“Information is the resolution of uncertainty.”

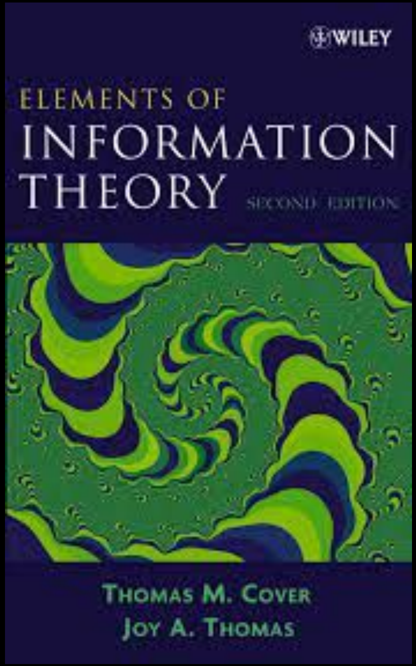
Shannon

Erfan Mirzaei

Information Theory Mini-Course

Nov 2022

References and Acknowledges



Session 3

Information Inequality, Max Entropy, Information never hurts, Data processing inequality, Sufficient statistics

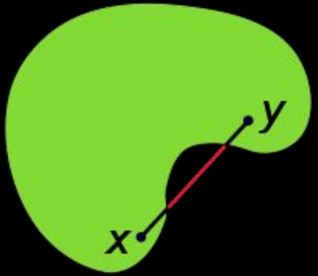
Convex combination/set

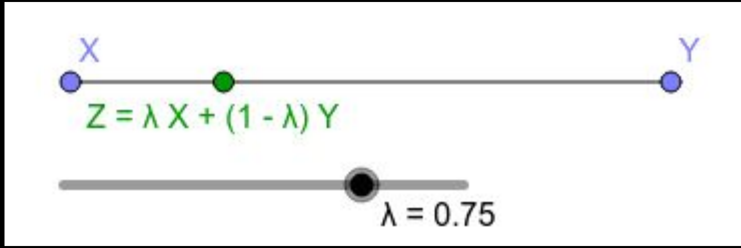
- Convex combination:

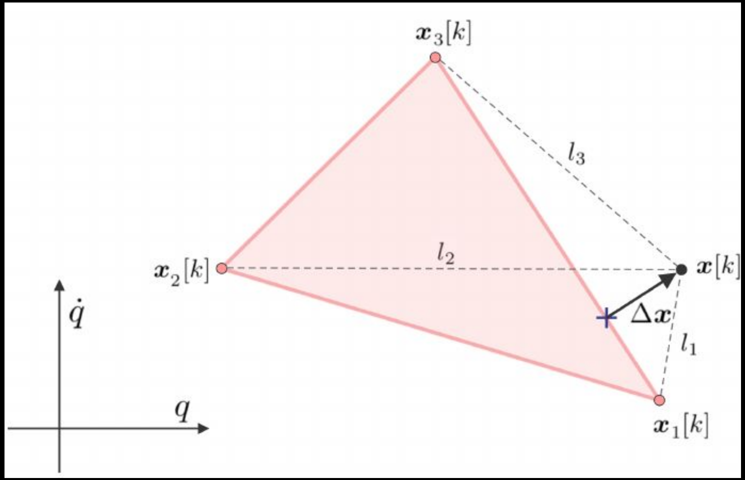
  ax + (1-a)y: 0 =< a =< 1

- Convex set:

  Convex combination of any arbitrary points are in the set.



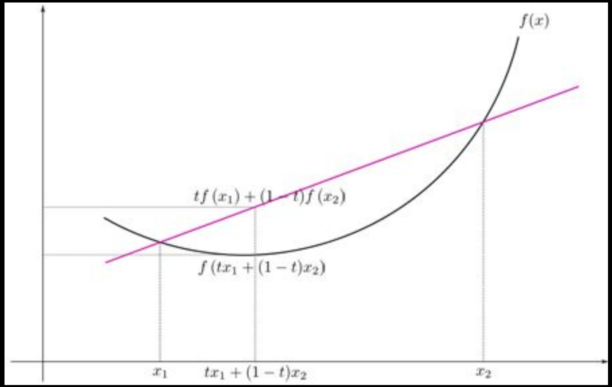


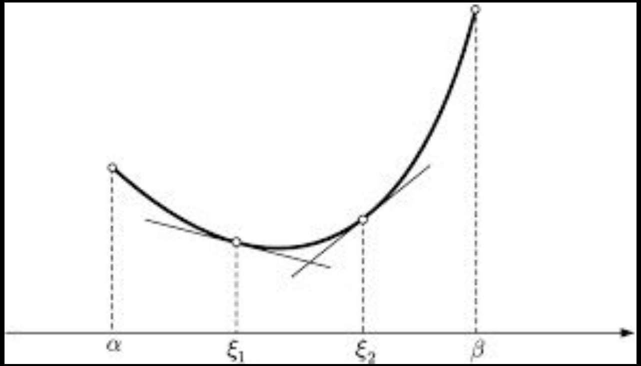


Convex function

- f'' >= 0, if it has second derivative everywhere on the domain.
- f(x) >= f(y) + f'(y)(x - y), if it has first derivative everywhere on the domain.
- For every 0 =< t =< 1, x1, x2 in the domain:

  f(tx1+(1-t)x2) =< tf(x1) + (1-t)f(x2)



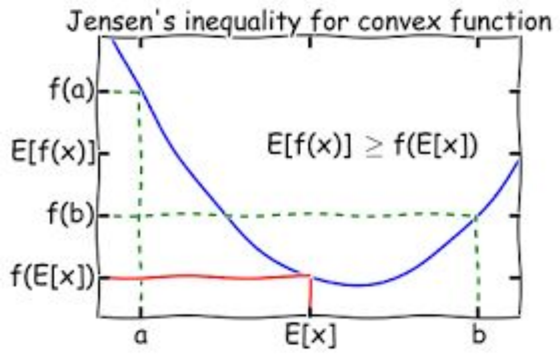


Jensen Inequality

Theorem 2.6.2 *(Jensen's inequality)* *If f is a convex function and X is a random variable,*

$$Ef(X) \geq f(EX). \tag{2.76}$$

Moreover, if f is strictly convex, the equality in (2.76) implies that $X = EX$ with probability 1 (i.e., X is a constant).



Jensen Inequality

Proof: We prove this for discrete distributions by induction on the number of mass points. The proof of conditions for equality when f is strictly convex is left to the reader.

For a two-mass-point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \tag{2.77}$$

which follows directly from the definition of convex functions. Suppose that the theorem is true for distributions with $k-1$ mass points. Then writing $p_i' = p_i/(1-p_k)$ for $i = 1, 2, \ldots, k-1$, we have

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) && (2.78)\\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) && (2.79)\\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) && (2.80)\\
&= f\left(\sum_{i=1}^{k} p_i x_i\right), && (2.81)
\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second follows from the definition of convexity.

The proof can be extended to continuous distributions by continuity arguments. □

Information Inequality

Theorem 2.6.3 *(Information inequality)* *Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. Then*

$$D(p||q) \geq 0 \tag{2.82}$$

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

Corollary *(Nonnegativity of mutual information)* *For any two random variables, X, Y,*

$$I(X;Y) \geq 0, \tag{2.90}$$

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

Information Inequality

$$
\begin{aligned}
D_{\mathrm{KL}}(p|q) &= \textstyle\sum_i p_i \log \frac{p_i}{q_i} \\
&= \textstyle\sum_i (-p_i \log q_i + p_i \log p_i) \\
&= -\textstyle\sum_i p_i \log q_i + \sum_i p_i \log p_i \\
&= -\textstyle\sum_i p_i \log q_i - \sum_i p_i \log \frac{1}{p_i} \\
&= -\textstyle\sum_i p_i \log q_i - H(p) \\
&= \textstyle\sum_i p_i \log \frac{1}{q_i} - H(p)
\end{aligned}
$$

Information Inequality

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

$$X = \frac{q(x)}{p(x)}$$

$$-D(p||q) = -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \quad (2.83)$$

$$= \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \longrightarrow f(X) \quad (2.84)$$

$$\le \log \sum_{x\in A} p(x) \frac{q(x)}{p(x)} \quad (2.85)$$

$$= \log \sum_{x\in A} q(x) \quad (2.86)$$

$$\le \log \sum_{x\in \mathcal{X}} q(x) \quad (2.87)$$

$$= \log 1 \quad (2.88)$$

$$= 0, \quad (2.89)$$

Maximum Entropy

Theorem 2.6.4 *$H(X) \le \log |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality if and only X has a uniform distribution over $\mathcal{X}$.*

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over $\mathcal{X}$, and let $p(x)$ be the probability mass function for X. Then

$$D(p \,\|\, u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X). \tag{2.93}$$

Hence by the nonnegativity of relative entropy,

$$0 \le D(p \,\|\, u) = \log |\mathcal{X}| - H(X). \quad \square \tag{2.94}$$

Information can't hurt

Theorem 2.6.5 *(Conditioning reduces entropy)(Information can't hurt)*

$$H(X|Y) \le H(X) \tag{2.95}$$

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

Proof: $0 \le I(X;Y) = H(X) - H(X|Y)$. □

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 the average $H(X|Y) = \sum_y p(y)H(X|Y = y) \le H(X)$. For example, in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.

Independence bound on entropy

Theorem 2.6.6 *(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*

$$H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i) \tag{2.96}$$

with equality if and only if the X_i are independent.

Proof: By the chain rule for entropies,

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) && (2.97) \\ &\le \sum_{i=1}^{n} H(X_i), && (2.98) \end{aligned}$$

Log sum inequality

Theorem 2.7.1 *(Log sum inequality)* *For nonnegative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.99}$$

with equality if and only if $\frac{a_i}{b_i} = const.$

We again use the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$. These follow easily from continuity.

Convexity of relative entropy

Theorem 2.7.2 *(Convexity of relative entropy)* *$D(p||q)$ 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*

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{2.105}$$

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

Proof: We apply the log sum inequality to a term on the left-hand side of (2.105):

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}$$
$$\le \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}. \tag{2.106}$$

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

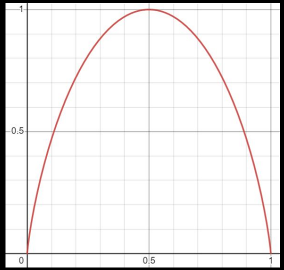
Concavity of entropy

Theorem 2.7.3 *(Concavity of entropy)* *$H(p)$ is a concave function of p.*

Proof

$$H(p) = \log |\mathcal{X}| - D(p||u), \tag{2.107}$$

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



Markov chain(process)

Definition Random variables X, Y, Z are said to *form a Markov chain in that order* (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y). \tag{2.117}$$

$X \to Y \to Z$ if and only if X and Z are conditionally independent given Y. Markovity implies conditional independence because

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y). \tag{2.118}$$

$X \to Y \to Z$ implies that $Z \to Y \to X$.

Data Processing inequality

Theorem 2.8.1 (*Data-processing inequality*) *If $X \to Y \to Z$, then* $I(X;Y) \geq I(X;Z)$.

Proof: By the chain rule, we can expand mutual information in two different ways:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \tag{2.119}$$

$$= I(X;Y) + I(X;Z|Y). \tag{2.120}$$

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

$$I(X;Y) \geq I(X;Z). \tag{2.121}$$

We have equality if and only if $I(X;Y|Z) = 0$ (i.e., $X \to Z \to Y$ forms a Markov chain). Similarly, one can prove that $I(Y;Z) \geq I(X;Z)$. □

Data Processing inequality

Corollary *In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$.*

Proof: $X \to Y \to g(Y)$ forms a Markov chain. □

Thus functions of the data Y cannot increase the information about X.

The data-processing inequality can be used to show that no clever manipulation of the data can improve the inferences that can be made from the data.

Sufficient statistics

Bernoulli distribution with theta parameter

Theta = ?

X = {0, 1, ...}

Theta_hat = sum(X)/len(X)

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{\sum_{i=1}^{n} 1-x_i}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \sum_{i=1}^{n} x_i \, \theta^{(\sum_{i=1}^{n} x_i)-1} (1-\theta)^{\sum_{i=1}^{n} 1-x_i} + (-1)\theta^{\sum_{i=1}^{n} x_i} \left(\sum_{i=1}^{n} 1-x_i\right)(1-\theta)^{(\sum_{i=1}^{n} 1-x_i)-1}$$

$$= \sum_{i=1}^{n} x_i (1-\theta) - \left(\sum_{i=1}^{n} 1-x_i\right)\theta$$

$$= \left(\sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} x_i\right) - \left(\theta n - \theta \sum_{i=1}^{n} x_i\right)$$

$$= \sum_{i=1}^{n} x_i - \theta n$$

$$\theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

X = sum(X)

What can we say about general case?

Sufficient statistics

- A family of probability mass functions $\{f\theta\ (x)\}$ indexed by θ
- X be a sample from a distribution in this family
- Let T (X) be any statistic (function of the sample) like the sample mean or sample variance. For any distribution on theta, we have:

$$\theta \to X \to T(X),$$

$$I(\theta; T(X)) \le I(\theta; X)$$

- However, if equality holds, no information is lost.
- A statistic T (X) is called sufficient for θ if it contains all the information in X about θ.

Sufficient statistics

Definition A function $T(X)$ is said to be a *sufficient statistic* relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ[i.e., $\theta \to T(X) \to X$ forms a Markov chain].

This is the same as the condition for equality in the data-processing inequality,

$$I(\theta; X) = I(\theta; T(X)) \tag{2.124}$$

for all distributions on θ. Hence sufficient statistics preserve mutual information and conversely.

Sufficient statistics

Let $X_1, X_2, \ldots, X_n$, $X_i \in \{0, 1\}$, be an independent and identically distributed (i.i.d.) sequence of coin tosses of a coin with unknown parameter $\theta = \Pr(X_i = 1)$. Given n, the number of 1's is a sufficient statistic for θ. Here $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$. In fact, we can show that given T, all sequences having that many 1's are equally likely and independent of the parameter θ. Specifically,

$$\Pr\left\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) \,\middle|\, \sum_{i=1}^{n} X_i = k\right\}$$

$$= \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } \sum x_i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.125}$$

Thus, $\theta \to \sum X_i \to (X_1, X_2, \ldots, X_n)$ forms a Markov chain, and T is a sufficient statistic for θ.

Sufficient statistics

Definition A statistic $T(X)$ is a *minimal sufficient statistic* relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistic U. Interpreting this in terms of the data-processing inequality, this implies that

$$\theta \to T(X) \to U(X) \to X. \tag{2.128}$$

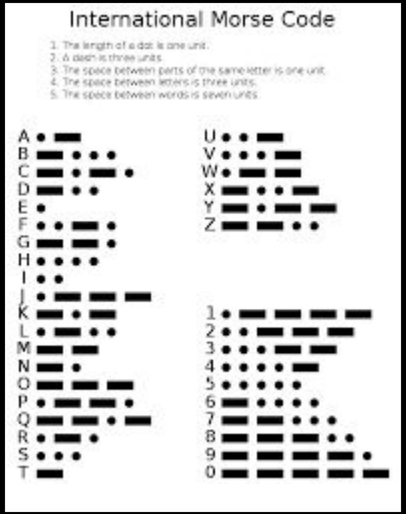
Hence, a minimal sufficient statistic maximally compresses the information about θ in the sample. Other sufficient statistics may contain additional irrelevant information. For example, for a normal distribution with mean θ, the pair of functions giving the mean of all odd samples and the mean of all even samples is a sufficient statistic, but not a minimal sufficient statistic. In the preceding examples, the sufficient statistics are also minimal.

Language

- You are lost in an island and can not hear the voice of each other:
-  OR  [You have enough of them.]
- Persian Language: 32 letters

  Sol: Send All the letters respectively, with how many stones?

- Can you do better?
  - Yes, you can :)
  - Log2 26 = 4.8 vs. 4.1257



Language

- You are lost in an island and can not hear the voice of each other:
-  OR  OR  OR  [You have enough of them.]
- Morse code for English = 4.1257 —> 2.0628
- Seems good, right? Why we don't use as many as stone patterns?
- To some extent, but is that simple?
  - Decoding error
- English: 26 characters, Persian: 32 characters
- Can we say Persian is more efficient for communication? No.
- Words length, their frequency, sentence size, ...

Thanks for your attention

Erfan Mirzaei
erfunmirzaei@gmail.com