Entropy and Mutual Information

(Discrete Random Variables)

Master Universitario en Ingeniería de Telecomunicación

I. Santamaría

Universidad de Cantabria

Contents

Introduction

Entropy

Joint Entropy and Conditional Entropy

Relative Entropy

Mutual Information

Jensen's inequality

Entropy and mutual information are key concepts in IT

- **Entropy**
  - The entropy $H(X)$ of a random variable X gives us the fundamental limit for data compression
  - A source producing i.i.d. realizations of X can be compressed up to $H(X)$ bits/realization
  - The entropy is the average shortest description of X

- **Mutual information**
  - The mutual information gives us the fundamental limit for transmission
  - The **capacity** of a channel is given by

$$C = \max_{p(x)} I(X;Y)$$

Definitions

Let X be a discrete random variable (r.v.) that takes values $x \in \mathcal{X}$. The probability mass function (pmf) of X will be denoted by

$$p(x) = Pr\{X = x\}$$

Example 1: Bernoulli r.v. $\mathcal{X} = \{0, 1\}$

$$Pr\{X = 1\} = p, \qquad Pr\{X = 0\} = 1 - p$$

Example 2: Binomial r.v. $X \sim \mathrm{B}(n, p)$ $(\mathcal{X} = \{0, 1, \ldots, n\})$

$$Pr\{X = k\} = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Note that $X \sim \mathrm{B}(1, p)$ is a Bernoulli r.v.
- If $X_1, \ldots, X_n$ are independent $\mathrm{B}(1, p)$, then $Y = \sum_{k=1}^{n} X_k \sim \mathrm{B}(n, p)$

Information

For a discrete random variable, X, the information of an outcome $X = x$ is

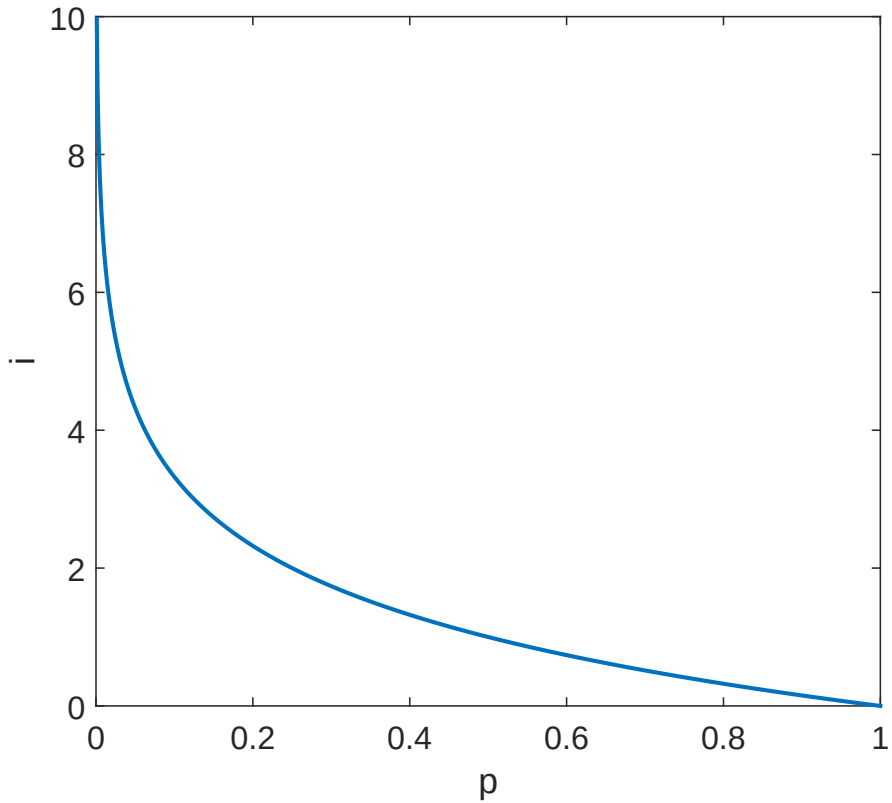
$$i(x) = -\log(p(x))$$

and it is measured in bits[1]

- Why the log of a probability measures information? Not a simple question, but there are strong reasons
  - Less probable outcomes are the most informative ones
  - For independent random variables: $i(x, y) = -\log(p(x, y)) = -\log(p(x)p(y)) = i(x) + i(y)$
- The outcome of a random experiment is most informative if the pmf is uniform $\rightarrow$ Experiment design, guessing games,...

[1] Unless otherwise indicated, in this course log denotes logarithm in base 2.

If X is Bernoulli with $Pr\{X = 1\} = p$, the information of the outcome $X = 1$ is $i = -\log(p)$



The information is always a positive quantity

Entropy

Definition: The entropy of a discrete random variable, X, is defined by[2]

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

- For discrete random variables $H(X) \geq 0$
- It is the average information of the random variable X:

$$H(X) = E[i(X)],$$

note that this can interpreted also as the mean value of a new (transformed) r.v. $Y = -\log(p(X))$

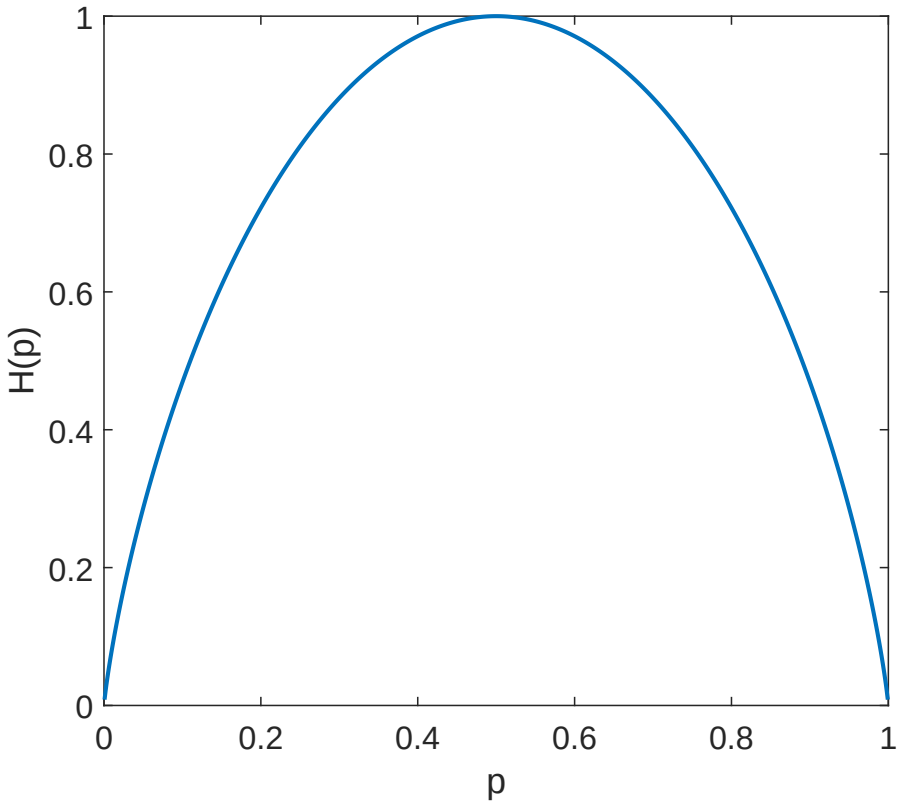
[2] We assume for convention that $0 \log(0) = 0$.

Interpretation

- Entropy is the measure of average uncertainty in X
- Entropy is the average number of bits needed to describe X
- Entropy is a lower bound on the average length of the shortest description of X

Example 1: Entropy of a Bernoulli r.v. with parameter p

$$H(X) = -p\log(p) - (1-p)\log(1-p) \triangleq H(p)$$



It is a concave function of p

Example 2: Entropy of a uniform r.v. taking on K values: e.g., $\mathcal{X} = (1, \ldots, K)$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) = \sum_{i=1}^{K} \frac{1}{K} \log(K) = \log(K)$$

- It does not depend of the values that X takes, it depends only on their probabilities (X and $X + a$ have the same entropy!)
- **Property**: For an arbitrary discrete r.v. $H(X) \leq \log(|\mathcal{X}|)$,where $|\mathcal{X}|$ denotes the cardinality of the set, and $H(X) = \log(|\mathcal{X}|)$ iff X has a uniform distribution over $\mathcal{X}$

  That is to say: $H(X)$ is a lower bound on the number of binary questions (bits) that are always guaranteed to identify an outcome from the ensemble $\mathcal{X}$

Example 3: The entropy of English

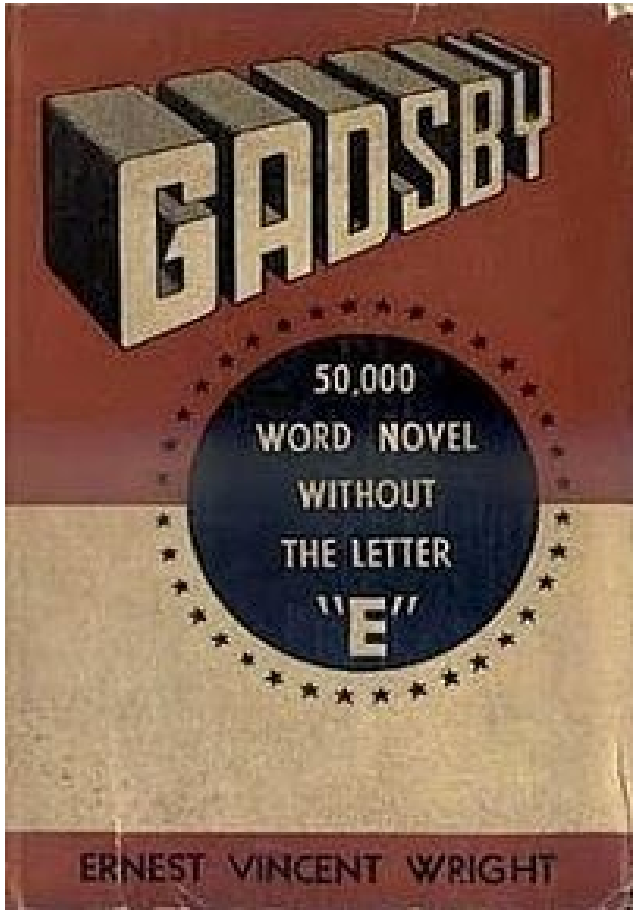
- Probabilities estimated from *The Frequently Asked Questions Manual for Linux*[a]
- 26 letters (a-z) and a space character '-'
- The entropy is

$$H = -\sum_i p_i \log(p_i) = 4.11 \; bits/letter$$

- The most informative letter is z: 11.48 bits (frequency of appearance 0.1%)
- The least informative letter is e: 3.5 bits (frequency of appearance 13%)
- Is it possible to write a book in English without using the letter e?

[a] Reprinted from MacKay's textbook

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



If youth, throughout all history, had had a champion to stand up for it; to show a doubting world that a child can think; and, possibly, do it practically; you wouldn't constantly run across folks today who claim that "a child don't know anything." A child's brain starts functioning at birth; and has, amongst its many infant convolutions, thousands of dormant atoms, into which God has put a mystic possibility for noticing an adult's act, and figuring out its purport.

Definitions

The **joint p.m.f** of two random variables X and Y taking values on alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, is

$$p(x,y) = Pr\{X = x, Y = y\}, \qquad (x,y) \in \mathcal{X} \times \mathcal{Y}$$

If $p(x) = Pr\{X = x\} > 0$, the **conditional probability** of $Y = y$ given that $X = x$ is defined as

$$p(y|x) = Pr\{Y = y \,|X = x\} = \frac{p(x,y)}{p(x)}$$

Independence

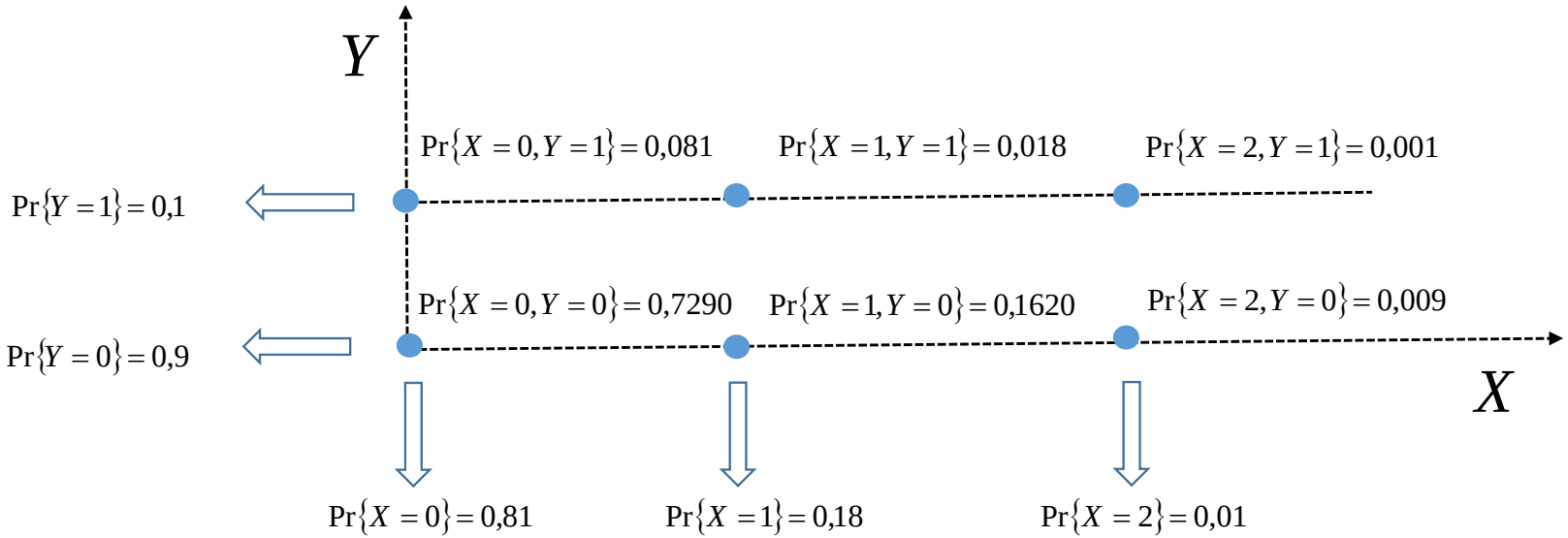
- The events $X = x$ and $Y = y$ are independent if

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

- The random variables X and Y are independent if

$$p(x,y) = p(x)p(y), \qquad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$$

Example



- You can easily check that X and Y are independent
- Actually, $X \sim \mathrm{B}(2, 0.1)$ and $Y \sim \mathrm{B}(1, 0.1)$

How is $Z = X + Y$ distributed?

Joint entropy and conditional entropy

Definition: The joint entropy, $H(X, Y)$ of two random variables (X, Y) with pmf $p(x, y)$ is defined as

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

Definition: The conditional entropy of Y given X is defined as

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

Note that

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

Chain rule

We know that $p(x, y) = p(x)p(y|x)$; therefore, taking logarithms and expectations on both sides we arrive to

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

and the so-called **chain rule** for conditional entropy follows

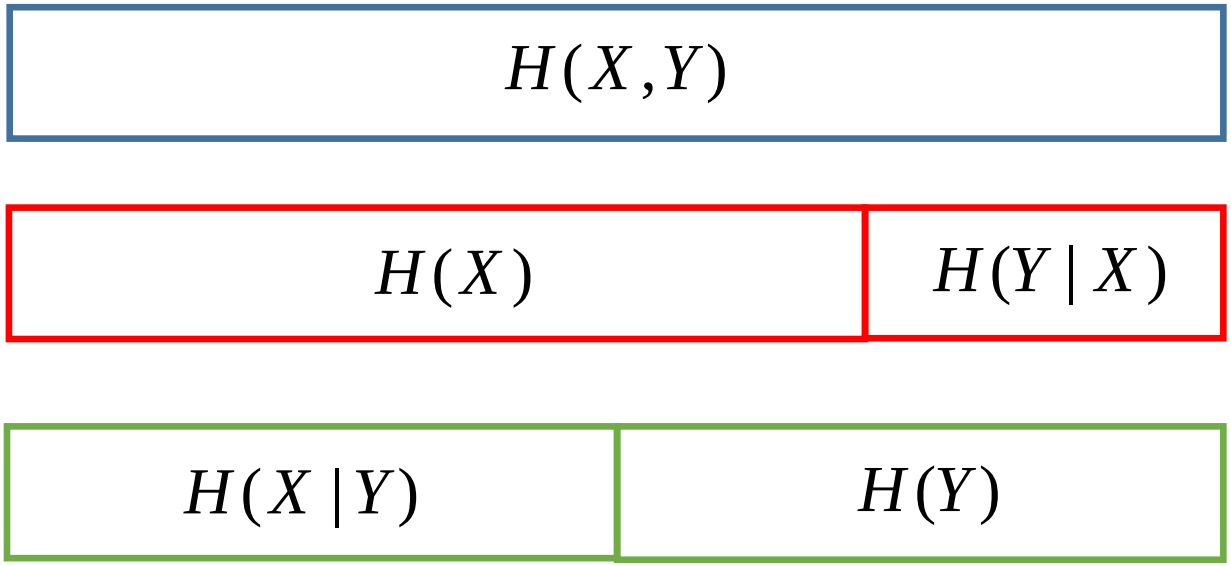
$$H(X, Y) = H(X) + H(Y|X)$$

Similarly, we have

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

Note that $H(Y|X) \neq H(X|Y)$, but

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



As a corollary of the chain rule, it is easy to prove the following

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

Generalization of the chain rule for entropy: Let $X_1, X_2, \ldots, X_n$ be a collection of random variables with joint pmf $p(x_1, x_2, \ldots, x_n)$. Then

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

The entropy is a sum of conditional entropies, for instance

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1)$$

Example (from Cover's textbook): Let (X,Y) have the following joint pmf

$p(x,y)$	$x=1$	$x=2$	$x=3$	$x=4$
$y=1$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
$y=2$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
$y=3$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
$y=4$	$\frac{1}{4}$	0	0	0

- $H(X) = 7/4$
- $H(Y) = 2$
- $H(X|Y) = 11/8$
- $H(Y|X) = 13/8$
- $H(X,Y) = 27/8$

Relative entropy, K-L divergence

Definition: The relative entropy or Kullback-Leibler divergence between two distributions, $p(x)$ and $q(x)$, taking values on the same alphabet is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \left[\log \frac{p(X)}{q(X)} \right]$$

We use the convention that $0 \log \frac{0}{0} = 0$, $p \log \frac{p}{0} = \infty$ and $0 \log \frac{0}{q} = 0$

- It is a measure of the "distance" between the two distributions
  - $D(p||q) \geq 0$
  - $D(p||q) = 0$ iff $p = q$
- However, it is not a true distance

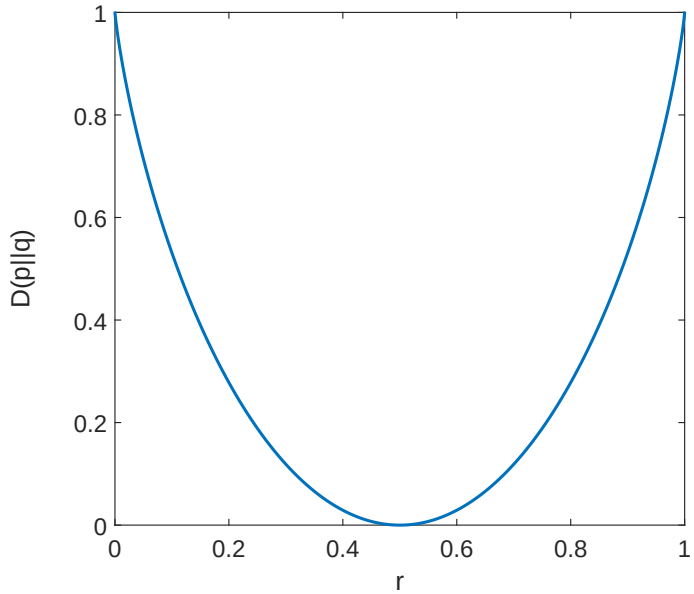
$$D(p||q) \neq D(q||p)$$

Example: Consider two different Bernoulli distributions:
$p(x) = (1-r)\delta(x) + r\delta(x-1)$ and $q(x) = (1-s)\delta(x) + s\delta(x-1)$

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

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

If we fix $s = 0.5$ and vary r from 0 to 1, $D(p||q)$ looks like this:



Mutual information

Definition 1: The mutual information $I(X;Y)$ between the random variables X and Y is given by

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

- It is the reduction in the uncertainty of X due to the knowledge of Y
- $I(X;Y) \geq 0$, which implies

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

  i.e., "information cannot hurt"
- $I(X;Y) = 0$ iff X and Y are independent

Definition 2: The mutual information $I(X;Y)$ can be defined as the relative entropy between the joint distribution $p(x,y)$ and the product of marginals $p(x)p(y)$

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

Properties

- It is symmetric (X says as much about Y as Y says about X)

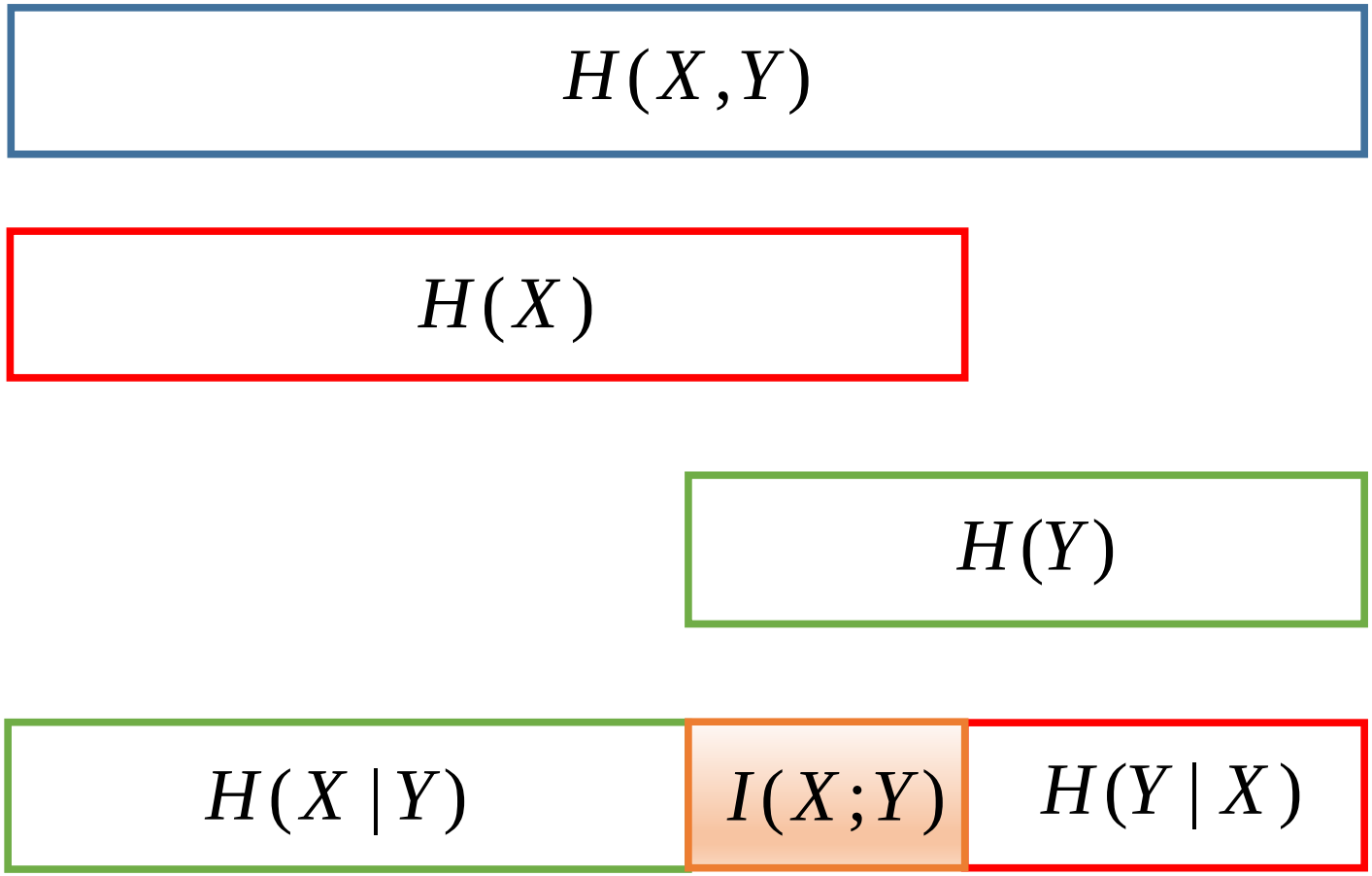
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \end{aligned}$$

- The mutual information of a random variable with itself is the entropy (**self-information**)

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

- It also satisfies

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


$H(X,Y)$
$H(X)$
$H(Y)$
$H(X \mid Y)$
$I(X;Y)$
$H(Y \mid X)$

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

$p(x,y)$	$x=1$	$x=2$	$x=3$	$x=4$
$y=1$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
$y=2$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
$y=3$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
$y=4$	$\frac{1}{4}$	0	0	0

- $I(X;Y) = 3/8$

Chain rule for mutual information

$$I(X_1, X_2, X_3; Y) = I(X_1; Y) + I(X_2; Y|X_1) + I(X_3; Y|X_2, X_1)$$

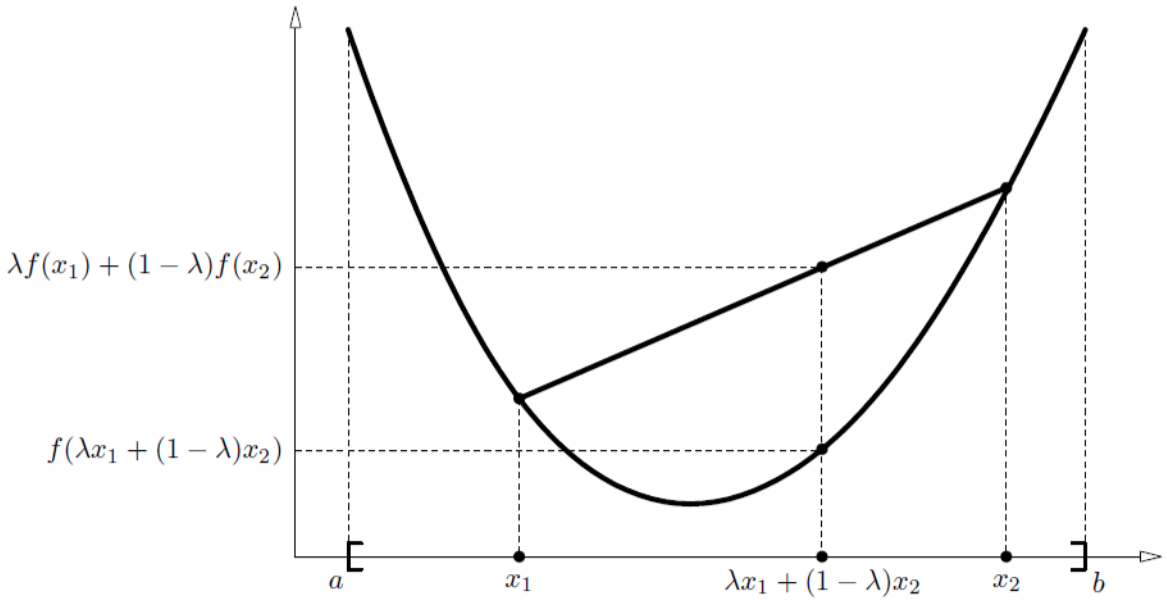
The proof follows from the definition of mutual information and the application of the chain rule for entropy

More generally, we can write

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

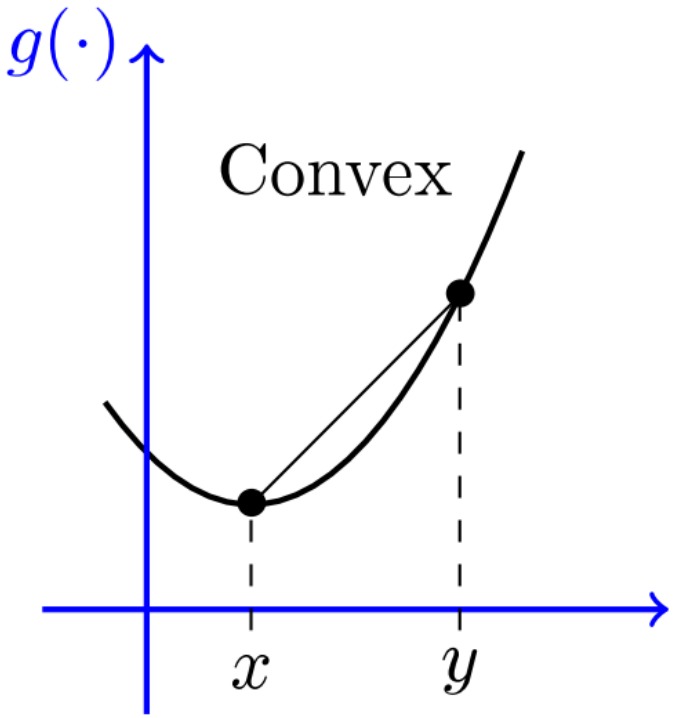
Jensen's inequality

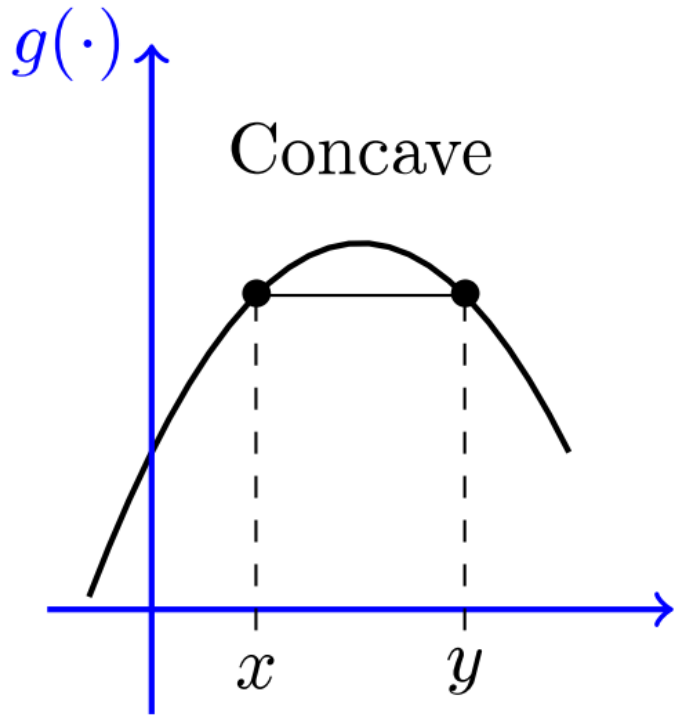
Jensen's inequality generalizes the idea that, for **convex** functions, the secant line lies above the graph



Many fundamental inequalities in information theory are consequences of Jensen's inequality, the most important one being

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





- Convex: x, x^2, $|x|$, e^x, $x\log(x)$ (for $x \geq 0$), ...
- Concave: x, $\log(x)$, $\sqrt{x}$ (for $x \geq 0$), ...

Definition: A function $f(x)$ is said to be convex on an interval (a, b) if, $\forall x_1, x_2 \in (a, b)$, $0 < \lambda < 1$

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

The function is **strictly convex** if equality holds only if $\lambda = 0$ or $\lambda = 1$

- A function $f(x)$ is concave if $-f(x)$ is convex
- If the function $f(x)$ has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval

Jensen's inequality

Theorem: If $f(x)$ is a convex function and X is a random variable

$$E[f(X)] \geq f(E[X])$$

Corollary: If $f(x)$ is strictly convex

$$E[f(X)] = f(E[X]) \Longrightarrow X = E[X] \Longrightarrow X \text{ is a constant}$$

Jensen's inequality follows directly from the fact that for a set of numbers $x_1, x_2, \ldots, x_n$ and a set of positive weights $a_1, a_2, \ldots, a_n$, a convex function satisfies

$$f\left(\frac{\sum_i a_i x_i}{\sum_j a_j}\right) \leq \frac{\sum_i a_i f(x_i)}{\sum_j a_j}$$

which can be easily proved by induction
If $p_i = \frac{a_i}{\sum_j a_j}$ is viewed as the pmf of a random variable X, the probabilistic form of Jensen's inequality follows

An application

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

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

where we have applied Jensen's inequality in the second line since $\log(x)$ is a strictly concave function