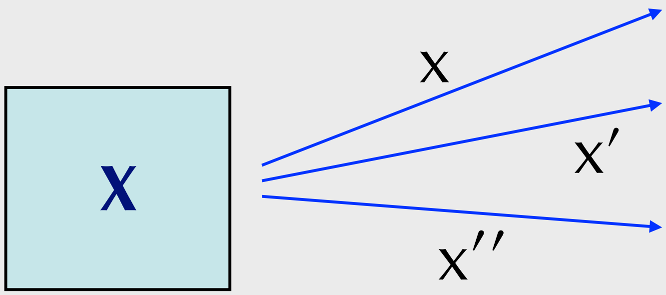
Entropy

Chapter 2, Elements of Information Theory

Discrete Random Variables

A discrete random variable $X \sim p(x)$ 'produces' letters $x \in \mathcal{X}$ from a countable (typically finite) alphabet $\mathcal{X}$ with probability mass function $p: \mathcal{X} \rightarrow \mathbb{R}$.



$p(x) = \Pr\{X=x\}$ with $x \in \mathcal{X}$

$p(x') = \Pr\{X=x'\}$ with $x' \in \mathcal{X}$

$p(x'') = \Pr\{X=x''\}$ with $x'' \in \mathcal{X}$

If we have several random variables we should write $p_X(x)$, $p_Y(y)$ and so on, but often we allow ourselves to drop the subscript and simply write $p(x)$, $p(y)$,…

Joint, Marginal Probabilities

If we have two or more random variables, then we can consider the *joint* and the *marginal distributions*.

For two random variables (X,Y) we have the joint distribution
$p(x,y) = \Pr\{x=X,y=Y\}$, such that $(X,Y) \sim p(x,y)$.

The marginal distributions $X{\sim}p(x)$ and $Y{\sim}p(y)$ are:
$p(x) = \Pr\{x=X\} = \Sigma_y\, p(x,y)$ and
$p(y) = \Pr\{y=Y\} = \Sigma_x\, p(x,y)$

The variables X and Y are *independent* if and only if
$p(x,y) = p(x) \cdot p(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Expectations

The fact that the random variable X has probability mass function p, is summarized by $X \sim p(x)$.

For a function g on $\mathcal{X}$ we can also look at the random variable g(X), which has the expected value
$E_p\, g(X) = \Sigma_x\, p(x) \cdot g(x)$ or simply "$E\, g(X)$".

Note for E g(X) to be meaningful, the range of g must allow multiplication by the reals p(x) and addition.

Bayes' Rule

Because we have $p(x,y) = p(y) \cdot p(x|y) = p(x) \cdot p(y|x)$ it holds that:

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

We call $p(y)$ the prior distribution, and $p(y|x)$ the posterior distribution (after having observed $X=x$).

Note that indeed $\Sigma_y\, p(y|x) = 1$ (using $\Sigma_y\, p(x,y)=p(x)$).

Entropy

It will be crucial to be able to quantify the amount of randomness of a probability distribution.

Definition: The entropy H(X) of a discrete random variable X is defined by (also denoted H(p)):

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

The entropy of a distribution is expressed in bits.

Note that because $\lim_{p \to 0} p \log p = 0$, the 'empty probabilities' $p(x)=0$ do not contribute to the entropy.

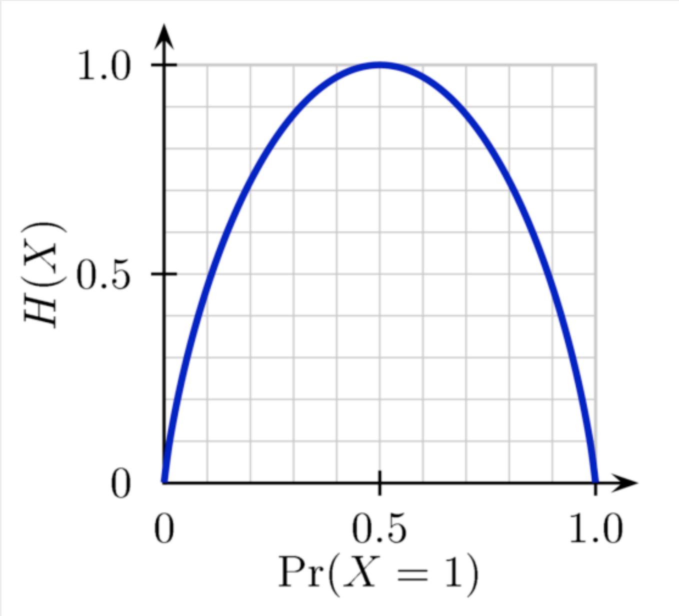
Entropy of a Bit

A completely random bit with p=(½,½) has
$H(p) = -(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}) = -(-\frac{1}{2} + -\frac{1}{2}) = 1.$

A deterministic bit with p=(1,0) has
$H(p) = -(1 \log 1 + 0 \log 0) = -(0+0) = 0.$

A biased bit with p=(0.1,0.9) has H(p) = 0.468996…

In general, the entropy looks as follows as a function of $0 \leq \Pr\{X=1\} \leq 1$:



Some Properties of H

Lemma 2.1: We always have $H(X)\geq 0$. Why?

$H(X)=0$ if and only if X is a 'deterministic variable' with $p(x)=1$ for one specific value $x\in\mathcal{X}$.

If $p(x) = 1/D$ for D different values $x\in\mathcal{X}$, then $H(X) = \log D$.

$H(X) \leq \log(\text{number of } x\in\mathcal{X} \text{ with } p(x)>0)$

You can view H as the expectation of $\log 1/p(x)$:
$H(X) = -\Sigma_x\, p(x) \log p(x) = E_p \log 1/p(X)$.

It measures the expected 'surprise' $\log 1/p(x)$.

Interpretation of Entropy

H(X) = "Expected surprise" = "Expected amount of information gain" when learning the $x \in \mathcal{X}$ value of a random variable X".

H(X) = "Expected number of required Yes/No questions to learn the value $x \in \mathcal{X}$ of a random variable X.

Asymptotic Equipartition Property (AEP), informally: When repeating X n times and n is big, the probability distribution $p(X^n)$ tends towards a uniform distribution over a typical set of size $2^{nH(X)}$ with typical probability $2^{-nH(X)}$ for each element.

History of Entropy

Historically, entropy was used before Shannon in the context of thermo-dynamics in the equality S = k ln W, where k is Boltzmann’s constant 1.38×10^{-23} Joule/Kelvin, W is the size of the state space of the system and S is its entropy.

Meaning of Entropy

"You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

— John von Neumann writing to Claude Shannon

Joint Entropy

If we have a two random variables $(X,Y) \sim p(x,y)$ with $p(x,y) = \Pr\{x=X, y=Y\}$, their joint entropy equals

$$H(X,Y) = -\Sigma_x \Sigma_y\, p(x,y) \log p(x,y),$$

which is equivalent with $H(X,Y) = -\, E_p \log p(X,Y)$.

For independent distributions with $p(x,y) = p(x)p(y)$ we have $H(X,Y) = H(X) + H(Y)$.

If X and Y are dependent then $H(X,Y) < H(X) + H(Y)$.

In fact, $H(X,Y) = H(X) + \Sigma_x\, p(x)\, H(Y|X=x)$.

Conditional Entropy

The expected entropy of Y after we have observed a value $x \in X$, is called the conditional entropy $H(Y|X)$:

$$\begin{aligned} H(Y|X) &= \sum_x p(x) \cdot H(Y|X=x) \\ &= -\sum_x p(x) \cdot \sum_y p(y|x) \log p(y|x) \\ &= -\sum_{x,y} p(x,y) \log p(y|x) \\ &= -E_{p(x,y)} \log p(Y|X) \end{aligned}$$

Chain rule: $H(X,Y) = H(X)+H(Y\,|\,X) = H(Y)+H(X\,|\,Y)$.

Example of H(X | Y)

Take p(X) over {0,…,500} with p = (½,1/1000,…,1/1000)
with entropy H(X) = ½ + ½·log 1000 ≈ 4.983 bits.

Take Y with $\mathcal{Y}$ = {"x=0","x≠0"}.

If we 'learn' that x is not 0, we increase the entropy:
p(x|x≠0) = (0,1/500,…,1/500) with H(X|x≠0) ≈ 8.966.

We learned information, yet the entropy increased?

Think: Not finding your wallet in the likely place.

The expected uncertainty (=conditional entropy) goes down:
H(X|Y) = ½ H(X|x=0) + ½ H(X|x≠0) ≈ 4.483.

Chain Rule for Entropy

For random variables $X_1,\ldots,X_n$ we have the Chain rule:

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

Think: the amount of information that you obtain by observing $X_1,\ldots,X_n$ equals the X_1 information $H(X_1)$, plus the additional X_2 information $H(X_2|X_1)$, et cetera.

Notice also the similarity with the multiplicative rules for joint probabilities: $p(x,y) = p(x)\cdot p(y|x)$.

About Conditional Entropy

Entropy: $H(X)$, $H(Y)$
Joint entropy: $H(X,Y)$
Conditional entropy: $H(Y|X) = \Sigma_x \, p(x) \cdot H(Y|X{=}x)$

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

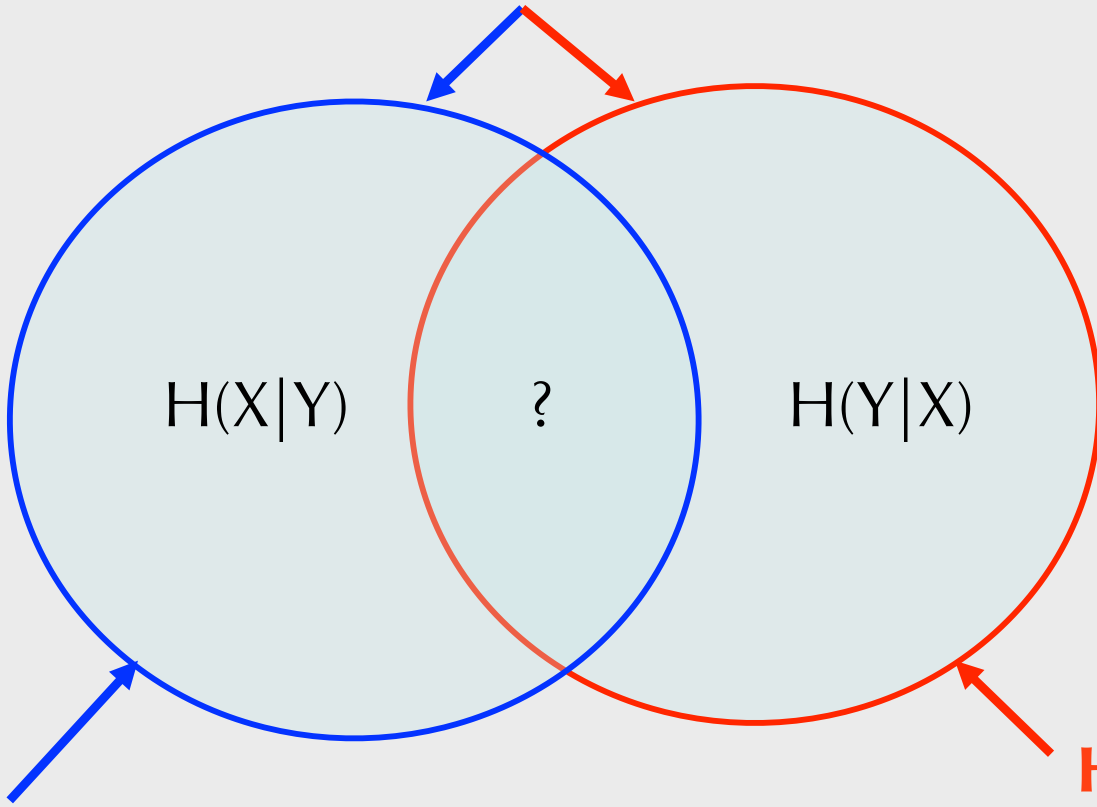
If X and Y are independent, then $H(X,Y) = H(X)+H(Y)$, hence then $H(Y|X) = H(Y)$.

In general: $0 \leqslant H(Y|X) \leqslant H(Y)$.

Possible asymmetry: $H(Y|X) - H(X|Y) = H(Y) - H(X)$

A Missing Piece

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



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

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

$$? = H(X) + H(Y) - H(X,Y)$$

Mutual Information

For two variables X,Y the mutual information I(X;Y) is the amount of certainty regarding X that we learned after observing Y. Hence $I(X;Y) = H(X)-H(X|Y)$.

Note how X and Y are symmetric:

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

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

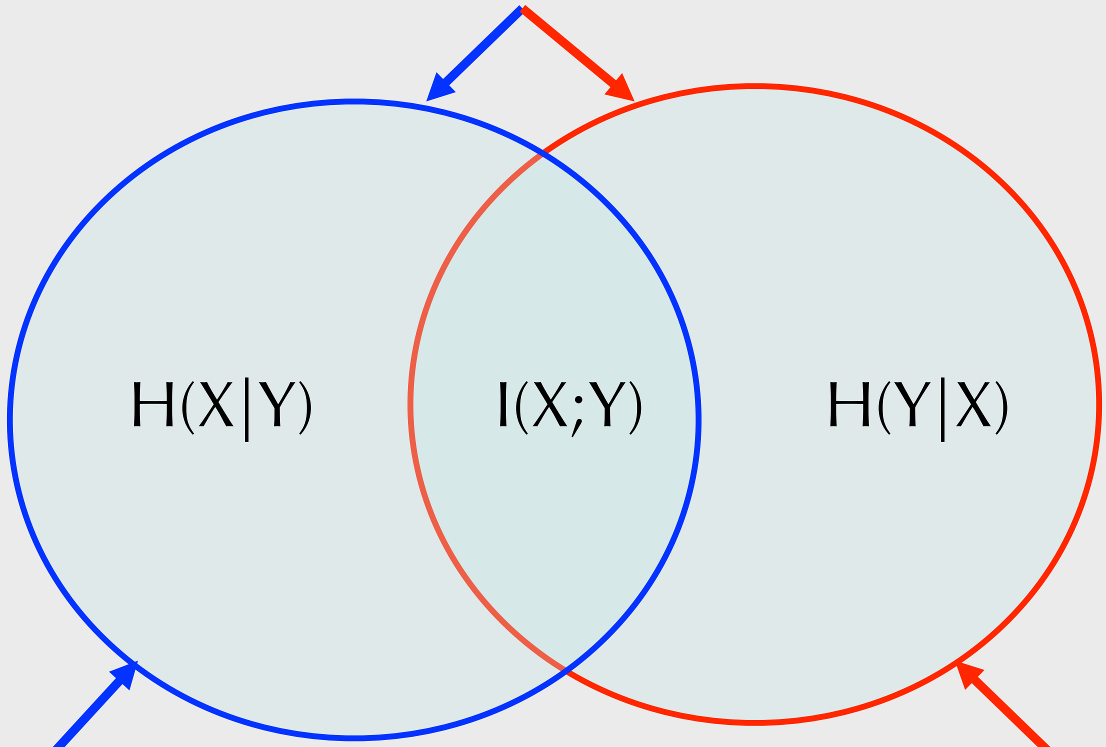
Also:

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

Think of I(X;Y) as the 'overlap' between X and Y; it is 0 if and only if X and Y are independent.

4 Pieces

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



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

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

About Mutual Information

Mutual information is the central notion in information theory. It quantifies how much we learn about X by observing Y.

When X and Y are the same we get: $I(X;X) = H(X)$, hence entropy is called 'self information'.

Expectation of What?

Mutual information can be viewed as an expectation:

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

This function is called the relative entropy between the probabilities p(x,y) and p(x)p(y) on $\mathcal{X} \times \mathcal{Y}$.

Relative Entropy

The *relative entropy* or Kullback-Leibler distance between two distributions p and q is defined by

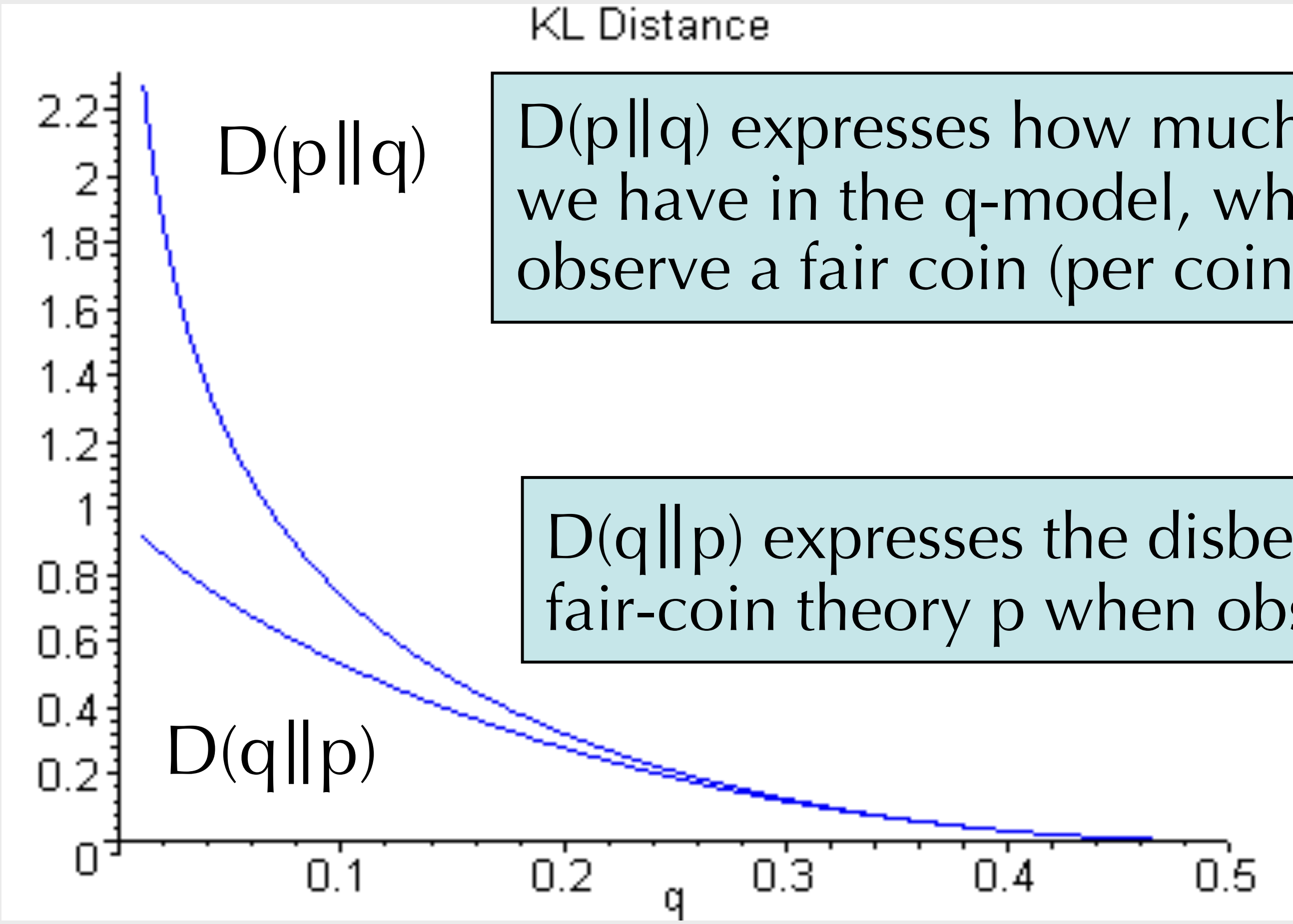
$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$
$$= E_p \log \frac{p(X)}{q(X)}$$

This 'distance' expresses our expected disbelief in q when we are observing probability distribution p.

It is not a true distance as $D(p\|q) \neq D(q\|p)$.
Also, $D(p\|q)$ can be infinite.

Example KL Distance

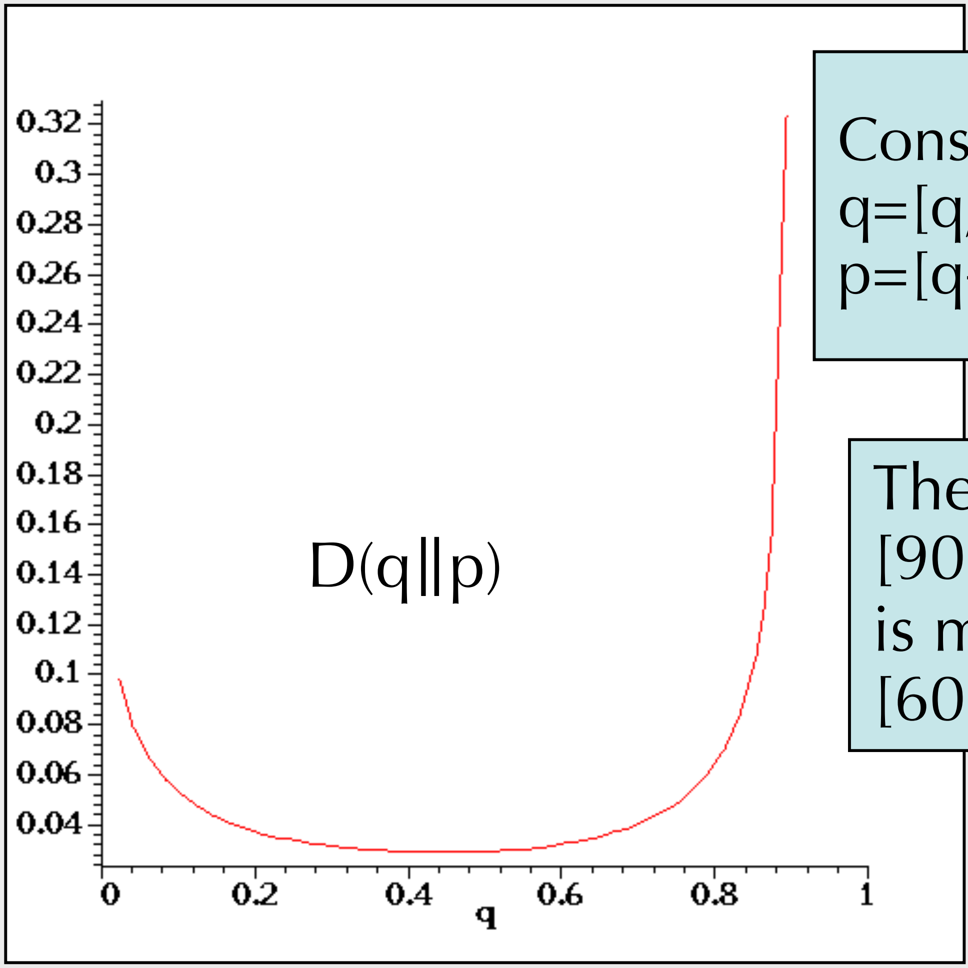
Compare a fair coin and biased coin.
Probabilities: $p = (\frac{1}{2},\frac{1}{2})$ versus $q = (q,1-q)$



D(p||q) expresses how much disbelief we have in the q-model, when we observe a fair coin (per coin flip).

D(q||p) expresses the disbelief in the fair-coin theory p when observing q.

Another D(q||p) Example



Consider a 10% difference:
q=[q,1−q] while
p=[q+1/10,1−q−1/10]

The difference between [90%,10%] and [80%,20%] is much bigger than between [60%,40%] and [50%,50%]

Try this…

Given a probability distribution p over $\mathcal{X}=\{1,\ldots,D\}$, prove that its entropy is upper bounded by $H(X) \leq \log D$.

.

.

.

Although 'obvious', proving this fact is not so easy.

Entropic Inequalities

Entropic definitions and equalities:

$H(X) = -\sum_x p(x) \log p(x)$, $H(X|Y) = -\sum_y p(y)\, H(X|Y{=}y)$,
$I(X;Y) = H(X) - H(X|Y)$, $D(p\|q) = \Sigma_x\, p(x) \log p(x)/q(x)$,…

Entropic inequalities that follow from $0 \leq p(x) \leq 1$:

$H(X) \geq 0$, $H(X|Y) \geq 0$, $H(X,Y) \geq H(X)$,…

Less obvious inequalities:

$I(X;Y) \geq 0$, $H(X) \leq \log|\mathcal{X}|$, $D(p\|q) \geq 0$,…

How do those inequalities relate?

How to prove the not-so-obvious ones?

Information Inequality

Theorem 2.6.3: For two probabilities distribution p and q we have $D(p\|q) \geq 0$ and $D(p\|q)=0$ if and only if $p=q$.

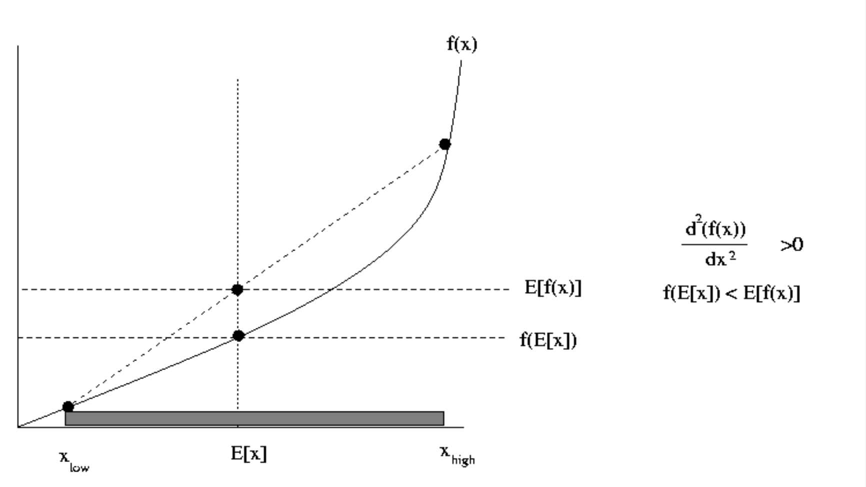
$$
\begin{aligned}
-D(p\|q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_x p(x) \log \frac{q(x)}{p(x)} \\
&\leqslant \log \sum_x p(x) \frac{q(x)}{p(x)} \\
&= \log 1 = 0
\end{aligned}
$$

Q: What happened here?
A: Application of "Jensen's inequality" to the concave function log:$\mathbb{R}^+ \to \mathbb{R}$.

Convex and Concave

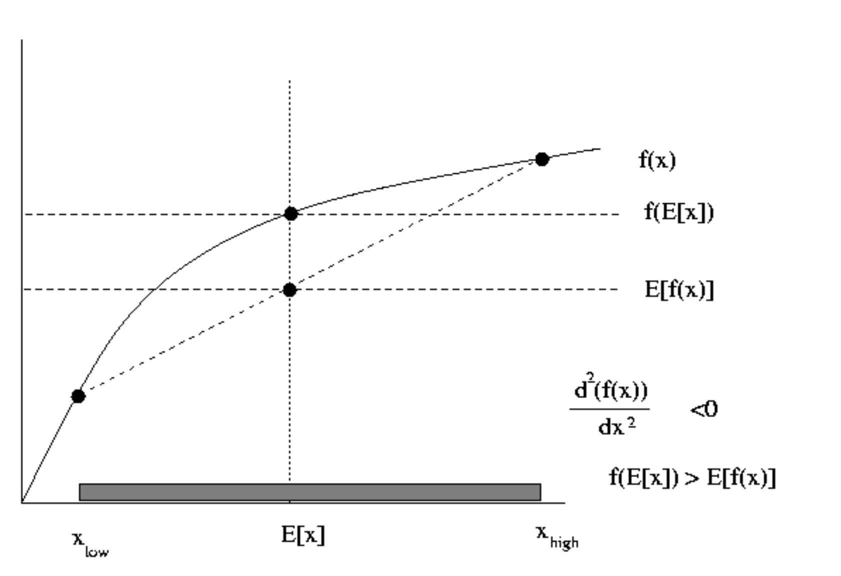
A function f is convex if for all points x_1 and x_2 it holds that $f(\lambda x_1+(1-\lambda)x_2) \leq \lambda f(x_1)+(1-\lambda)f(x_2)$.

Equivalently: $f''\geq 0$
Strictly convex: $f''>0$

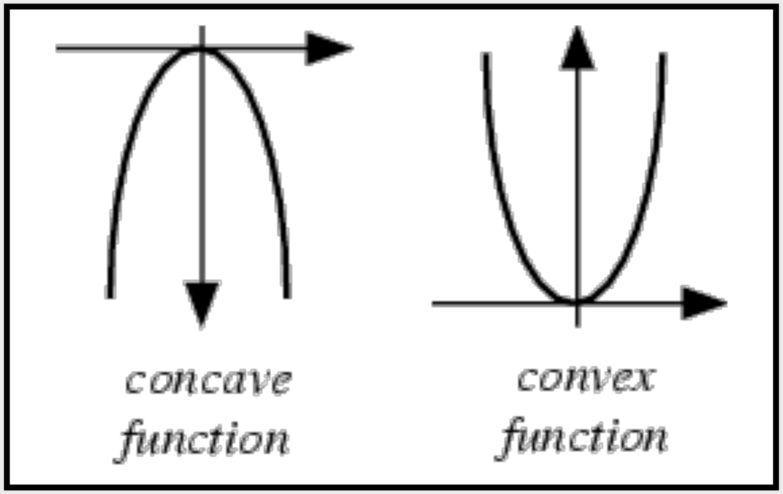


A function f is concave if for all points x_1 and x_2 it holds that $f(\lambda x_1+(1-\lambda)x_2) \geq \lambda f(x_1)+(1-\lambda)f(x_2)$.

Equivalently: $f''\leq 0$
Strictly concave: $f''<0$



Jensen's Inequality



Theorem 2.6.2: If f is a convex function on a random variable Z, then $E[f(Z)] \geq f(E[Z])$.

If f is a concave function on a random variable Z, then $E[f(Z)] \leq f(E[Z])$.

If f is strictly convex or strictly concave, then $E[f(Z)] = f(E[Z])$ implies that Z is a deterministic variable.

Proof: See Cover and Thomas, Theorem 2.6.2.

Note: f is convex if and only if $-f$ is concave.

Jensen for the Log Function

The function log(z) is concave for $z \in \mathbb{R}^+$, hence
$E[\log Z] \leq \log(E[Z])$ and also $E[\log 1/Z] \geq \log(1/E[Z])$.

For our purposes we typically have variables X,Y and some additional function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ such that (with $Z = g(X,Y)$) Jensen's inequality tells us:

$E[\log g(X,Y)] \leq \log(E[g(X,Y)])$.

Important example: $D(p\|q) = \Sigma_x\, p(x) \log q(x)/p(x) =$
$E_p[\log q(x)/p(x)] \leq \log(E_p[q(x)/p(x)]) =$
$\log(\Sigma_x\, p(x) \cdot q(x)/p(x)) = \log(\Sigma_x q(x)) = 0$.

Using Jensen's Inequality

Using Jensen's inequality on the log function, we get

$-\log(E(1/g(X))) \leq E(\log g(X)) \leq \log(E(g(X)))$

for all functions $g:\mathcal{X}\rightarrow\mathbb{R}^+$ and distributions $p(X)$.

Using this we can prove for all random variables X, Y and all distributions p,q:

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

$I(X;Y) \geq 0$

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

Proving $H(X) \leqslant \log|\mathcal{X}|$, Twice

Directly using Jensen's Inequality on $E[\log(1/p(X))]$:

$$
\begin{aligned}
H(X) &= E\,[-\log p(X)] \\
&= E\,[\log 1/p(X)] \\
&\leqslant \log(E\,[1/p(X)]) \\
&= \log \sum p(x)/p(x) \\
&= \log|\mathcal{X}|
\end{aligned}
$$

Using nonnegativity of Relative entropy $D(p\|q)$ between $p(X)$ and $q=1/|\mathcal{X}|$:

$$
\begin{aligned}
0 &\leqslant D(p\|q) \\
&= E\,[\log p(X)/q(X)] \\
&= E\,[\log p(X)|\mathcal{X}|] \\
&= E\,[\log p(X)] + \log|\mathcal{X}| \\
&= -\,H(X) + \log|\mathcal{X}|
\end{aligned}
$$

The upper bound on the entropy $H(X)=\log|\mathcal{X}|$ is achieved with $p(x) = 1/|\mathcal{X}|$.