# (Classical) Information Theory I

**Sibasish Ghosh**

*The Institute of Mathematical Sciences*
*CIT Campus, Taramani, Chennai 600 113, India.*

# Abstract

**Classical information theory provides useful methods to process and transmit informations using laws of classical physics. It was Shannon who had laid the foundational structure of classical information theory, back in 1948, by providing (i) an optimal data compression scheme for storing classical information through a process of removing all the redundancies in the data, and (ii) an optimal channel encoding-decoding scheme for transmitting classical information in an error-free manner through a process of inserting redundancies into the encoding. In this set of three lectures, I will discuss about these two schemes after a brief discussion about Shannon's entropy. If time permits, I will then briefly discuss about Kolmogorov's complexity and its relation with Shannon's entropy.**

- **First lecture: What is classical information; Shannon's entropy as information content; Coding schemes; Examples.**
- **Second lecture: Shannon's source coding theorem.**
- **Third lecture: Shannon's noisy channel coding theorem.**

# What is classical information?

**Statement 1:** **The sun will rise on the east today.**

**Statement 2:** **We may have rain fall tonight at Chennai.**

- **Statement 1 does not add any useful information to our knowledge, it is a certain event.**
- **Statement 2 does add some useful information to our knowledge, as rain fall at Chennai is not a certain event.**

## Information is all about probabilities

- **Statement 3: One-third of the world's population is India's present population.**
- **Statement 4: Two-third of the earth's surface is covered by water.**
- **Is there any difference between the information contents of these two statements?.**
- **No! Given any person, our ignorance about his/her nationality is *same* as our ignorance about the character (viz. water or solid) of any given surface area of the Earth.**
- **So information is all about the probabilities of occurrances of different events: probability that a person (from all over the world) is Indian is $1/3$, probability that a surface area of the Earth is solid is equal to $1/3$, etc.**

## Information is ignorance

- **Thus the amount of information about an event is the amount of ignorance (or, uncertainty) about that event.**
- **Ignorance increases with increase of the inverse of the** $probability\ p$ **of occurance of the event.**
- **Total amount of ignorance of two** $independent$ **events is sum of the ignorances.**
- **So the amount of ignorance** $I(p)$ **should be an additive function of** $p$**.**

# Shannon's entropy

- **Using continuity and additivity properties of $I(p)$, Shannon (1948) has shown that $I(p) = \log_2(1/p)$ upto some additive and/or multiplicative constant.**

- **So the average information content of a set $X$ of $n$ mutually exclusive but exaustive events $x_1$, $x_2$, , $x_n$ with respective probabilities $p_1$, $p_2$, . . ., $p_n$ is the given by the Shannon entropy $H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$.**

- **$H(X)$ depends only on $p_i$'s, not on event names $x_i$'s.**

- **Base 2 in the logarithm is used in order to express every event as a string of bit values 0 and 1, and thereby, $H(X)$ has its unit as no. of bits.**

## Information content of a random variable

- **As the names of the events (e.g., Sun will rise today on the East, we will not have rain fall tonight in Chennai, Barrack Obama is a non-Indian national, Mediterranian sea is not a portion of the land area of Earth's surface) are immaterial so far as there information contents are concerned, it is enough to consider an abstract variable $X$ having values ($x$, say) as the events (in a statement) with associated probabilities $\mathrm{Prob}(X = x) \equiv p_x$ being the corresponding probabilities of occurances of the events.**

- **Such an $X$ is called as a *random variable* in probability theory.**

## Information content of ....

- **Thus, according to Shannon, the average information content of a random variable $X$, having** *probability mass density* $p_x$**, is the entropy** $H(X) = -\sum_x p_x \log_2 p_x$**.**

- **For a joint probability mass density** $p_{x_1,x_2,\ldots,x_n} \equiv \text{Prob}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ **of an unordered sequence of $n$ random variables $X_1, X_2, \ldots, X_n$, the average information content is given by the joint Shannon entropy:** $H(X_1, X_2, \ldots, X_n) = -\sum_{x_1,x_2,\ldots,x_n} p_{x_1,x_2,\ldots,x_n} \log_2 p_{x_1,x_2,\ldots,x_n}$**.**

- **In case $X_1, X_2, \ldots, X_n$ are independent random variables (i.e.,** $p_{x_1,x_2,\ldots,x_n} = p_{x_1} \times p_{x_2} \times \ldots \times p_{x_n}$**) we have** $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$**.**

## Certain event vs. most disordered event

- **For any random variable $X$ with value set $\{x_1, x_2, \ldots, x_n\}$ and $\mathrm{Prob}(X = x_i) = p_i$, if for some $i$, $x_i$ is a certain event then $H(X) = 0$; so we have no ignorance about $X$!**

- **Thus, in the case of statement 1, as the phenomenon that the Sun will rise on the East is a certain event, the value "Sun will rise on the East" of a two-valued random variable $X$ occurs with probability $1$, while its complementary value "Sun will not rise on the East" does never occur. So $H(X) = -(1 \times \log_2 1 \ + 0 \times \log_2 0) = 0$.**

## Certain event vs. most ....

- **For equally probable events $x_1$, $x_2$, ..., $x_n$, we have $H(X) = \log_2 n$; so we have maximum ignorance about $X$.**
- **Thus, for example, in an experiment of throwing a unbiased coin once, as the probability of occurance of a 'head' is same as that of occurance of a 'tail', the average information content of the corresponding random variable $X$ (which takes the values 'head' or 'tail' with probability $1/2$) is given by $H(X) = -(2 \times \frac{1}{2} \times \log_2 \frac{1}{2}) = 1$.**
- **For all other probability distributions, $0 \leq H(X) \leq \log_2 n$.**
- **Thus, in the case of statement 3, we have: $0 < H(X) = -(\frac{1}{3} \times \log_2 \frac{1}{3} + \frac{2}{3} \times \log_2 \frac{2}{3}) = \log_2 3 - 2/3 < \log_2 2 = 1$.**

# Example 1

- **Given a proper fraction $p$, let $X$ be a $\{0,1\}$-valued random variable with $\mathrm{Prob}(X=0)=1-p$ and $\mathrm{Prob}(X=1)=p$. Then $H(X)=-p\log_2 p-(1-p)\log_2(1-p)\equiv H_2(p)$. Note that $H_2(p)$ is a *concave* function of the variable $p$ on the unit interval $0\leq p\leq 1$, *i.e.*, with $0\leq p_1\leq p_2\leq 1$ together with $0\leq\lambda\leq 1$, we have $H_2(\lambda p_1+(1-\lambda)p_2)\geq\lambda H_2(p_1)+(1-\lambda)H_2(p_2)$.**

## Example 2

- **Let $X$ be a $\{a, b, c, d\}$-valued random variable with $\text{Prob}(X = a) = 1/2,\ \text{Prob}(X = b) = 1/4,\ \text{Prob}(X = c) = 1/8, \text{Prob}(X = d) = 1/8$. Then $H(X) = 7/4$ bits.**
- **We now want to determine the value of $X$ with a min. no. of questions: e.g., "Is $X = a$?" If not, "is $X = b$?" If both are false, "is X = c?" Note that each question is of yes/no - type ($i.e.$, binary). The 1st question is single binary, 2nd is double binary, 3rd is triple binary with respective probabilities $1/2$, $1/4$, $1/8 + 1/8 = 1/4$ $\Rightarrow$ the expected no. of binary questions is $7/4$, which is same as $H(X)$!**
- **In general, for any random variable $X$, the expected no. of min. of of binary questions required to determine the value of $X$, lies between $H(X)$ and $H(X) + 1$.**

## Joint vs. conditional entropy

- **Joint entropy $H(X_1, X_2, \ldots, X_n)$: already defined.**
- **Conditional entropy:** $H(Y|X) \equiv \sum_x p_x \times H(Y|X = x) = -\sum_x p_x \times \sum_y p_{y|x} \times \log_2 p_{y|x} = -\sum_x p_{x,y} \times \log_2 p_{y|x}$.
- **Chain rule: Joint entropy of $(X, Y)$ is the sum of the entropy of $X$ (equivalently, $Y$) and the conditional entropy of $Y$ given $X$ (equivalently, $X$ given $Y$),** , *i.e.*, $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.
- **Prove it!**
- **As a consequence:** $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.
- **Prove it!**

## Example

- **Consider the motion of a system of two coupled classical harmonic oscillators of equal mass $m$, equal frequency $\omega$, both are moving along the $x$-axis under simple harmonic potentials $(1/2)m\omega^2 x_1^2$ and $(1/2)m\omega^2 x_2^2$, and where the coupling potential is $K(x_1 - x_2)^2$. So the total Hamiltonian of the system is $H = p_1^2/(2m) + p_2^2/(2m) + (1/2)m\omega^2 x_1^2 + (1/2)m\omega^2 x_2^2 + K(x_1 - x_2)^2$, where $p_i$ is the momentum of the $i$-th oscillator and $x_i$ being its position.**

## Example ....

- **Let $X_1$ be the random variable having values as the position ($x$, say) of the centre of mass of the entire systems, while $X_2$ be the random variable having values as the distance ($r$, say) of the two oscillators.**
- **Show that both the probability mass functions $p_x \equiv \text{Prob}(X_1 = x)$ and $p_r \equiv \text{Prob}(X_2 = r)$ are Gaussians. Show that the joint probability $p_{x,r} = p_x \times p_r$ (so $X_1, X_2$ are independent random variables). Show that $H(X_2|X_1) = H(X_1|X_2) = 0$.**
- **In general, $H(X|Y) \neq H(Y|X)$, but we always have: $H(X) - H(X|Y) = H(Y) - H(Y|X)$.**

## Shannon's entropy, relative entropy, mutual information

- **Shannon's entropy is all about probability distributions: same probability distributions give rise to the same entropy.**

- **How to compare two different probability distributions** $p_x = \text{Prob}(X = x)$ **and** $q_x = \text{Prob}(Y = x)$**? Relative entropy:** $D(p||q) \equiv \sum_x p_x \times \log_2\{p_x/q_x\}$**.**

- **Relative entropy is not a distance as it is not symmetric with respect to** $p$ **and** $q$**. But it is always non-negative and is equal to zero iff** $p_x = q_x$**.**

## Shannon's entropy, relative entropy ....

- **Mutual information $I(X;Y)$ is the measure of information content of one random variable about the other one. It is the relative entropy of the joint distribution $p_{x,y}$ and the product distribution $p_x \times p_y$: $I(X;Y) = \sum_x \sum_y p_{x,y} \times \log_2\{p_{x,y}/(p_x \times p_y)\}$.**
- **For finding out capacities of noisy channels, by looking at the outputs of the channel over all inputs, the mutual information of the inputs and the outputs is the quantity one should calculate – according to Shannon.**
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$**. (Prove it!) Thus $I(X:Y)$ is the reduction in the amount of uncertainty about $X$ due to the knowledge about $Y$.**
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$**. (Prove it!)**

## Example

- **Let $X$ and $Y$ be two $\{0,1\}$-valued random variables with $p_0 = \mathrm{Prob}(X = 0) = 1 - r$ and $q_0 = \mathrm{Prob}(Y = 0) = 1 - s$. Show that $D(p||q) = 0$ iff $p_0 = q_0$. When $r = 1/2$ and $s = 1/4$, show that $I(X;Y) = 0.375$ bits.**

## Chain rules for entropy, relative entropy and mutual information

- **Chain rule for entropy: For a string of random variables $X_1, X_2, \ldots, X_n$ with joint probability distribution $p_{x_1,x_2,\ldots,x_n}$, we have:**
$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$.

- **Conditional mutual information $I(X;Y|Z)$: It is the reduction in the uncertainty in $X$ due to the knowledge of $Y$ given $Z$. So $I(X;Y|Z) \equiv H(X|Z) - H(X|Y,Z)$.**

- **Chain rule for mutual information:**
$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1)$.

## Chain rules for entropy, relative entropy .....

- **Conditional relative entropy** **It is the average relative entropy of the conditional probabilities $p_{y|x}$ and $q_{y|x}$, averaged over $p_x$:** $D(p_{y|x}||q_{y|x}) \equiv \sum_x \sum_y p_{y|x} \times \log_2\{p_{y|x}/q_{y|x}\}$.

- **Chain rule for relative entropy:**
$D(p_{x,y}||q_{x,y}) = D(p_x||q_x) + D(p_{y|x}||q_{y|x})$.

## Jensen's inequality

- **If $f$ is a convex function and if $X$ is a random variable, we have: expectation of $f(X) \geq f(\text{expectation of } X)$. Moreover, if $f$ is strictly convex, then the equality implies that $X = \text{ expectation of } X$ with probability $1$, *i.e.*, $X$ is a constant random variable.**

- **Jensen's isequality can be used to prove the non-negativity of relative entropy: $D(p||q) \geq 0$ with equality holds iff $p_x = q_x$ for all $x$.**

- **As a consequence: $I(X;Y) \geq 0$ with equality iff $X$ and $Y$ are independent.**

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

- **It follows from the fact that** $I(X;Y) = H(X) - H(X|Y) \geq 0$**.**
- **But this result holds on an average: for some value** $y$**,** $H(X|Y = y)$ **may be greater than** $H(X)$**.**
- **Example: In a police investigation of a murder case, a new evidence regarding past history of the victim might increase the uncertainty about proceedings of the investigation, but on an average, any evidence decreases this uncertainty.**

## Convextity of relative entropy, concavity of entropy

- $D(p||q)$ **is jointly convex in** $p$ **and** $q$**:**

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

**for two pairs** $(p_1, q_1)$**,** $(p_2, q_2)$ **of probability mass distributions.**

- $H(X)$ **is a concave function:**

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

- **Example: Mixing of two gases of equal entropy (regarding the vel. distribution of the gas particles) results in a gas with higher entropy.**

# Information about next lecture

- **In the next lecture, we will discuss about Shannon's source coding theorem, which quantifies the minimum amout of bit space required, on an average, to store a large string of values of a random variable.**