Lecture 6 - Information theory

Jan Bouda

FI MU

April 19, 2009

Part I

Uncertainty and entropy

Uncertainty

- Given a random experiment it is natural to ask how uncertain we are about an outcome of the experiment.
- Compare two experiments - tossing an unbiased coin and throwing a fair six-sided dice. First experiment attains two outcomes and the second experiment has six possible outcomes. Both experiments have the uniform probability distribution. Our intuition says that we are more uncertain about an outcome of the second experiment.
- Let us compare tossing of an ideal coin and a binary message source emitting 0 and 1 both with probability $1/2$. Intuitively we should expect that the uncertainty about an outcome of each of these experiments is the same. Therefore the uncertainty should be based only on the probability distribution and not on the concrete sample space.
- Therefore, the uncertainty about a particular random experiment can be specified as a function of the probability distribution $\{p_1, p_2, \ldots, p_n\}$ and we will denote it as $H(p_1, p_2, \ldots, p_n)$.

Uncertainty - requirements

1. Let us fix the number of outcomes of an experiment and compare the uncertainty of different probability distributions. Natural requirement is that the most uncertain is the experiment with the uniform probability distribution, i.e. $H(p_1, \ldots p_n)$ is maximal for $p_1 = \cdots = p_n = 1/n$.
2. Permutation of probability distribution does not change the uncertainty, i.e. for any permutation $\pi : \{1 \ldots n\} \to \{1 \ldots n\}$ it holds that $H(p_1, p_2, \ldots, p_n) = H(p_{\pi(1)}, p_{\pi(2)} \ldots, p_{\pi(n)})$.
3. Uncertainty should be nonnegative and equals to zero if and only if we are sure about the outcome of the experiment. $H(p_1, p_2, \ldots, p_n) \geq 0$ and it is equal if and only of $p_i = 1$ for some i.
4. If we include into an experiment an outcome with zero probability, this does not change our uncertainty, i.e. $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$

Uncertainty - requirements

5. As justified before, having the uniform probability distribution on n outcomes cannot be more uncertain than having the uniform probability distribution on $n+1$ outcomes, i.e.

$$H(\overbrace{1/n,\ldots,1/n}^{n\times}) \leq H(\overbrace{1/(n+1),\ldots,1/(n+1)}^{(n+1)\times}).$$

6. $H(p_1,\ldots,p_n)$ is a continuous function of its parameters.
7. Uncertainty of an experiment consisting of a simultaneous throw of m and n sided die is as uncertain as an independent throw of m and n sided die implying

$$H(\overbrace{1/(mn),\ldots,1/(mn)}^{mn\times}) = H(\overbrace{1/m,\ldots,1/m}^{m\times}) + H(\overbrace{1/n,\ldots,1/n}^{n\times}).$$

Entropy and uncertainty

8. Let us consider a random choice of one of $n+m$ balls, m being red and n being blue. Let $p=\sum_{i=1}^{m} p_i$ be the probability that a red ball is chosen and $q=\sum_{i=m+1}^{m+n} p_i$ be the probability that a blue one is chosen. Then the uncertainty which ball is chosen is the uncertainty whether red of blue ball is chosen plus weighted uncertainty that a particular ball is chosen provided blue/red ball was chosen. Formally,

$$\begin{aligned} &H(p_1,\ldots,p_m,p_{m+1},\ldots,p_{m+n}) = \\ =&H(p,q)+pH\left(\frac{p_1}{p},\ldots,\frac{p_m}{p}\right)+qH\left(\frac{p_{m+1}}{q},\ldots,\frac{p_{m+n}}{q}\right). \end{aligned} \tag{1}$$

It can be shown that any function satisfying Axioms $1-8$ is of the form

$$H(p_1,\ldots,p_m) = -(\log_a 2)\sum_{i=1}^{m} p_i \log_2 p_i \tag{2}$$

showing that the function is defined uniquely up to multiplication by a constant, which effectively changes only the base of the logarithm.

Entropy and uncertainty

Alternatively, we may show that the function $H(p_1, \ldots, p_m)$ is uniquely specified through axioms

1. $H(1/2, 1/2) = 1$.
2. $H(p, 1-p)$ is a continuous function of p.
3. $H(p_1, \ldots, p_m) = H(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$

as in Eq. (2).

Entropy

The function $H(p_1, \ldots, p_n)$ we informally introduced is called the (Shannon) entropy and, as justified above, it measures our uncertainty about an outcome of an experiment.

Definition

Let X be a random variable with probability distribution $p(x)$. Then the **(Shannon) entropy** of the random variable X is defined as

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

In the definition we use the convention that $0 \log 0 = 0$, what is justified by $\lim_{x \to 0} x \log x = 0$. Alternatively, we may sum only over nonzero probabilities.

As explained above, all required properties are independent of multiplication by a constant what changes the base of the logarithm in the definition of the entropy. Therefore, in the rest of this part we will use logarithm without explicit base. In case we want to measure information in bits, we should use logarithm base 2.

Entropy

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a function. Let us recall that the expectation of the transformed random variable is $E[\phi(X)] = \sum_{x \in \mathbf{Im}(X)} \phi(x) P(X = x)$. Using this formalism we may write most of the information-theoretic quantities. In particular, the entropy can be expressed as

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

where $p(x) = P(X = x)$.

Lemma

$H(X) \geq 0$.

Proof.

$0 < p(x) \leq 1$ implies $\log(1/p(x)) \geq 0$. □

Part II

Joint and Conditional entropy

Joint entropy

In order to examine an entropy of more complex random experiments described by correlated random variables we have to introduce the entropy of a pair (or n–tuple) of random variables.

Definition

Let X and Y be random variables distributed according to the probability distribution $p(x, y) = P(X = x, Y = y)$. We define the **joint (Shannon) entropy** of random variables X and Y as

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

or, alternatively,

$$H(X, Y) = -E[\log p(X, Y)] = E\left[\frac{1}{\log p(X, Y)}\right].$$

Conditional entropy

Important question is how uncertain we are about an outcome of a random variable X given an outcome of a random variable Y. Naturally, our uncertainty about an outcome of X given $Y = y$ is

$$H(X|Y = y) = -\sum_{x \in \mathbf{Im}(X)} P(X = x|Y = y) \log P(X = x|Y = y). \qquad (3)$$

The uncertainty about an outcome of X given an (unspecified) outcome of Y is naturally defined as a sum of equations (3) weighted according to $P(Y = y)$, i.e.

Conditional Entropy

Definition

Let X and Y be random variables distributed according to the probability distribution $p(x,y) = P(X = x, Y = y)$. Let us denote $p(x|y) = P(X = x|Y = y)$. The **conditional entropy** of X given Y is

$$\begin{aligned} H(X|Y) = & \sum_{y \in \mathbf{Im}(Y)} p(y) H(X|Y = y) = \\ & = - \sum_{y \in \mathbf{Im}(Y)} p(y) \sum_{x \in \mathbf{Im}(X)} p(x|y) \log p(x|y) = \\ & = - \sum_{x \in \mathbf{Im}(X)} \sum_{y \in \mathbf{Im}(Y)} p(x,y) \log p(x|y) \\ & = - E[\log p(X|Y)]. \end{aligned} \tag{4}$$

Conditional Entropy

Using the previous definition we may raise the question how much information we learn on average about X given an outcome of Y. Naturally, we may interpret it as the decrease of our uncertainty about X when we learn outcome of Y, i.e. $H(X) - H(X|Y)$. Analogously, the amount of information we obtain when we learn the outcome of X is $H(X)$.

Theorem (Chain rule of conditional entropy)

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

Chain rule of conditional entropy

Proof.

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathbf{Im}(X)}\sum_{y\in\mathbf{Im}(Y)} p(x,y)\log p(x,y) = \\
&= -\sum_{x\in\mathbf{Im}(X)}\sum_{y\in\mathbf{Im}(Y)} p(x,y)\log[p(y)p(x|y)] = \\
&= -\sum_{\substack{x\in\mathbf{Im}(X)\\ y\in\mathbf{Im}(Y)}} p(x,y)\log p(y) - \sum_{\substack{x\in\mathbf{Im}(X)\\ y\in\mathbf{Im}(Y)}} p(x,y)\log p(x|y) = \\
&= -\sum_{y\in\mathbf{Im}(Y)} p(y)\log p(y) - \sum_{\substack{x\in\mathbf{Im}(X)\\ y\in\mathbf{Im}(Y)}} p(x,y)\log p(x|y) = \\
&= H(Y) + H(X|Y).
\end{aligned}
\tag{5}
$$

□

Chain rule of conditional entropy

Proof.

Alternatively we may use $\log p(X,Y) = \log p(Y) + \log p(X|Y)$ and take the expectation on both sides to get the desired result. □

Corollary (Conditioned chain rule)

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

Note that in general $H(Y|X) \neq H(X|Y)$. On the other hand, $H(X) - H(X|Y) = H(Y) - H(Y|X)$ showing that information is symmetric.

Part III

Relative Entropy and Mutual Information

Relative entropy

Let us start with the definition of the relative entropy, which measures inefficiency of assuming that a given distribution is $q(x)$ when the true distribution is $p(x)$.

Definition

The **relative entropy** or **Kullback-Leibler distance** between two probability distributions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_{x \in \mathbf{Im}(X)} p(x) \log \frac{p(x)}{q(x)} = E\left[\log \frac{p(X)}{q(X)}\right].$$

In the definition we use the convention that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Important is that the relative entropy is always nonnegative and it is zero if and only if $p(x) = q(x)$. It is not a distance in the mathematical sense since it is not symmetric in its parameters and it does not satisfy the triangle inequality.

Mutual information

Mutual information measures information one random variable contains about another random variable. It is the decrease of the uncertainty about an outcome of a random variable given an outcome of another random variable, as already discussed above.

Definition

Let X and Y be random variables distributed according to the probability distribution $p(x,y)$. The **mutual information** $I(X;Y)$ is the relative entropy between the joint distribution and the product of marginal distributions

$$\begin{aligned} I(X;Y) =& \sum_{x \in \mathbf{Im}(X)} \sum_{y \in \mathbf{Im}(Y)} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ =& D(p(x,y) \| p(x)p(y)) = E\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]. \end{aligned} \tag{6}$$

Mutual Information and Entropy

Theorem

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

Proof.

$$\begin{aligned} I(X;Y) =& \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} == \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} = \\ =& -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) = \\ =& -\sum_{x,y} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right) = \\ =& H(X) - H(X|Y). \end{aligned} \tag{7}$$

□

Mutual information

From symmetry we get also $I(X;Y) = H(Y) - H(Y|X)$. X says about Y as much as Y says about X. Using $H(X,Y) = H(X) + H(Y|X)$ we get

Theorem

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

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

Part IV

Properties of Entropy and Mutual Information

General Chain Rule for Entropy

Theorem

Let $X_1, X_2, \ldots, X_n$ be random variables. Then

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

Proof.

We use repeated application of the chain rule for a pair of random variables

$$\begin{aligned} H(X_1, X_2) =& H(X_1) + H(X_2|X_1), \\ H(X_1, X_2, X_3) =& H(X_1) + H(X_2, X_3|X_1) = \\ =& H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1), \\ & \vdots \end{aligned} \tag{8}$$

□

General Chain Rule for Entropy

Proof.

$$\vdots$$

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

□

Conditional Mutual Information

Definition

The **conditional mutual information** between random variables X and Y given Z is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E\left[\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right],$$

where the expectation is taken over $p(x,y,z)$.

Theorem (Chain rule for mutual information)

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

Conditional Relative Entropy

Definition

The **conditional relative entropy** is the average of the relative entropies between the conditional probability distributions $p(y|x)$ and $q(y|x)$ averaged over the probability distribution $p(x)$. Formally,

$$D\left(p(y|x) \| q(y|x)\right) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} = E\left[\log \frac{p(Y|X)}{q(Y|X)}\right].$$

The relative entropy between two joint distributions can be expanded as the sum of a relative entropy and a conditional relative entropy.

Theorem (Chain rule for relative entropy)

$$D(p(x,y) \| q(x,y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)).$$

Chain Rule for Relative Entropy

Proof.

$$
\begin{aligned}
D(p(x,y)\|q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} = \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} = \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} = \\
&= D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)).
\end{aligned}
\tag{9}
$$

□

Part V

Jensen's inequality

Convex and concave functions

Before introducing Jensen's inequality let us briefly refresh definitions of convex and concave function, what is crucial in this part.

Definition

A function $f(x)$ is said to be **convex** on a set $\mathbf{S}$ if for every $x_1, x_2 \in \mathbf{S}$ and $0 \leq \lambda \leq 1$

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

A function is **strictly convex** if the equality holds only if $\lambda = 0$ or $\lambda = 1$. A function f is **concave** if $-f$ is convex. A function f is **strictly concave** if $-f$ is strictly convex.

Theorem

If the function has a second derivative which is nonnegative (positive) everywhere, then the function is convex (strictly convex).

Convex and Concave Functions

Proof.

We use the Taylor series expansion of the function around x_0

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2,$$

where x^* lies between x_0 and x. By our initial assumption the term $f''(x^*)$ is always nonnegative and the same holds for the last addend. Let $x_0 = \lambda x_1 + (1 - \lambda)x_2$, $\lambda \in [0, 1]$ and $x = x_1$ and we have

$$f(x_1) \geq f(x_0) + f'(x_0)[(1 - \lambda)(x_1 - x_2)]. \tag{10}$$

Similarly, taking $x = x_2$ we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]. \tag{11}$$

Multiplying (10) by λ and (11) by $(1 - \lambda)$ and adding we obtain the convexity. The proof for the strict convexity is analogous. □

Convex and Concave Functions

Proof.

Multiplying (10) by λ and (11) by $(1-\lambda)$ and adding we obtain the convexity

$$
\begin{aligned}
&\lambda f(x_1) + (1-\lambda) f(x_2) \geq \\
\geq &\lambda(f(x_0) + f'(x_0)[(1-\lambda)(x_1-x_2)]) + (1-\lambda)(f(x_0) + f'(x_0)[\lambda(x_2-x_1)]) = \\
= &\lambda f(x_0) + (1-\lambda) f(x_0) + \lambda f'(x_0)[(1-\lambda)(x_1-x_2)] - (1-\lambda) f'(x_0)[\lambda(x_1-x_2)] = \\
= &f(x_0) = f(\lambda x_1 + (1-\lambda) x_2).
\end{aligned}
$$

The proof for the strict convexity is analogous. □

Jensen's Inequality

Last theorem shows immediately the strict convexity for x^2, e^x and $x \log x$ for $x \geq 0$, and the strict concavity of $\log x$ and $\sqrt{x}$ for $x \geq 0$.
The following inequality is behind most of the fundamental theorems in information theory and in mathematics in general.

Theorem (Jensen's inequality)

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

$$E[f(X)] \geq f(E(X)). \tag{12}$$

Moreover, if f is strictly convex, the equality in (12) implies that $X = E(X)$ occurs with probability 1, i.e. X is a constant.

Jensen's Inequality

Proof.

We prove this inequality by induction on the number of elements in $\mathbf{Im}(X)$. For probability distribution on two points we have

$$E(f(X)) = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(E(X)) \tag{13}$$

what follows directly from convexity. Suppose the theorem holds for $k-1$ points. Then we put $p'_i = p_i/(1-p_k)$ for $i = 1, 2 \ldots, k-1$ and we have

$$\begin{aligned} E(f(X)) = \sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \geq \\ &\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) \geq \end{aligned} \tag{14}$$

□

Jensen's Inequality

Proof.

$$
\begin{aligned}
&\geq f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p_i' x_i\right) = \\
&= f\left(\sum_{i=1}^{k} p_i x_i\right) = f(E(X)),
\end{aligned}
$$

where the first inequality follows from the induction hypothesis and the second one from convexity of f. □

Information Inequality

Theorem (Information inequality)

Let $p(x)$ and $q(x)$, $x \in \mathbf{X}$, be two probability distributions. Then

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

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

Information Inequality

Proof.

Let $\mathbf{A} = \{x|p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p\|q) &= -\sum_{x\in\mathbf{A}} p(x)\log\frac{p(x)}{q(x)} = \\
&= \sum_{x\in\mathbf{A}} p(x)\log\frac{q(x)}{p(x)} \leq \\
&\overset{(*)}{\leq} \log\sum_{x\in\mathbf{A}} p(x)\frac{q(x)}{p(x)} = \\
&= \log\sum_{x\in\mathbf{A}} q(x) \leq \log\sum_{x\in\mathbf{X}} q(x) = \\
&= \log 1 = 0,
\end{aligned} \tag{15}$$

where $(*)$ follows from Jensen's inequality. □

Information Inequality

Proof.

Since $\log t$ is a strictly concave function (implying $-\log t$ is strictly convex) of t, we have equality in $(*)$ if and only if $q(x)/p(x) = 1$ everywhere, i.e. $p(x) = q(x)$. Also, if $p(x) = q(x)$ the second inequality also becomes equality. □

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.

Proof.

$I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0$ with equality if and only if $p(x,y) = p(x)p(y)$, i.e. X and Y are independent. □

Consequences of Information Inequality

Corollary

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

with equality if and only if $p(y|x) = q(y|x)$ *for all* y *and* x *with* $p(x) > 0$*.*

Corollary

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

with equality if and only if X *and* Y *are conditionally independent given* Z*.*

Theorem

$H(X) \leq \log|\mathbf{Im}(X)|$ *with equality if and only if* X *has a uniform distribution over* $\mathbf{Im}(X)$*.*

Consequences of Information Inequality

Proof.

Let $u(x) = 1/|\mathbf{Im}(X)|$ be a uniform probability distribution over $\mathbf{Im}(X)$ and let $p(x)$ be the probability distribution of X. Then

$$D(p\|u) = \sum p(x) \log \frac{p(x)}{u(x)} =$$

$$= -\sum p(x) \log u(x) - \left(-\sum p(x) \log p(x)\right) = \log |\mathbf{Im}(X)| - H(X).$$

□

Theorem (Conditioning reduces entropy)

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

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

Consequences of Information Inequality

Proof.

$0 \leq I(X;Y) = H(X) - H(X|Y)$. □

Previous theorem says that on average knowledge of a random variable Y reduces our uncertainty about other random variable X. However, there may exist y such that $H(X|Y=y) > H(X)$.

Theorem (Independence bound on entropy)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

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

with equality if and only if X_i's are mutually independent.

Consequences of Information Inequality

Proof.

We use the chain rule for entropy

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ &\leq \sum_{i=1}^{n} H(X_i), \end{aligned} \tag{16}$$

where the inequality follows directly from the previous theorem. We have equality if and only if X_i is independent of all $X_{i-1}, \ldots, X_1$. □

Part VI

Log Sum Inequality and Its Applications

Log Sum Inequality

Theorem (Log sum inequality)

For a nonnegative numbers $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$ *it holds that*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $a_i/b_i = const.$

In the theorem we used again the convention that $0 \log 0 = 0$, $a \log(a/0) = \infty$ if $a > 0$ and $0 \log(0/0) = 0$.

Log Sum Inequality

Proof.

Assume WLOG that $a_i > 0$ and $b_i > 0$. The function $f(t) = t \log t$ is strictly convex since $f''(t) = \frac{1}{t} \log e > 0$ for all positive t. We use the Jensen's inequality to get

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right)$$

for $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. Setting $\alpha_i = b_i / \sum_{j=1}^n b_j$ and $t_i = a_i/b_i$ we obtain

$$\sum_i \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} \geq \left(\sum_i \frac{a_i}{\sum_j b_j}\right) \log \sum_i \frac{a_i}{\sum_j b_j},$$

what is the desired result. □

Consequences of Log Sum Inequality

Theorem

$D(p\|q)$ is convex in the pair (p, q), i.e. if (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions, then

$$D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)$$

for all $0 \le \lambda \le 1$.

Theorem (Concavity of entropy)

$H(p)$ is a concave function of p

Theorem

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

Part VII

Data Processing inequality

Data Processing Inequality

Theorem

$X \to Y \to Z$ is a Markov chain if and only if X and Z are independent when conditioned by Y, i.e.

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

Note that $X \to Y \to Z$ implies $Z \to Y \to X$. Also, if $Z = f(Y)$, then $X \to Y \to Z$.

Theorem (Data processing inequality)

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

Data Processing Inequality

Proof.

We expand mutual information using the chain rule in two different ways as

$$\begin{aligned} I(X;Y,Z) =& I(X;Z) + I(X;Y|Z) \\ =& I(X;Y) + I(X;Z|Y). \end{aligned} \tag{17}$$

Since X and Z are conditionally independent given Y we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$ we have

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

□