Lecture 3 More properties of entropy and mutual information

October 10, 2024

Outline

The following parametric family of entropies was introduce by Alfred Renyi in the mid 1950s as a generalization of Shannon entropy. Renyi wanted to find the most general class of information measure that preserved by additivity of statistically independent systems and were compatible with Kolmogrov's probability axioms.

Definition (Rényi entropy)

Given the parameter $\alpha > 0$ with $\alpha \neq 1$, and given a discrete random variable X with alphabet $\mathcal{X}$ and distribution P_X, its Rényi entropy of order α is given by

$$H_\alpha = \frac{1}{1-\alpha} \log\Big(\sum_{x \in \mathcal{X}} P_X(x)^\alpha\Big).$$

Let us assume that the outcomes of some experimental discrete random variable occur with probabilities $p_1, \cdots, p_N$, and if the kth outcome delivers I_k bits of information then the total amount of information for the set $\Gamma = \{I_1, \cdots, I_N\}$ is

$$I(P) = \sum_{k=1}^{N} p_k I_k \tag{1.1}$$

which can be recognized as Shannon's entropy $H(X)$. Here we have assumed the linear average operator in this formulation.

In the general theory of means for any monotonic function $g(x)$ with an inverse $g^{-1}(x)$ one can define the general associated with $g(x)$ for a set of real values $\{x_k, k = 1, \cdots, N\}$ with probabilities of $\{p_k\}$ as

$$g^{-1}\Big(\sum_{k=1}^{N} p_k g(x_k)\Big).$$

Applying this definition to the information $I(P)$, we obtain

$$I(P) = g^{-1}\Big(\sum_{k=1}^{N} p_k g(I_k)\Big)$$

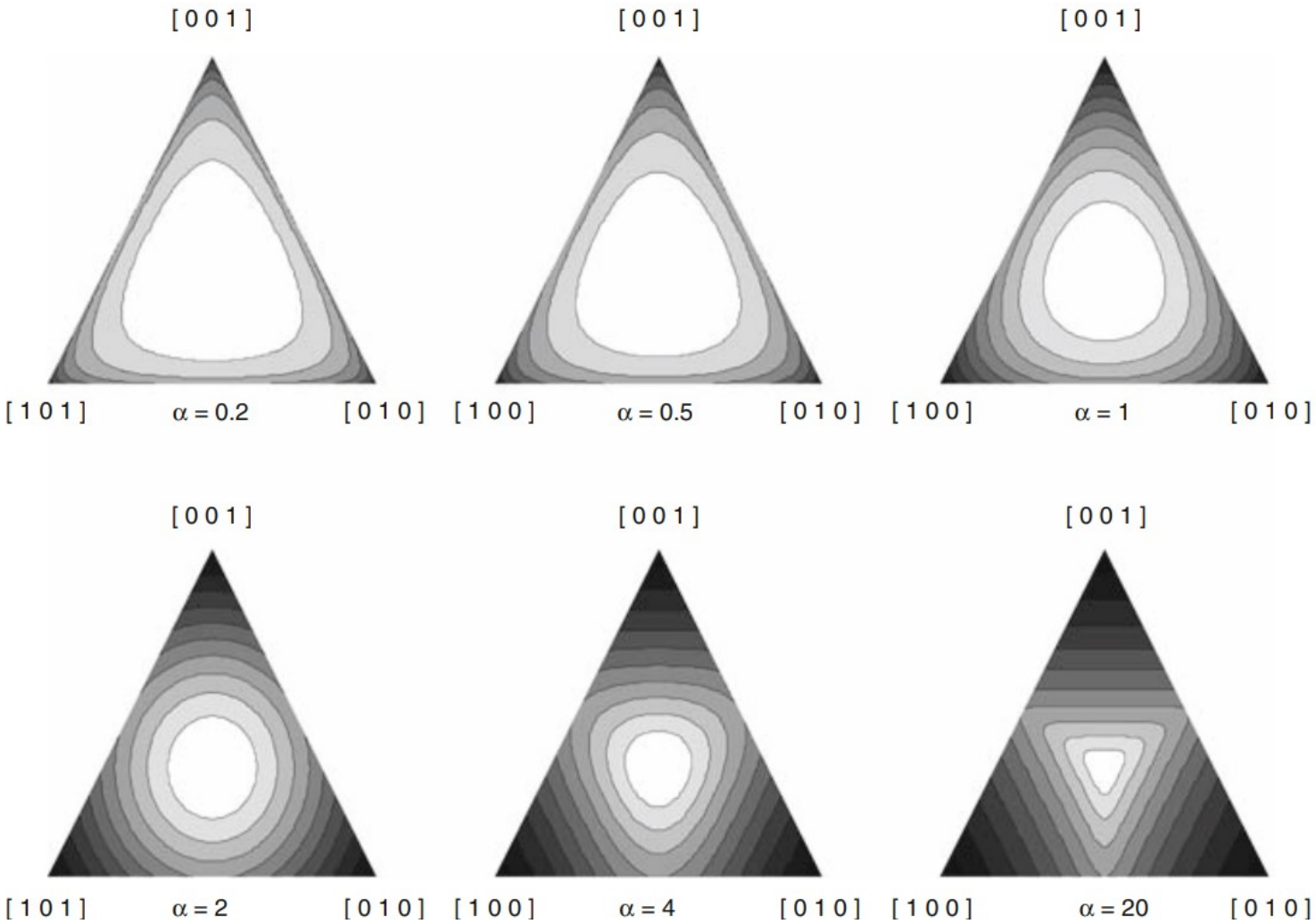
where $g(x)$ is a Kolmogorov-Nagumo invertible function. This $g(x)$ is the so called quasi-linear mean and it constitutes the most general mean compatible with Kolmogorov's axiomatics.

- Rényi then proved that when the postulate of additivity for independent events is applied to the above equation it dramatically restricts the class of possible $g(x)$. In fact, only two classes are possible.
- One is $g(x) = cx$ with c constant, which states that for linear $g(x)$ the quasi-linear mean reduces to the ordinary mean and yields the shannon information measure (1.1).
- The other functional class is $g(x) = c \cdot 2^{(1-\alpha)x}$ which implies

$$I_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{k=1}^{N} p_k^\alpha \right)$$

with $\alpha \neq 1$ and $\alpha \geq 0$, and it is called Rényi's information measure of order α.

Alfréd Rényi (20 March 1921 – 1 February 1970) was a Hungarian mathematician known for his work in probability theory, though he also made contributions in combinatorics, graph theory, and number theory.


[0 0 1]
[1 0 1]
α = 0.2
[0 1 0]
[0 0 1]
[1 0 0]
α = 0.5
[0 1 0]
[0 0 1]
[1 0 0]
α = 1
[0 1 0]
[0 0 1]
[1 0 1]
α = 2
[0 1 0]
[0 0 1]
[1 0 0]
α = 4
[0 1 0]
[0 0 1]
[1 0 0]
α = 20
[0 1 0]

Definition (Rényi divergence)

Given a parameter $0 < \alpha < 1$, and two discrete random variables X and $\hat{X}$ with common alphabet $\mathcal{X}$ and distribution P_X and $P_{\hat{X}}$, respectively, then the Rényi divergence of order α between X and $\hat{X}$ is given by

$$D_\alpha(X\|\hat{X}) = \frac{1}{\alpha - 1} \log\Big(\sum_{x\in\mathcal{X}} [P_X^\alpha(x) P_{\hat{X}}^{1-\alpha}(x)]\Big).$$

This definition can be extended to $\alpha > 1$ if $P_{\hat{X}}(x) > 0$ for all $x \in \mathcal{X}$.

Lemma

When $\alpha \to 1$, we have the following:

$$\lim_{\alpha \to 1} H_\alpha(X) = H(X)$$

and

$$\lim_{\alpha \to 1} D_\alpha(X \| \hat{X}) = D(X \| \hat{X}).$$

Fundamental inequality

Lemma (Fundamental inequality (FI))

For any $x > 0$ and $D > 1$, we have that

$$\log_D(x) \leq \log_D(e) \cdot (x - 1),$$

with equality if and only if $x = 1$.

Proof.

We shall show that for any $a > 0$,

$$\ln a \leq a - 1$$

with equality if and only if $a = 1$. Let $f(a) = \ln a - a + 1$. Then $f'(a) = 1/a - 1$ and $f''(a) = -1/a^2$. Since $f(1) = 0$, $f'(1) = 0$, and $f'(1) = -1 < 0$, we see that $f(a)$ attains its maximum value 0 when $a = 1$. □

Setting $y = 1/x$ and using FI above directly that for any $y > 0$, we also have that

$$\log_D(y) \geq \log_D(e)(1 - \frac{1}{y}),$$

also with equality iff $y = 1$. In the above the base-D logarithm was used. Specifically, for a logarithm with base-2, the above inequalities become

$$\log_2(e)(1 - \frac{1}{x}) \leq \log_2(x) \leq \log_2(e) \cdot (x - 1),$$

with equality iff $x = 1$.

Information inequality

Let X be a random variable taking values in an alphabet $\mathcal{X}$. The **support** of X, denoted by $\mathcal{S}_X$, is the set of all $x \in \mathcal{X}$ such that $p(x) > 0$. If $\mathcal{S}_X = \mathcal{X}$, we say that p is **strictly positive**.

Theorem

For any two probability distributions p and q on a common alphabet $\mathcal{X}$,

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

with equality if and only if $p = q$.

Proof.

If $q(x) = 0$ for some $x \in \mathcal{S}_p$, then $D(p\|q) = \infty$ and the theorem is trivially true. Therefore, we assume that $q(x) > 0$ for all $x \in \mathcal{S}_p$. Consider

$$\begin{aligned} D(p\|q) &= (\log e) \sum_{x \in \mathcal{S}_p} p(x) \ln \frac{p(x)}{q(x)} \\ &\geq \sum_{x \in \mathcal{S}_p} p(x) \big(1 - \frac{q(x)}{p(x)}\big) \\ &= (\log e) \big[\sum_{x \in \mathcal{S}_p} p(x) - \sum_{x \in \mathcal{S}_p} q(x) \big] \\ &\geq 0, \end{aligned}$$

where the inequality in the second line results from an application of the equality $\ln a \geq 1 - \frac{1}{2}$ for all $a > 0$, and the inequality in the last line follows from

$$\sum_{x \in \mathcal{S}_p} q(x) \leq 1 = \sum_{x \in \mathcal{S}_p} p(x).$$

This proves the inequality.

Proof.

For equality of the conclusion to hold, equality must hold in the above inequalities for all $x \in \mathcal{S}_p$. For the first inequality, we see that the equality holds if and only if $p(x) = q(x)$ for all $x \in \mathcal{S}_p$, which implies

$$\sum_{x \in \mathcal{S}_p} q(x) = \sum_{x \in \mathcal{S}_p} p(x) = 1,$$

Thus the equality in the conclusion holds if and only if $p(x) = q(x)$ for all $x \in \mathcal{S}_p$.

It is immediate that $p = q$ implies that $p(x) = q(x)$ for all $x \in \mathcal{S}_p$, so it remains to prove the converse. Since $\sum_x q(x) = 1$ and $q(x) \geq 0$ for all x, $p(x) = q(x)$ for all $x \in \mathcal{S}_p$ implies $q(x) = 0$ for all $x \notin \mathcal{S}_p$, and therefore $p = q$. The theorem is proved. □

Corollary

For any two random variables X, Y,

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

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

Proof.

We have that

$$I(X;Y) = D(p(x,y)\|p(x)p(y)) \geq 0,$$

with equality iff $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent). □

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 *such that* $p(x) > 0$.

Corollary

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

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

Theorem

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

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

Theorem

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

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

Proof.

$0 \leq I(X;Y) = H(X) - H(X|Y).$ □

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 average $H(X|Y) = \sum_y p(y)H(X|Y = y) \leq H(X)$.

Example

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

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

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

Theorem

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

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

with equality if and only if the X_i are independent.

Proof.

By applying the chain rule for entropy,

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

Equality holds iff each conditional entropy is equal to its associated entropy, that iff X_i is independent of$(X_{i-1}, \cdots, X_1)$ for all i. □

Upper bound on entropy

Theorem

If a random variable X takes values from a finite set $\mathcal{X}$, then

$$H(X) \leq \log_2 |\mathcal{X}|,$$

where $|\mathcal{X}|$ is the size of the set $\mathcal{X}$. Equality holds if and only if X is equiprobable or uniformly distributed over $\mathcal{X}$ (i.e. $P_X(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$).

Proof.

$$
\begin{aligned}
\log_2 |\mathcal{X}| - H(X) &= \sum_{x\in\mathcal{S}_X} P_X(x)\cdot \log_2 |\mathcal{X}| + \sum_{x\in\mathcal{S}_X} P_X(x)\log_2 P_X(x) \\
&= \sum_{x\in\mathcal{S}_X} P_X(x)\cdot \log_2[|\mathcal{X}|\cdot P_X(x)] \\
&\geq \sum_{x\in\mathcal{S}_X} P_X(x)\cdot \log_2(e)(1-\frac{1}{|\mathcal{X}|\cdot P_X(x)}) \\
&= \log_2(e)\sum_{x\in\mathcal{S}_X}(P_X(x)-\frac{1}{|\mathcal{X}|}) \\
&= \log_2(e)(1-\frac{|\mathcal{S}_X|}{|\mathcal{X}|}) \geq 0.
\end{aligned}
$$

with equality if and only if $\mathcal{S}_X = X$ and $|\mathcal{X}|\cdot P_X(x) = 1$, which means $P_X(\cdot)$ is a uniform distribution on $\mathcal{X}$. □

- Intuitively, entropy tells us how random X is.

- Intuitively, entropy tells us how random X is.
- X is deterministic if and only if $H(X) = 0$.

- Intuitively, entropy tells us how random X is.
- X is deterministic if and only if $H(X) = 0$.
- If X is uniform (equiprobable), $H(X)$ is maximized and equal to $\log_2 |\mathcal{X}|$.

Bound on mutual information

Theorem

If $\{(X_i, Y_i)\}_{i=1}^n\}_{i=1}^n$ *is a process satisfying the conditional independence assumption* $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$, *then*

$$I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \leq \sum_{i=1}^n I(X_i; Y_i),$$

with equality holding iff $\{X_i\}_{i=1}^n$ *are independent.*

Proof.

From the independence bound on entropy, we have the inequality $H(Y_1, \cdots, Y_n) \leq \sum_{i=1}^{n} H(Y_i)$. By the conditional independence assumption, we have

$$\begin{aligned} H(Y_1, \cdots, Y_n | X_1, \cdots, X_n) &= E[-\log_2 P_{Y^n|X^n}(Y^n|X_n)] \\ &= E[-\sum_{i=1}^{n} \log_2 P_{Y_i|X_i}(Y_i|X_i)] \\ &= \sum_{i=1}^{n} H(Y_i|X_i). \end{aligned}$$

Hence,

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i).$$

with equality holding iff $\{Y_i\}_{i=1}^{n}$ are independent, which holds iff $\{X_i\}_{i=1}^{n}$ are independent. □

Theorem (Log-sum inequality)

For non-negative numbers $a_1, a_2, \cdots, a_n$ *and* $b_1, b_2, \cdots, b_n$,

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

with equality if and only if $\frac{a_i}{b_i} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$, *which is a constant that does not depend on* i.

Proof.

Assume without loss of generality that $a_i > 0$ and $b_i > 0$. Let $a := \sum_{i=1}^n a_i$ and $b := \sum_{i=1}^n b_i$. Then

$$\sum_{i=1}^n a_i \log_D \frac{a_i}{b_i} - a \log_D \frac{a}{b} = a\Big[\sum_{i=1}^n \frac{a_i}{a} \log_D \frac{a_i}{b_i} - \underbrace{(\sum_{i=1}^n \frac{a_i}{a})}_{=1} \log_D \frac{a}{b}\Big]$$

$$= a \sum_{i=1}^n \frac{a_i}{a} \log_D \big[\frac{a_i}{b_i} \frac{b}{a}\big]$$

$$\geq a \log_D(e) \sum_{i=1}^n \frac{a_i}{a} \big[1 - \frac{b_i}{a_i} \frac{a}{b}\big]$$

$$= a \log_D(e) \big(\sum_{i=1}^n \frac{a_i}{a} - \sum_{i=1}^n \frac{b_i}{b}\big)$$

$$= a \log_D(e)(1-1) = 0.$$

with equality holding iff $\frac{a_i}{b_i} \frac{b}{a} = 1$ for all i; i.e., $\frac{a_i}{b_i} = \frac{a}{b}$ $\forall i$. □

- We can use the log sum inequality to prove various convexity results.
- Here we prove the information inequality, which says that $D(p\|q) \geq 0$ with equality if and only if $p(x) = q(x)$.
- By the log sum inequality,

$$\begin{aligned} D(p\|q) &= \sum p(x) \log \frac{p(x)}{q(x)} \\ &\geq \left(\sum p(x)\right) \log \sum p(x) / \sum q(x) \\ &= 1 \log \frac{1}{1} = 0 \end{aligned}$$

with equality iff $\frac{p(x)}{q(x)} = c$. Since both p and q are probability mass functions, $c = 1$, and hence we have $D(p\|q) = 0$ iff $p(x) = q(x)$ for all x.

Outline

Convex and concave function

Definition

A function $f(x)$ is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ 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 f is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Convex and concave function

Definition

A function $f(x)$ is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

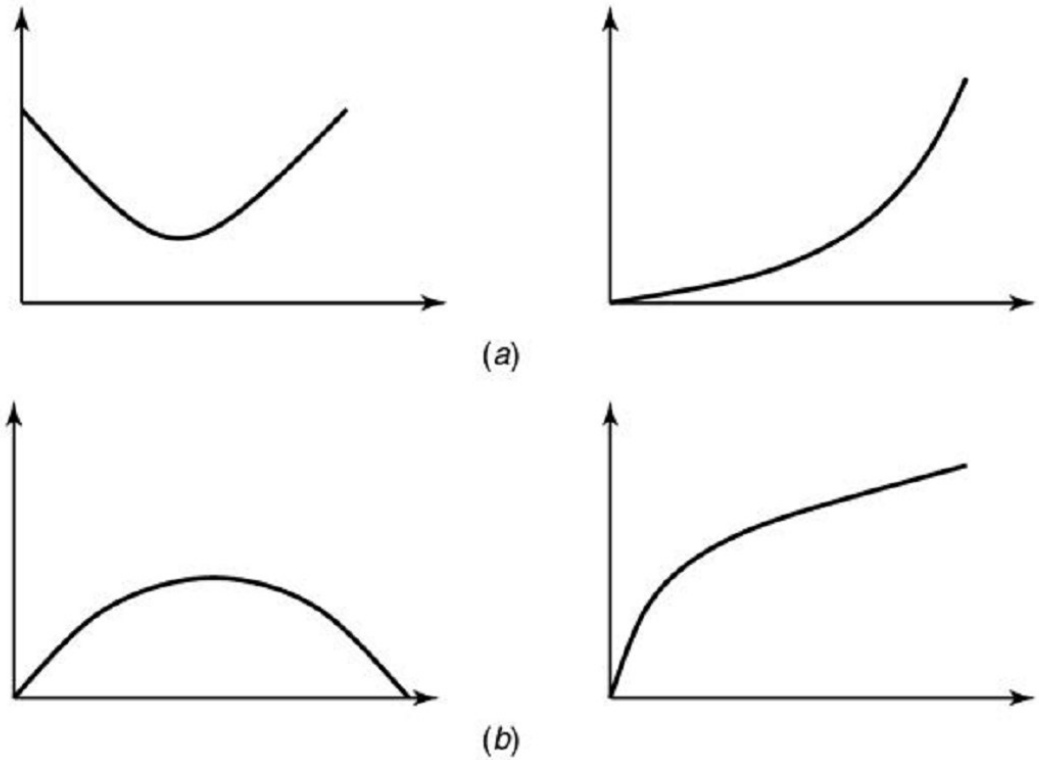
$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

A function f is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition

A function f is concave if $-f$ is convex.

A function is convex if it always lies below any chord. A function is concave if it always lies above chord.



Theorem

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

$$Ef(X) \geq f(EX).$$

Moreover, if f is strictly convex, the above inequality implies that $X = EX$ with probability 1.

- All the inequalities in last section can be also proved using Jensen's inequality. We shall prove the log sum inequality.
- Assume without loss of generality that $a_i > 0$ and $b_i > 0$.
- Let f be a strictly convex function, $\alpha_i \geq 0$, and $\sum_{i=1}^{n} \alpha_i = 1$. Jensen's inequality states that

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f(\sum_{i=1}^{n} \alpha_i t_i).$$

- Equality holds if and only if t_i is a constant for all i.
- To prove the log-sum inequality, set $\alpha_i = b_i / \sum_{j=1}^{n} b_j$, $t_i = a_i/b_i$, and $f(t) = t \cdot \log_D(t)$, we obtain the desired result.

Outline

Convexity of relative entropy

Theorem

$D(p\|q)$ is convex in the pair (p,q); that is, if (p_1,q_1) and (p_2,q_2) are two pairs of probability mass functions, then

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

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

Proof.

We apply the log sum inequality to a term on the left-hand side of the above equation:

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}$$
$$\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}.$$

Summing this over all x, we obtain the desired property. □

Theorem

$H(p)$ is a concave function of p. More precisely, $H(P_X)$ is a concave function of P_X, namely

$$H(\lambda P_X + (1-\lambda)P_{\tilde{X}}) \geq \lambda H(P_X) + (1-\lambda)H(P_{\tilde{X}})$$

for all $\lambda \in [0,1]$.

Proof.

We have that

$$H(p) = \log|\mathcal{X}| - D(p\|u),$$

where u is the uniform distribution on $|\mathcal{X}|$ outcomes. The concavity of H then follows directly from the convexity of D. □

Alternative proof.

Let X_1 be a random variable with distribution p_1, taking on values in a set A. Let X_2 be another random variable with distribution p_2 on the same set. Let

$$\theta = \begin{cases} 1 & \text{with probability } \lambda, \\ 2 & \text{with probability } 1-\lambda. \end{cases}$$

Let $Z = X_\theta$. Then the distribution of Z is $\lambda p_1 + (1-\lambda)p_2$. Now since conditioning reduces entropy, we have $H(Z) \geq H(Z|\theta)$, or equivalently,

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2),$$

which proves the concavity of the entropy as a function of the distribution. □

Theorem

Noting that $I(X;Y)$ *can be written as* $I(P_X, P_{Y|X})$*, where*

$$I(P_X, P_{Y|X}) := \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)P_X(x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{a\in\mathcal{X}} P_{Y|X}(y|a)P_X(a)},$$

then $I(X;Y)$ *is a concave function of* P_X *(for fixed* $P_{Y|X}$*, and a convex function of* $P_{Y|X}$ *(for fixed* P_X*).*

Next lecture

- Markov Chains (4.1)
- Entropy Rate (4.2-4.3)