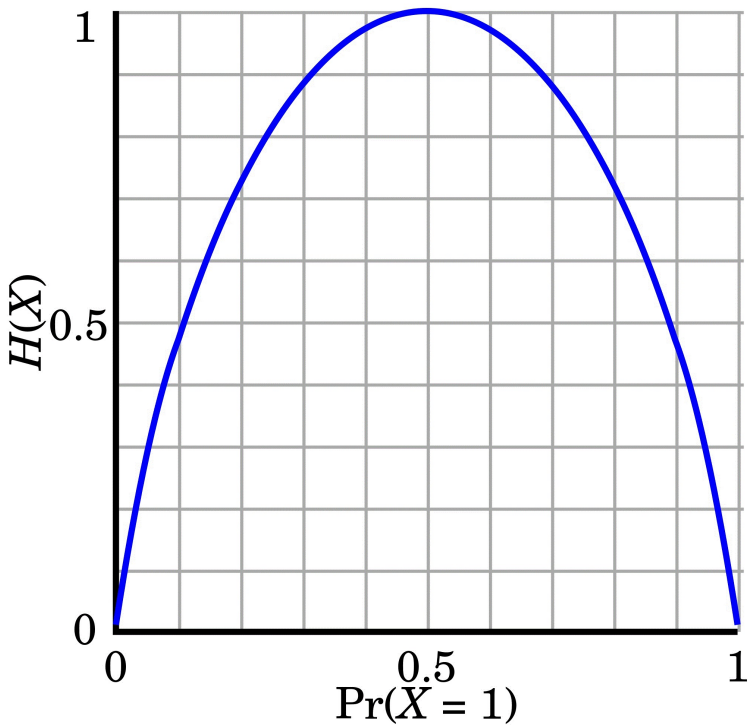
For example, the result of flipping k coins has k bits of entropy: 2^k possible outcomes all with the same probability.

Example: Bernoulli distribution

If X has Bernoulli distribution with probability p of success:

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$

We also usually call it $H(p)$.



Maximum average information for $p = 0.5$ (ex: a fair coin flip).

Average information contained in winning the lottery with probability $p = 10^{-5}$:

$$H(X) = 0.00018052\ldots$$

bits of information.

Jensen's inequality

In order to further derive properties of information measures we need to formulate the following theorem...

Theorem (Jensen's inequality)

If f is a convex function and X is a random variable, then

$$\mathbb{E}\left[f\left(X\right)\right] \geq f\left(\mathbb{E}\left[X\right]\right)$$

If f is strictly convex, equality implies $\mathbb{E}\left[X\right] = X$ with probability 1, that is X is constant.

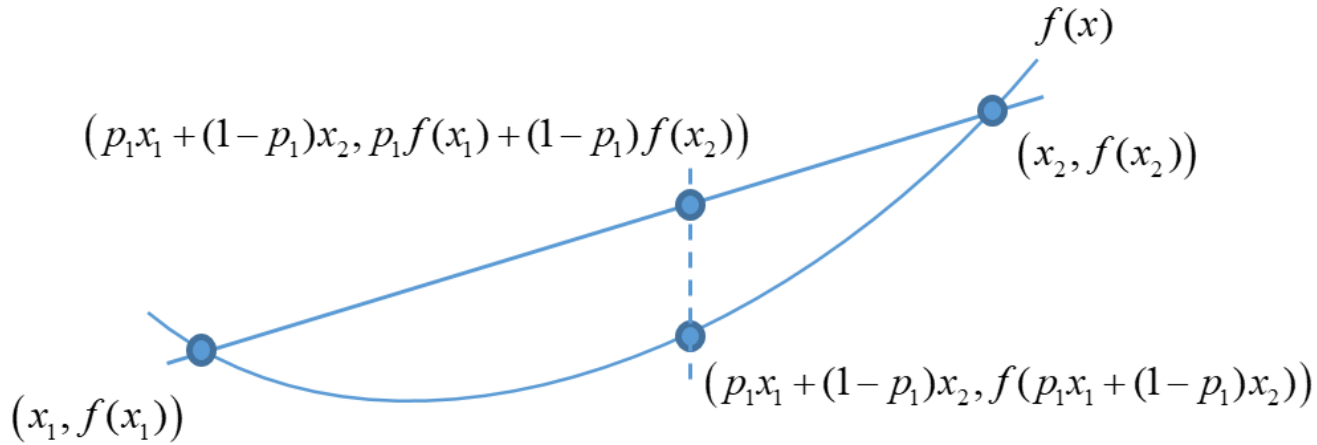
Proof. For the sake of a simpler formulation let us denote $p_i = p(x_i)$. Consider a discrete distribution and proceed by induction. For $n = 2$, inequality is

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

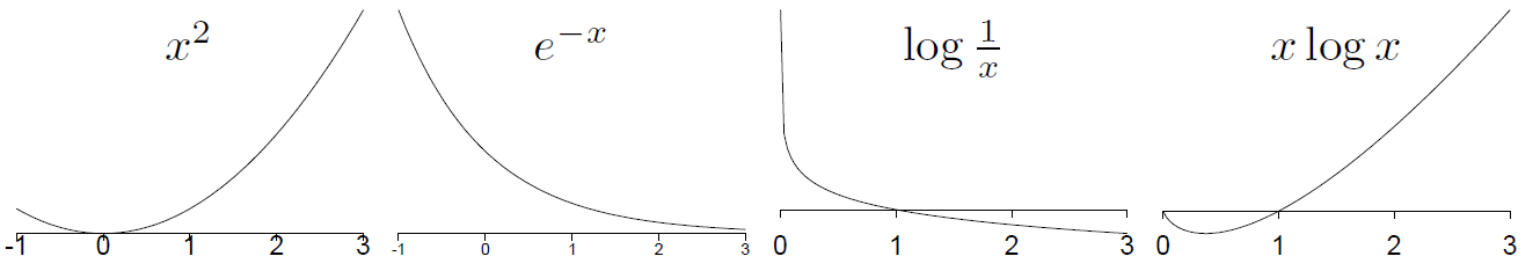
which is true as far as f is convex and $p_1 = 1 - p_2$:

Jensen's inequality

Proof (cont.)



Remember that for a convex function, $\frac{\partial^2 f}{\partial x^2} \geq 0 \quad \forall x$. Some examples of convex functions are...



Jensen's inequality

Proof (cont.) By applying convexity on f,

$$f\left(\sum_{i=1}^{k} p_i x_i\right) = f\left(p_1 x_1 + (1-p_1)\sum_{i=2}^{k} p'_i x_i\right)$$
$$\leq p_1 x_1 + (1-p_1) f\left(\sum_{i=2}^{k} p'_i x_i\right)$$

where $p'_i = \frac{p_i}{1-p_1}$ for $i = 2, \ldots, k$. Since $\sum_{k=1}^{k} p'_k = 1$, we can apply induction to get

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

The proof can be extended to continous-valued random variables and exchange summation by integrals. □

Instead, if f is concave, inequality is reversed.

Maximum value of entropy

Theorem (Maximum value of entropy)

$$H(X) \leq \log |\mathcal{X}|$$

with equality iff X has a uniform distribution over $\mathcal{X}$.

Proof. From Jensen's inequality, and since $\log(x)$ is concave:

$$\begin{aligned} H(X) - \log |\mathcal{X}| &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} - \sum_{x \in \mathcal{X}} p(x) \log |\mathcal{X}| \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)|\mathcal{X}|} \\ &\leq \log \left(\sum_{x \in \mathcal{X}} p(x) \frac{1}{p(x)|\mathcal{X}|} \right) = 0 \end{aligned}$$

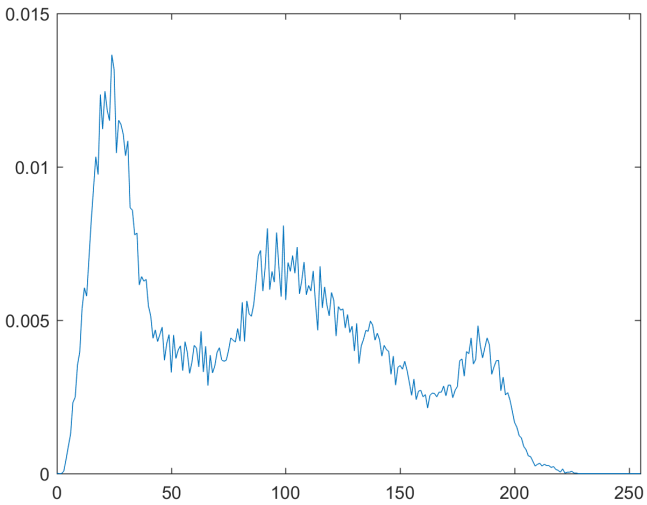
□

Exercise

Three squares have an average area $A = 100\,m^2$.
The average of the lenghts of their sides is $l = 10\,m$.
What can be said about the size of the largest of the three squares?

Hint: Consider that $f(x) = x^2$ is a convex function.

Application: Information in a B/W image



The pixels of a black and white image take values between 0 and 255, so 8 binary digits/pixel are needed to represent it.

If each pixel is considered an observation of a single random variable X, what is the average amount of information conveyed by X?

It is given by the entropy, that we can be computed from the normalized *histogram*:

$$H(X) = \sum_{x=0}^{255} h_x \log \frac{1}{h_x} = 7.519 \text{ bits/pixel}$$

where $h_x = \Pr(X = x) = p(x) = n_x/N$ is the number of pixels taking the value x over the total number of pixels N.

Application: Information in a B/W image

Let us binarize the image: set a threshold value and assign 0 or 255 to each pixels accordingly. The information contained in the image must be much lower, at most 1 bit/pixel.

Setting the threshold at $x = 120$, the entropy is 0.879 bits/pixel.

If the threshold is at $x = 160$, the entropy is 0.575 bits/pixel.

Why?

Joint entropy

Let X and Y be a pair of random variables with value sets $\mathcal{X} = \{x_1, \ldots, x_q\}$ and $\mathcal{Y} = \{y_1, \ldots, y_r\}$, joint probability distribution $p(x, y)$ and marginal probabilities $p(x)$ and $p(y)$.

The *joint entropy* of the two variables is

$$H(X, Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}.$$

It measures the total amount of information contained in the two variables together.

Analogously, the joint entropy can be defined for an n-tuple of random variables, $H(X_1, ..., X_n)$.

Conditional entropy

Assume the outcome x_i taken by the variable X. Consider the variable $Y|x_i$ with possible values $\mathcal{Y}$ and probabilities

$$\Pr(Y = y_j|X = x_i) = p(y_j|x_i) = \frac{p(x_i, y_j)}{p(x_i)}.$$

Its entropy is given by:

$$H(Y|x_i) = \sum_{y\in\mathcal{Y}} p(y|x_i) \log \frac{1}{p(y|x_i)}$$

In an analogous way one may consider the variable $X|y_j$.

Conditional entropy

The *conditional entropy* $H(Y|X)$ is defined as the average of the entropies $H(Y|x_i)$:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|x) \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x)p(y|x)\log\frac{1}{p(y|x)} \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(y|x)}. \end{aligned}$$

It measures the average amount of information in Y if X is known.

In general $H(X|Y) \neq H(Y|X)$. Can you derive a relation between them?

Conditional entropy: properties

The *chain rule* relates joint and conditional entropies:

Proposition (Chain rule for entropy)

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

Proof. By developing the definition of joint entropy. □

That is: the information conveyed by (X,Y) is the information of X plus the information of Y, assuming the value taken by X is known. In general:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$$

Theorem (Conditioning reduces uncertainty)

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

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

Conditional entropy: properties

Proof. It results from Jensen's inequality:

$$
\begin{aligned}
-H(X) + H(X|Y) &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x|y)} \\
&= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\
&\leq \log \left(\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \frac{p(x)p(y)}{p(x,y)} \right) = \log 1 = 0. \ \square
\end{aligned}
$$

Corollary

$$H(X,Y) \leq H(X) + H(Y)$$

Relative entropy

The relative entropy or *Kullback-Leibler divergence* of two distribution functions p, q defined in the domain $\mathcal{X}$ is defined as:

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

where, in order to guarantee the continuity the following conventions are adopted: $\log\left(\frac{0}{0}\right) = 0$, $0 \cdot \log\left(\frac{0}{q}\right) = 0$, $p \cdot \log\left(\frac{p}{0}\right) = \infty$

Theorem (Gibb's inequality)

$D(p||q) \geq 0$ *with equality iff* $p = q$

Proof. Use Jensen's inequality in the definition of $D(p||q)$.

Theorem (Asymmetry)

In general, $D(p||q) \neq D(q||p)$

Mutual information

The *mutual information* of Y with respect to X is defined as

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

that is, the mutual information is the amount of information that one of the variables in a pair conveys about the other.

One can easily check the formula

$$I(Y;X) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D\left(p(x,y)||p(x)p(y)\right)$$

As information is the relative entropy between a joint density and the product of marginals, it is a measure of independence between X and Y. The following properties hold

- Symmetry: $I(X;Y) = I(Y;X)$
- Positivity: $I(X;Y) \geq 0$
- $I(X;X) = H(X)$
- $I(X;Y) = 0$ if X and Y are independent.

Conditional mutual information

The *conditional mutual information* of Y with respect to X given Z is defined as:

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

which can be written as

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

from which the following properties hold

- Symmetry: $I(X;Y|Z) = I(Y;X|Z)$
- Positivity: $I(X;Y|Z) \geq 0$
- $I(X;X|Z) = H(X|Z)$
- $I(X;Y|Z) = 0$ if X and Y are independent conditioned to Z.

Chain rule for information

Proposition (Chain rule for information)

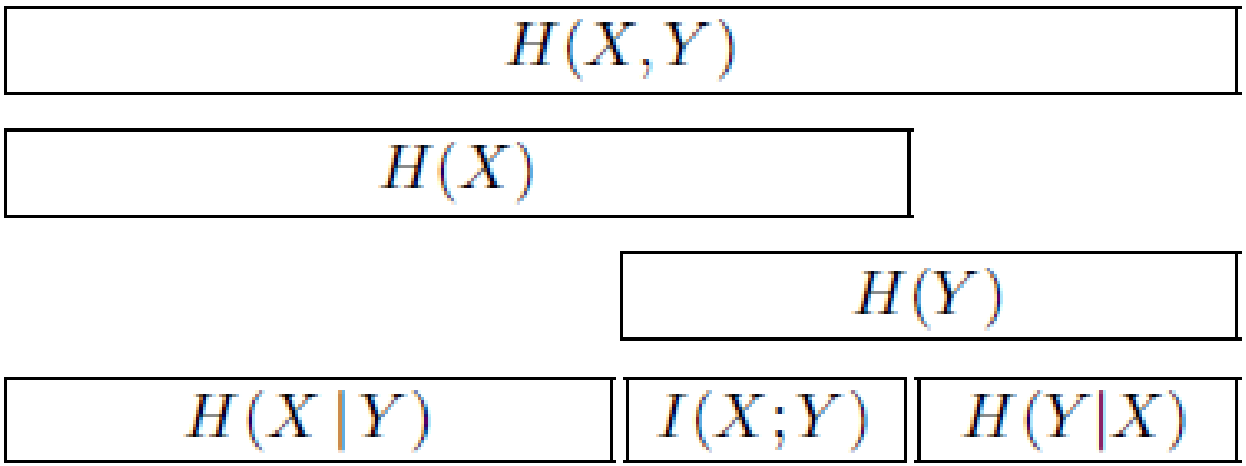
$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$

Proof.

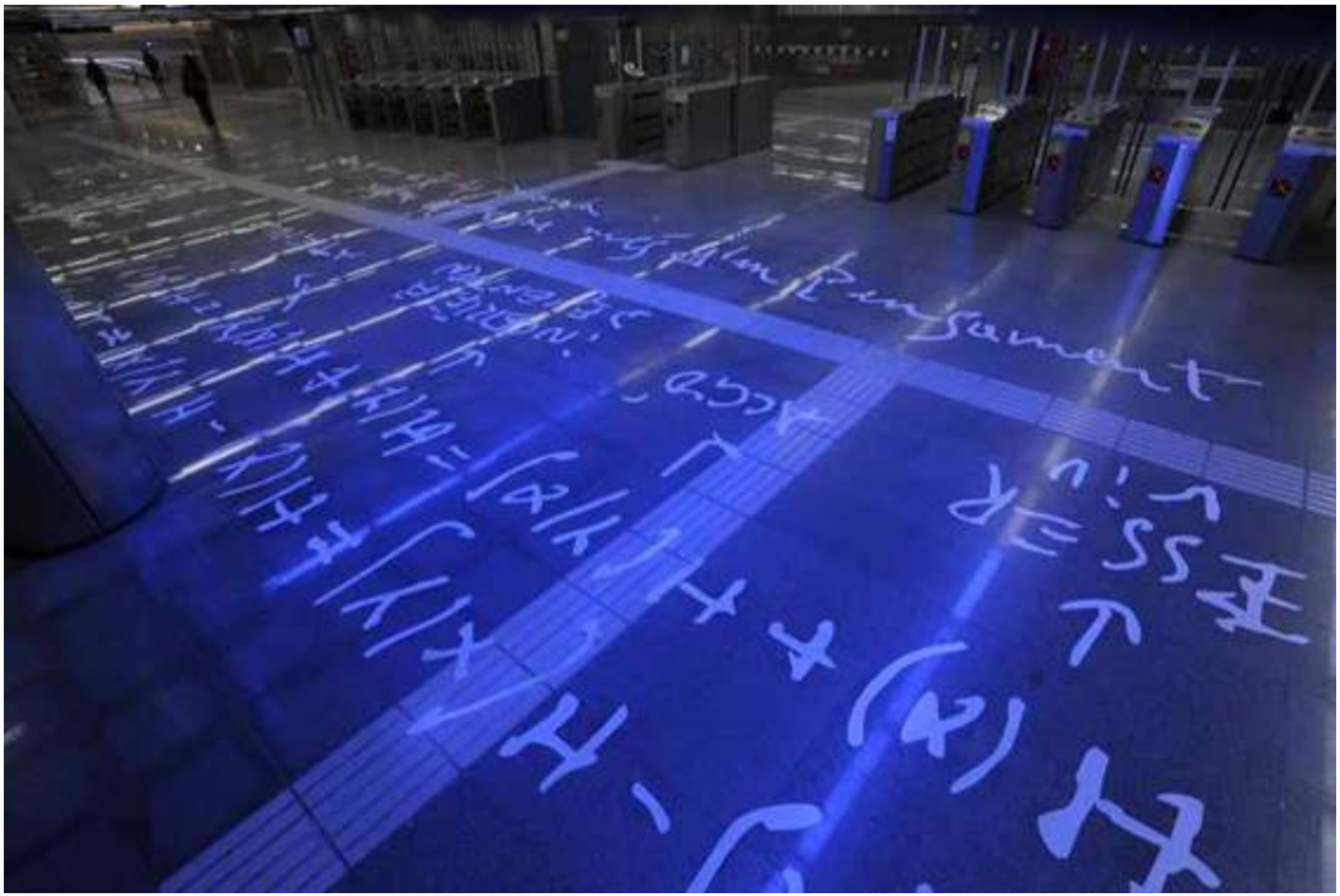
$$\begin{aligned} I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y) \\ &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y) \\ &= \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1) \end{aligned}$$

□

Relation between entropies and information

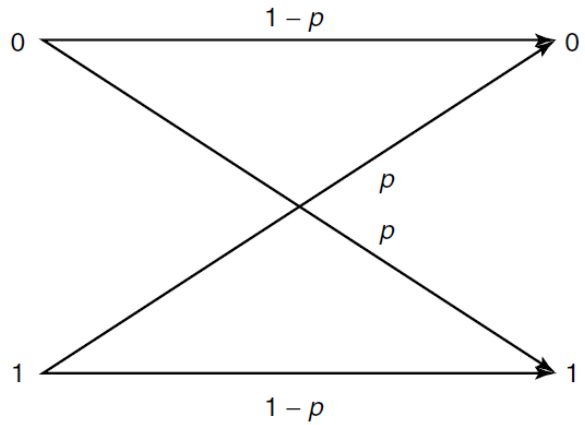


Where was this picture taken?



Application: mutual info in a binary symmetric channel

Let us evaluate the uncertainty in the determination of the binary input $X \in \{0, 1\}$ from the observations of the output of a noisy channel $Y \in \{0, 1\}$.



Input symbols are equally probable and the conditional distribution can be defined as

$p(y\|x)$	0	1
0	$1-p$	p
1	p	$1-p$

Application: mutual info in a binary symmetric channel

$$
\begin{aligned}
H(X) &= 1 \\
H(X|Y) &= \sum_{x,y\in\{0,1\}} p(x,y)\log\frac{1}{p(x|y)} \\
&= \sum_{x,y\in\{0,1\}} p(y|x)p(x)\log\frac{p(y)}{p(y|x)p(x)} \\
&= (1-p)\log\frac{1}{1-p} + p\log\frac{1}{p} = H(p)
\end{aligned}
$$

where $p(y)$ has been evaluated using $p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x)$.

Therefore, $I(X;Y) = 1 - H(p)$. Note that when $p = \frac{1}{2}$ no information about X can be extracted from Y. Otherwise stated, if $I(X;Y) = 0$ bits, nothing can be said about X when observing Y.

Data processing inequality

Let us assume that random variables X, Y, Z form *Markov chain* in that order $X \to Y \to Z$, which means that the conditional density of Z depends on Y, not on X:

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

that is, given Y, X and Z are independent. This entails that the joint probability is given by

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

As a consequence, no clever manipulation of data can improve the amount of information:

Theorem (Data processing inequality)

If $X \to Y \to Z$, *then* $I(X; Y) \geq I(X; Z)$

Data processing inequality

Proof. Using the definition of the mutual information based on entropies:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned}$$

The term $I(X;Z|Y) = 0$ because X is independent of Z given Y. Therefore:

$$I(X;Y|Z) = I(X;Y) - I(X;Z) \geq 0$$

and hence $I(X;Y) \geq I(X;Z)$. □

Note that if random variables do not form a Markov chain, it is possible that $I(X;Y|Z) > I(X;Y)$.

Exercise

Let X and Y be independent fair binary random variables. Let $Z = X + Y$.

- Check that the three variables do not form a Markov chain
- Check that $I(X;Y) = 0$
- Prove that $I(X;Y|Z) = \frac{1}{2}$ bit

Hint: use $H(X|Z) = \sum_{z_i \in \mathcal{Z}} p(z_i) H(X|Z = z_i)$

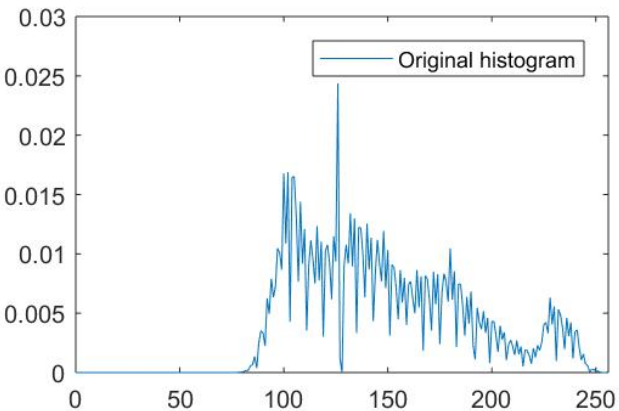
Application: Image equalization

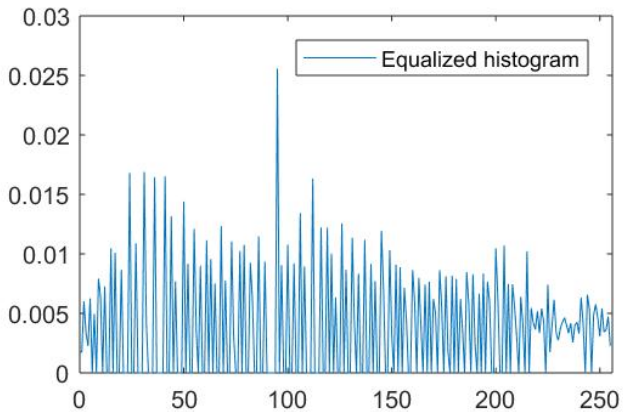
Image equalization is a procedure whereby the pixels of an image (considered as independent observations of a single random variable X) are transformed through a deterministic function g in such a way that the resulting image (with pixels Y) has a flatter normalized *histogram* than the original:

$$Y = g(X) = round\left(255 \sum_{x=0}^{X} h_x\right)$$

That is, the function $g(X)$ assigns to Y rounded values of the discrete integral of the histogram h_x of the original image.

Application: Image equalization





$H = 7.2$ bits/pixel
175 non-empty histogram bins

$H = 7.038$ bits/pixel
136 non-empty histogram bins

Application: Image equalization

On one side, the histogram is flatter and it may seem that the entropy should increase. However, there are many empty bins in the histogram so the uncertainty decreases! What is the overall effect on entropy?

Let us prove that entropy decreases:

$$\begin{aligned} H(X|Y) &= H(X) - I(X;Y) \\ &= H(X) - H(Y) + H(Y|X) \text{ by definition of } I(.) \\ &= H(X) - H(Y) \; g \text{ is deterministic, uncertainty on } Y \text{ given } X \text{ is zero} \\ &\geq 0 \text{ by positivity of entropy} \end{aligned}$$

Therefore, the entropy of any image processed with an injective function can only decrease.

Fano's inequality

Suppose we observe a random variable Y and wish to guess the value of a correlated random variable X. In the guessing of X, we use a (possibly random) estimation $\hat{X}$ taking values in $\hat{\mathcal{X}}$ (possibly different from $\mathcal{X}$). Note that $X \to Y \to \hat{X}$.

Intuitively, we may expect to estimate X with low error only if $H(X|Y)$ is small.

Theorem (Fano's inequality)

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, we can bound the probability of error $P_e = \Pr\left(X \neq \hat{X}\right)$ as

$$H(P_e) + P_e \log\left(|\mathcal{X}| - 1\right) \geq H(X|\hat{X}) \geq H(X|Y).$$

Fano's inequality

Proof. Define an error random variable $E = \begin{cases} 1 & \text{if} \quad \hat{X} \neq X \\ 0 & \text{if} \quad \hat{X} = X \end{cases}$

Use the chain rule to expand in two ways the conditioned entropy:

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X})$$

Note that in this expression:

- $H(E|X, \hat{X}) = 0$ since E is a function of both X and $\hat{X}$
- $H(E|\hat{X}) \leq H(E) = H(P_e)$ since conditioning reduces uncertainty
- $H(X|E, \hat{X}) = \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1)$
  $\leq P_e \cdot \log(|\mathcal{X}| - 1)$

and therefore $H(P_e) + P_e \cdot \log(|\mathcal{X}| - 1) \geq H(X|\hat{X})$.

By the data processing inequality, $I(X; \hat{X}) \leq I(X; Y)$ and hence $H(X|\hat{X}) \geq H(X|Y)$. □

Application: inference in the binary symmetric channel

The probability of error P_e is defined as the probability of $X \neq Y$,

$$P_e = \Pr(Y = 1|X = 0)\Pr(X = 0) + \Pr(Y = 0|X = 1)\Pr(X = 1)$$

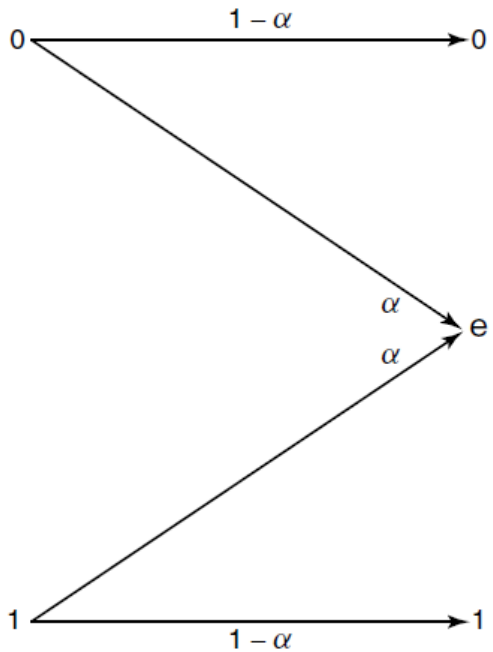
A bound on the probability of error is given by Fano's theorem: $H(p) \leq H(P_e)$.

$$\begin{array}{lccc} \text{If } p \leq \frac{1}{2} & P_e \geq p & \hat{X} = Y \rightarrow & P_e = p \\ \text{If } p \geq \frac{1}{2} & P_e \geq 1 - p & \hat{X} = 1 - Y \rightarrow & P_e = 1 - p \end{array}$$

The bound is tight in this case.

Application: inference in the binary erasure channel

Take the binary erasure channel with equally probable input symbols:



Let us relate the error in the estimation of X when Y is observed, with the conditional entropy. Using the Fano's bound, since $|\mathcal{X}| = 2$, then $H(X|Y) \leq H(P_e)$.

Application: inference in the binary erasure channel

The channel is characterized by the conditional distribution

$p(y\|x)$	0	1
0	$1-\alpha$	0
1	0	$1-\alpha$
e	α	α

The receiver decides $\hat{X}$ randomly 0 or 1 when $Y = \mathsf{e}$, so the true probability of error is given by

$$P_e = \Pr(Y = \mathsf{e}) \Pr(\hat{X} \neq X) = \frac{\alpha}{2}$$

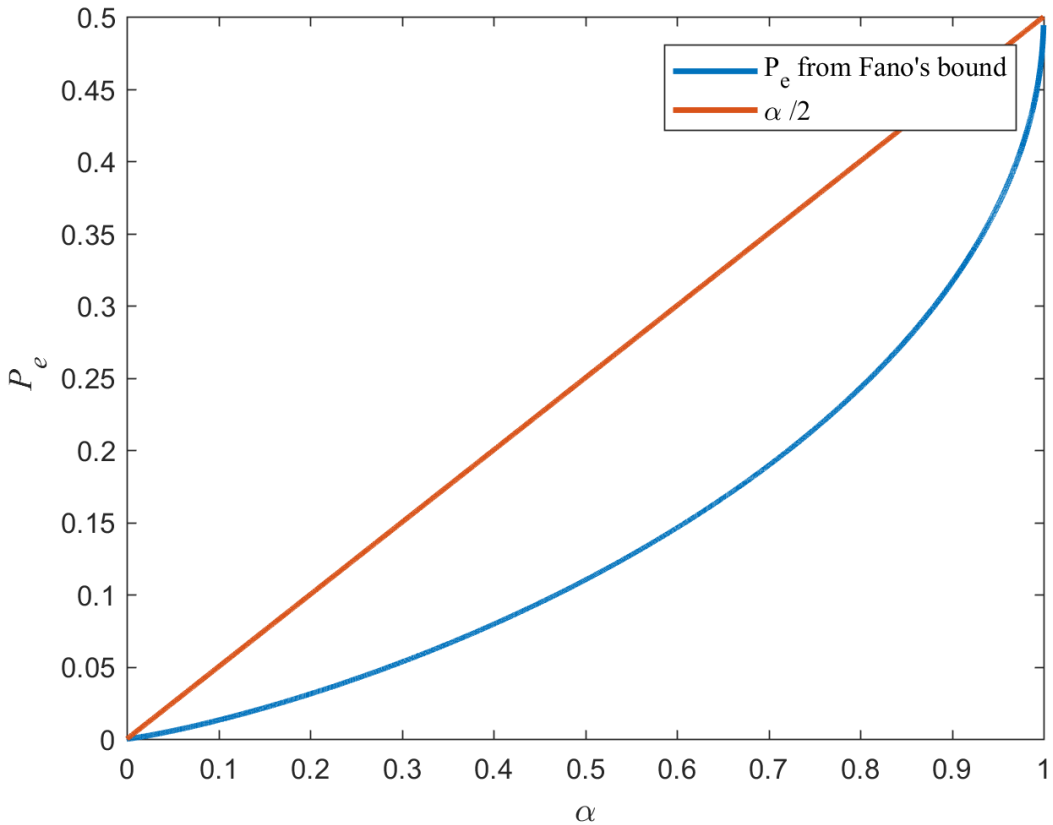
On the other hand, the conditional entropy is computed as

$$H(X|Y) = \sum_{y \in \{0,1,\mathsf{e}\}} p(y) H(X|y) = \Pr(Y = \mathsf{e}) = \alpha$$

Derive it!

Application: inference in a binary erasure channel

We now can plot the probability of error provided by the Fano's bound as a function of α



The bound is not tight this time.

Entropy of continuous random variables

What is the information conveyed by continuous random variables?

Consider a random variable X taking values in the domain $\mathcal{X}$, and its probability density function $f(x)$ with a countable number of discontinuities in $\mathcal{X}$ (then it is *Riemann integrable*). Let us divide $\mathcal{X}$ into bins of length Δ. By the mean value theorem there exist a value x_i within each bin such that

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

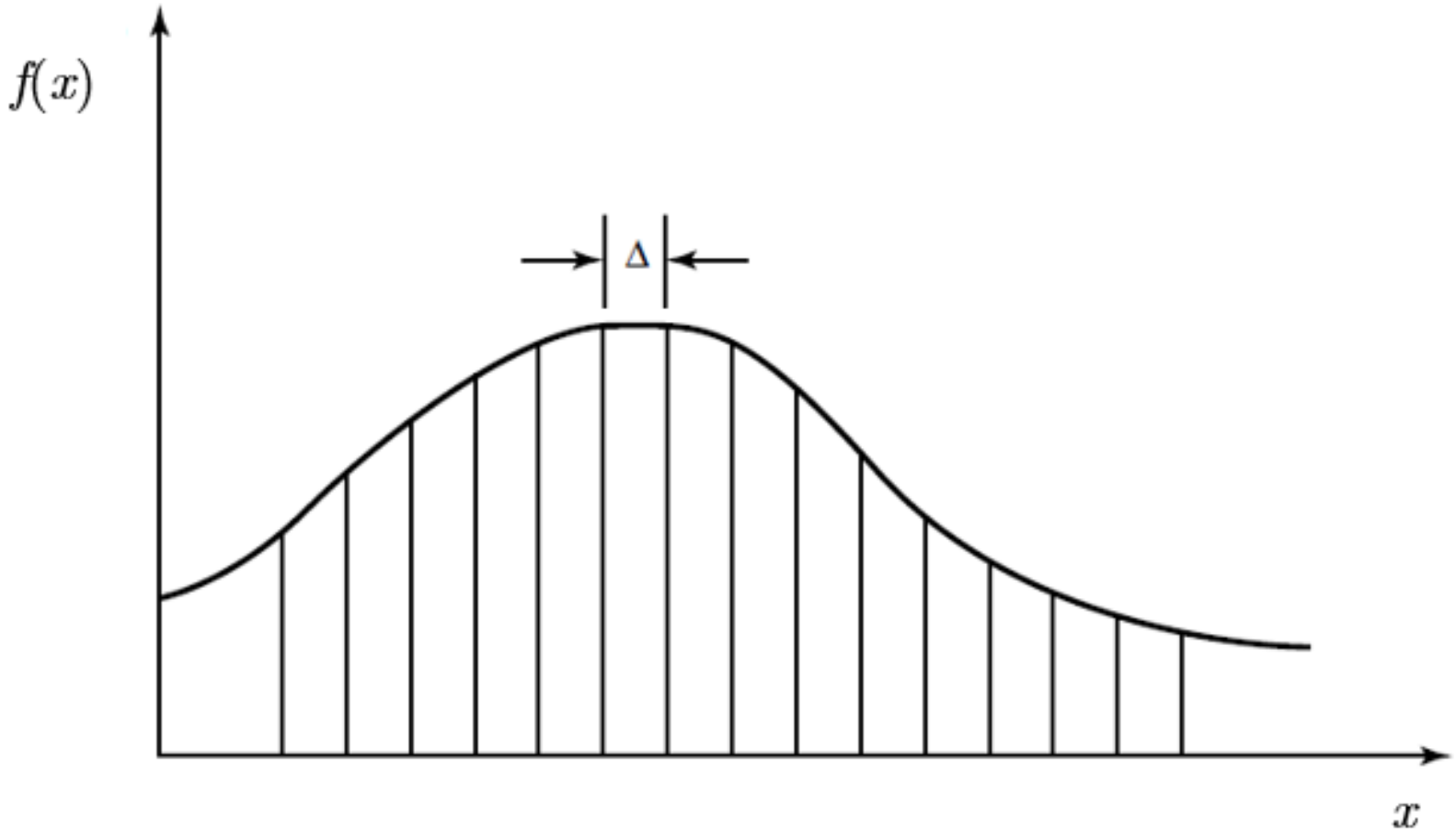
Define the quantized random variable X^Δ as

$$X^\Delta = x_i \quad \text{if } i\Delta \leq X < (i+1)\Delta$$

and hence its probability distribution is

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

Entropy of continuous random variables



Entropy of continuous random variables

The entropy of the quantized version of X is given by

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{i=-\infty}^{\infty} p(x_i)\log p(x_i) = -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log f(x_i) - \sum_{i=-\infty}^{\infty} f(x_i)\Delta \log \Delta \\ &= -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log f(x_i) - \log \Delta \end{aligned}$$

When $\Delta \to 0$ the second term tends to infinity, while the first term is the differential entropy if $f(x)$ is Riemann integrable

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx$$

We can interpret $h(X) + \log \Delta$ as the number of bits required on average to describe X to a certain accuracy given by Δ.

Entropy of continuous random variables

The differential entropy is specific of the continous-valued random variable X and it is measured in bits. However, it may be negative, which can be checked with a uniform random variable defined in the interval $[a, b]$.

Some other properties are:

Proposition (Translation invariance)

For any constant a, $h(X + a) = h(X)$

Proposition (Scaling)

For any non-zero constant a, $h(aX) = h(X) + \log |a|$

Entropy of continuous random variables

Accordingly we can define the conditional differential entropy as

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

and using $f(x,y) = f(y|x)f(x)$ it turns out $h(X|Y) = h(X,Y) - h(Y)$.

Similarly for the mutual information

$$I(X;Y) = \int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

From the definition we can write

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

Can you derive the mutual information of the quantized versions of X and Y?

How to estimate entropy?

Imagine we define the random variable X as the value of the pixel of any image containing a human face, and that we have a limited number of images (a total of N pixels) to characterize $H(X)$. The value of $H(X)$ computed from the observations depends on the set of images available, and hence it will be a random number.

Take n_x as the number observed of samples for each possible value of X. The distribution of n_x is *binomial* of parameter $p(x)$, and $h_x = n_x/N$ is its unbiased estimator of the distribution function, that is $E[h_x] = p(x)$.

Theorem (Bias of entropy estimator)

The plug-in estimator $\hat{H}_N = \sum_{x\in\mathcal{X}} h_x \log \frac{1}{h_x}$ underestimates in average the true entropy.

Proof. Take the concave function $f(x) = -x\log x$ and let us apply Jensen's inequality:

$$H = \sum_{x\in\mathcal{X}} f\left(p(x)\right) = \sum_{x\in\mathcal{X}} f\left(E\left[h_x\right]\right) \geq \sum_{x\in\mathcal{X}} E\left[f(h_x)\right] = E\left[\sum_{x\in\mathcal{X}} f\left(h_x\right)\right] = E\left[\hat{H}_N\right]$$

How to estimate entropy?

This does not imply that each particular finite sample estimate is below the true entropy however. The bias has been proven to be

$$H - E\left[\hat{H}_N\right] = \frac{|\mathcal{X}| - 1}{2N} - \frac{1}{12N^2}\left(1 - \sum_{x \in \mathcal{X}} \frac{1}{p(x)}\right) + \mathcal{O}(N^{-3}) \geq 0$$

so a modified estimator was proposed by Miller:

$$\hat{H}_N = \sum_{x \in \mathcal{X}} h_x \log \frac{1}{h_x} + \frac{|\mathcal{X}| - 1}{2N}$$

G. Miller, "Note on the bias of information estimates", in H. Quastler (Ed.), Information Theory in Psychology II-B, Glencoe, IL: Free Press, 95–100, 1955