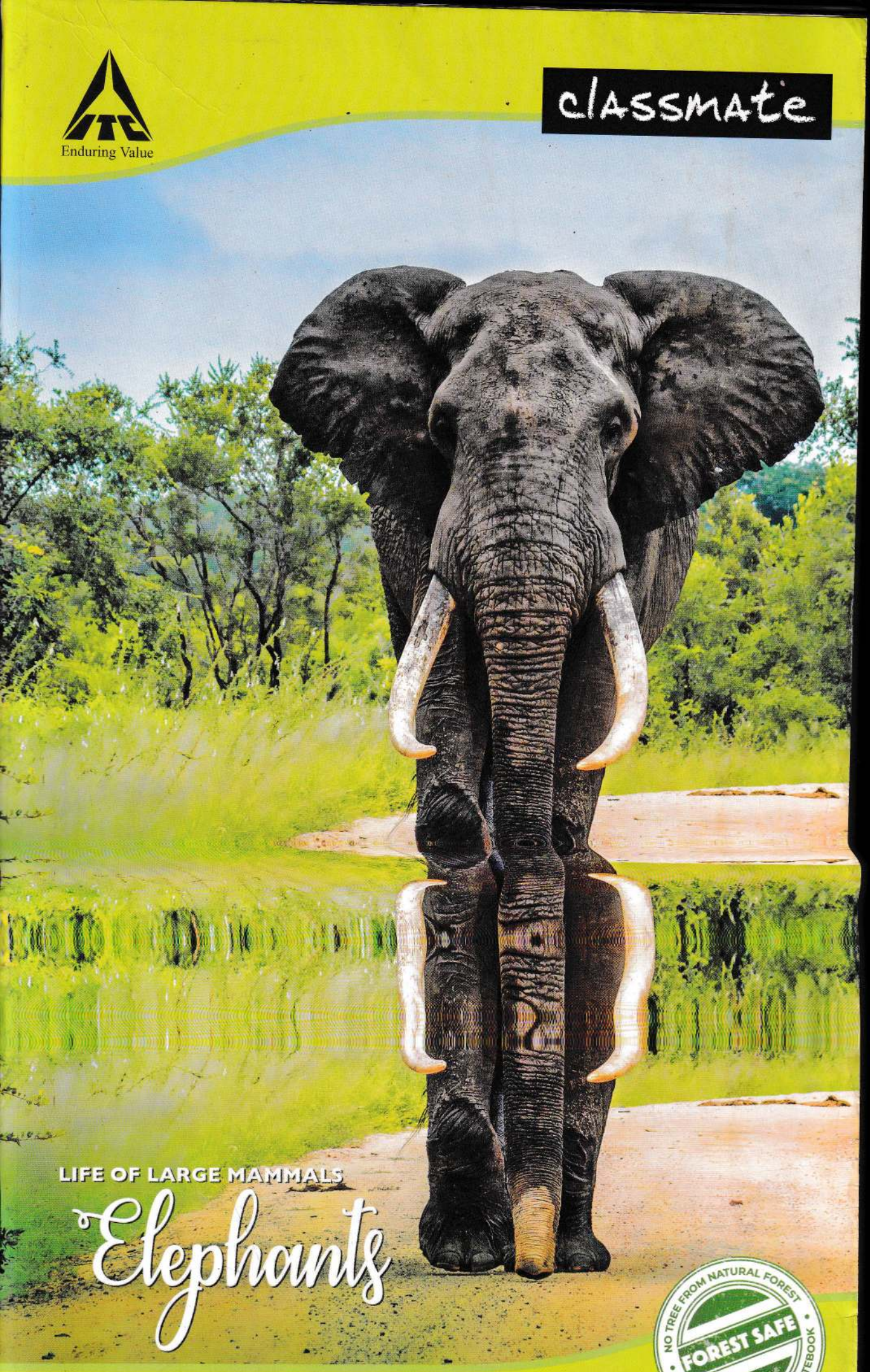
ITC
Enduring Value

classmate



LIFE OF LARGE MAMMALS

Elephants

NO TREE FROM NATURAL FOREST WAS CUT TO MAKE THIS NOTEBOOK
FOREST SAFE

Customize your notebook covers at
www.classmateshop.com

soorajss17129@gmail.com +91-9400635788

INDEX

11·1

NAME: SOORAJ·S· STD.: SEC.: ROLL NO.: SUB.:

S. No.	Date	Title	Page No.	Teacher's Sign / Remarks
		QUANTUM COMPOTATION & QUANTUM INFORMATION		
		– Nielsen & Chuang		

11

ENTROPY & INFORMATION

Suppose we learn the value of a random variable X. The Shannon entropy of X quantifies how much information we gain, on average, when we learn the value of X.

Alternatively, the entropy of X measures the amount of uncertainty about X before we learn its value.

Conceptually,

information can be thought of as being stored in or transmitted, as variables that can take on different values.

A variable can be thought of as a unit of storage that can take on, at different times, one of several different specified values, following some process for taking on those values.

Informally, we get information from a variable by looking at its value, and the information is about the process behind the variable.

The entropy of a variable is the "amount of information" contained in the variable.

This amount is determined not just by the # of different values the variable can take on, just as the information in an email is quantified not just by the # of words in the email or the different possible words in the language of the email.

Informally, the amount of information in an email is proportional to the amount of "surprise" its reading causes.

Ex:- if an email is simply a repeat of an earlier email, then it is not informative at all. On the other hand, if say the email reveals the outcome of a cliff-hanger election, then it is highly informative.

Similarly, the information in a variable is tied to the amount of surprise that value of the variable causes when revealed.

Shannon entropy quantifies the amount of information in a variable.

Storage and transmission of information can intuitively be expected to be tied to the amount of information involved.

Ex:- information may be about the outcome of a coin toss. This information can be stored in a Boolean variable that can take on the values 0 or 1. We can use the variable to represent

the raw data corresponding to the coin toss – whether the coin toss came up heads or not. This Boolean variable can be represented in a single bit. This bit directly stores the value of the variable, ie., the raw data corresponding to the outcome of the coin toss. It does not succinctly capture the information in the coin toss, e.g., whether the coin is biased or unbiased and, if biased, how biased.

Whereas, Shannon entropy quantifies, among other things, the absolute minimum amount of storage and transmission needed for succinctly capturing any information (as opposed to the raw data), and in typical cases that amount is less than what is required to store or transmit the raw data behind the information. Shannon entropy also suggests a way of representing the information in the calculated fewer # of bits.

the raw data corresponding to the coin toss – whether the coin toss came up heads or not. This Boolean variable can be represented in a single bit. This bit directly stores the value of the variable, ie., the raw data corresponding to the outcome of the coin toss. It does not succinctly capture the information in the coin toss, e.g., whether the coin is biased or unbiased and, if biased, how biased.

Whereas, Shannon entropy quantifies, among other things, the absolute minimum amount of storage and transmission needed for succinctly capturing any information (as opposed to the raw data), and in typical cases that amount is less than what is required to store or transmit the raw data behind the information. Shannon entropy also suggests a way of representing the information in the calculated fewer # of bits.

At a conceptual level, Shannon entropy is simply the amount of information in a variable. More mundanely, that translates to the amount of storage (e.g. # of bits) required to store the variable, which can intuitively be understood to correspond to the amount of information in that variable.

The calculation of the Shannon entropy, and therefore the amount of information in a variable, is not simply the # of bits required to represent all the different values a variable might take on, which is just the raw data.

Ex:-

a variable may take on any of 4 different values. In digital storage, 2 bits would be sufficient to uniquely represent the 4 different values, and thus the variable can be stored in 2 bits. However, this is an upper limit on the required storage; it is the amount of storage required to store the raw data of the variable, not the information in that data.

Less storage might be sufficient to store the information, depending on the process by which the variable takes on different values. Shannon entropy helps identify that amount of storage needed for the information.

→ Entropy can be looked as a measure of "compressibility" of the data.

ie,
a compression matric: how much can the raw data of a variable be compressed without losing the information in the variable?

Concept of Amount of information

Intuitively, one way to understand the concept of "amount of information" in a variable is to tie it to how difficult or easy it is to guess that information without having to look at the variable — the easier it is to guess the value of the variable, the less surprise is the variable and so the less information the variable has.

Another way to view information, is to contrast it with the amount of data. For example, two different Boolean variables could be stored in 1 bit each, but the amount of information in the two may be quite different:

Ex:-

① Suppose a coin is completely biased and always comes up heads when tossed. The random variable representing the coin toss's outcome has probability 1 of coming up heads (it is constant), and thus there is no need to store or transmit that variable as it can be trivially guessed at any time. ∴ The amount of information in that variable is zero.

② If we had a perfect coin with 50:50 chances of coming up heads or tails upon a coin toss, then we can guess the outcome of a toss with only 50% accuracy (probability = ½)

∴ it is necessary to store/transmit the actual value of that coin toss outcome's random variable in order to know its value with better than 50% accuracy.

∴ The amount of information in this 2nd random variable is much higher than in the 1st case.

Contrast the fact that, for both of the above coins, the raw data regarding their toss outcomes need 1 bit each to be stored.

③ Suppose we had a perfect die with 6 possible outcomes on the toss (roll) of the die. The amount of information in the corresponding random variable is even higher, as it is even harder to guess the outcome of the die roll, we have only $\frac{1}{6}^{th}$ chance of guessing the outcome correctly.

Quantifying the Amount of Information

One way to represent the "amount of information" is the # of bits it takes to represent/express the variable.

If a variable can take only 2 values, it can be represented with just 1 bit; if it can take on any of 4 values, 2 bits are needed; if 9 values then 4 bits are needed. etc.

But, this is the storage required for the raw data, not for the information content of the data.

If a variable is easier to guess, then we can leverage that fact to reduce the # of. bits needed to store/transmit that information.

Ex:-

If a die is 80% likely to come up with the # "3", then

store/transmit only single bit with the value 0 whenever the die actually comes up with "3", and store/transmit more bits, starting with a 1st bit of value 1, when the die comes up with other values.

When the 1st bit is a 0, the receiver knows not to look for more bits associated with this particular storage/transmission.

ie.,

While in a naive representation we would need 3 bits to represent each outcome of a die having 6 faces, but by leveraging probabilities we use only 1 bit whenever the die comes up with "3" - a more frequently occurring event; thus reducing the average # of bits needed (over multiple dice throws) to store the die throw outcomes.

One can leverage the probabilities of the different values to reduce the # of bits needed iff the variable has a non-uniform distribution. If not, there will be no reduction in the # of bits needed.

⇒ If a variable is more likely to take on one value than another, then it is easier to guess the value of the variable without looking at it, thus the variable has less information in it, and thus it takes fewer bits to store/transmit the value.

Example of Information Quantity

Suppose a variable can take on 3 different values a, b, c, but half the time it takes on the value 'a', and quarter of the time each the values 'b' and 'c'.

$$P(a) = 0.5 = 1/2 = 2/4$$

$$P(b) = 0.25 = 1/4$$

$$P(c) = 0.25 = 1/4$$

We can represent 'a' with just 1 bit having the value "0"; 'b' with 2 bits "10", and "c" with 2 bits "11". This is Huffman coding.

When the 1st bit is a "0", the receiver knows to stop reading that "word" right there; when the 1st bit is a "1", the reader knows to also read the next bit to complete the "word".

With this, what's the # of bits needed to represent this variable?

We need just 1 bit half the time (when the value is 'a'), and 2 bits each the other 2 times (when the value is either 'b' or 'c')

∴

The average # of bits needed is,

$$= P(a) \times 1 + P(b) \times 2 + P(c) \times 2$$

$$= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = \frac{3}{2} = 1.5$$

⟹ The entropy of the above variable having those specified probabilities of taking on different values is 1.5

The Entropy Formula

Rationale Behind using Logarithm for Entropy

Defining information

If we encode an outcome of higher probability in such a way that a smaller codeword (ie., a smaller piece of information) can represent it, we are better off.

Can we say that it is the length of the codeword that is its information ?

But in defining the length of the codeword that represents the outcome as its information we have not made use of the probability of the outcome at all which was seen to have a direct reciprocal effect on information that we needed to send.

Let's denote information with i and probability with p.

What if we say $i = 1/p$?

There are at least 2 problems now!

1st: if the probability of an outcome = 1, we don't need to send any information, and so i should be 0, which is not happening with $i = 1/p$.

Let's say we use $i = (1/p) - 1$

Say, the probability of Norway, Canada, Russia & Germany winning are all 0.25 each. We can encode all 4 of these outcomes using only 2 bits.
If we go by $i = 1/p$ (or) $i = 1/p - 1$, then i_{Norway} itself is 4 or 3.

Let's denote information with i and probability with p.

What if we say $i = 1/p$?

There are at least 2 problems now!

1st: if the probability of an outcome = 1, we don't need to send any information, and so i should be 0, which is not happening with $i = 1/p$.

Let's say we use $i = (1/p) - 1$

Say, the probability of Norway, Canada, Russia & Germany winning are all 0.25 each. We can encode all 4 of these outcomes using only 2 bits.
If we go by $i = 1/p$ (or) $i = 1/p - 1$, then i_{Norway} itself is 4 or 3.

At this point of time, we do have a relationship that uses the probability of the outcome, but the "information" we defined this way seem to have no connection with the length of the codeword. It depends only on probability now

- We are looking for a formula for i that should be related to the length of the codeword that represents the outcome, and that length be somehow related to the probability of the outcome.

For finding the length of the codeword in order to describe an outcome in an event, we need to know the total # of outcomes. For 2 outcomes, we need 1 bit. For 4, we need 2 bits, and so on.

How can we find the total # of outcomes using the probability of each outcome?

Say, we have an event that can have n possible outcomes with each outcome denoted as O_i having a probability P_i.
With $i = \log_2 n$ bits we'll be able to encode all the possible outcomes of this event, where $n = 2^i$.

What can be the upper limit of n if we know all the P_i's ?

- The smallest p solely limits the upper bound of n

Proof

Let P_j is the smallest. Then,

$P_i = P_j + a_i$ for every i, where $a_i = 0$ (or) a small +ve number.

$$1 = P_1 + P_2 + \cdots + P_n$$
$$= (P_j + a_1) + (P_j + a_2) + \cdots + (P_j + a_n)$$
$$= n P_j + (a_1 + a_2 + \cdots + a_n)$$

$$1 \geq n P_j$$

$$\Longrightarrow \boxed{n \leq \frac{1}{P_j}}$$

∴

The total # of outcomes of such an event cannot be more than $1/p$.

(O_p is the smallest probability outcome having a probability p)

$\Rightarrow i = \log_2 1/p$ bits can be used to encode all the outcomes of such an event, including O_p

$$n \leq \frac{1}{p} \Rightarrow \log_2 n \leq \log_2(1/p) = i$$

We'll be able to encode O_p and every other outcome in such an event using $\log_2(1/p)$ bits.

So, can we define information for an outcome that has a probability p, as $\log_2(1/p)$, in an event where this p is the smallest?

For an outcome with a probability p, $i = \log_2(1/p)$ bits are sufficient to encode it in an event in which this p is the smallest (among the probabilities of all other outcomes).

What if this p is not the smallest.

① Do these many bits still suffice for encoding this outcome? ($\log(1/p)$ bits).

For the # of outcomes n_p in an event in which p is the lowest probability outcome

$n_p \leq 1/p$ where we chose $i = \log_2(1/p)$

$\Longrightarrow n_p \leq 2^i$

$\Longrightarrow n_p = 2^i - \epsilon$, where ϵ is zero (or) a +ve integer.

Here, ϵ number of encodings are free
ie., not being used for representing any outcome.

Let's say we have added more outcomes to the event with p' being the new lowest probability outcome.

$\Rightarrow$ The max. number of outcomes now becomes $1/p'$.

Let, i' : new # of bits needed to code any of these $1/p'$ outcomes

Now, the max. # of additional outcomes will be given by,

$$\Delta n = n_{p'} - n_p \qquad \left[n_{p'} \leq 1/p' \right.$$

$$\leq \frac{1}{p'} - n_p$$

$$= \frac{1}{p'} - (2^i - \epsilon) \quad , \text{where } i = \log_2(1/p)$$

$$= \frac{1}{p'} - 2^i + \epsilon$$

$$\Delta n \leq 2^{i'} - 2^i + \epsilon \quad , \text{where } i' = \log_2(1/p')$$

2^i : # of outcomes that can be encoded using i bits

$2^{i'} - 2^i$: # of outcomes that can be encoded strictly using the additional $(i'-i)$ bits

ϵ : # of outcomes that are anyway encodable using the unused arrangement of i bits.

Ex:-

We added 5 new outcomes to 3 outcomes (encodable by 2 bits), and now we have 8 outcomes (encodable by 3 bits)

Among the 5 newly added outcomes $2^3 - 2^2 = 4$ is the max. # of outcomes that will strictly need the additional $3-2=1$ bit.

And the remaining 1 outcome was already encodable using 2 bits itself.

Thus,

the addition of 5 new outcomes (even of lower probability than any present before) doesn't affect the codability of the original 3 outcomes. Those are still encodable using $i=2$ bits only.

All the additional outcomes can be encoded using only the additional bits (and the unused arrangements of i bits if needed), and the original i bits are free to be used for encoding all the original outcomes.

$\Rightarrow$ For an outcome that has a probability p, $i = \log_2(1/p)$ bits can encode it in any event.

We have the length of the codeword that can represent an outcome based upon its probability p in any event – it is fair to call this length the information of the outcome.

$$i = \log_2(1/p)$$

⇒ The information i associated with any outcome of probability p is the minimum # of bits that can be used for encoding this outcome in any event having any # of outcomes.

And this # is solely given by the probability of the outcome as, $i = \log_2\left(\frac{1}{p}\right)$.

- This definition depends upon the unit we are using, but given a unit, we have a well defined information.

- The information of an outcome (ie, the amount of data needed to represent this outcome) is independent of the total # of outcomes of the event.

Ex:-

$p = 0.25 \Rightarrow i = \log_2\left(\frac{1}{0.25}\right) = \log_2(4) = 2$: we need 2 bits for encoding this outcome in any event (that can have 2, 4 or 4000 outcomes).

- It is more correct to name the specific units (along with information)

  Ex: a given state has i bits of information associated to it.

Information (or) Entropy of an Event (or) System

What would be the information content of any event, ie., the information of all its outcomes combined ?

What's the information content of a system that can have different states with different probabilities ?

- The information content of the event or the system is called Entropy.

- In statistics, such a system (event) that can have different states (outcomes) with different probabilities is called a probability distribution.

For example, if there are 3 states (outcomes) with information i_1, i_2 and i_3, what is the entropy of this system (event) ?

From the system's point of view these states are not equal, since these states or outcomes are not occurring with equal probability.

∴ We take the weighted average of the information (or entropy) of individual states of a system.

$$H = \sum_i P_i \, i$$

$$= \sum_i P_i \log_2 \left(1/P_i\right)$$

$$= -\sum_i P_i \log_2 (P_i)$$

⇒ The entropy of a system describes the minimum number of bits that can be used to describe any state of the system on average.

Shannon entropy

Suppose we learn the value of a random variable X. The Shannon entropy of X quantifies how much information we gain, on average, when we learn the value of X.

Alternatively, the entropy of X measures the amount of uncertainty about X before we learn its value.

Intuitively, the information content of a random variable should not depend on the labels attached to the different values that may be taken by the random variable.

Ex:-

We expect that a random variable taking the values 'heads' and 'tails' with respective probabilities $1/4$ and $3/4$ contains the same amount of information as a random variable that takes the values 0 and 1 with respective probabilities $1/4$ and $3/4$.

∴ The entropy of a random variable is defined to be a function of the probabilities of the different possible values of the random variable takes, and is not influenced by the labels used for those values.

We often write the entropy as a function of a probability distribution $P_1, P_2, \ldots, P_n$.

The Shannon entropy associated with the probability distribution $P_1, P_2, \ldots, P_n$ is defined by,

$$H(X) = H(P_1, P_2, \ldots, P_n) = -\sum_x P_x \log(P_x)$$

When $P_x = 0$,

$\log 0$ is undefined

Intuitively, an event which can never occur should not contribute to the entropy, so by convention we agree that $0 \log(0) \equiv 0$.

More formally, $\lim_{x \to 0} x \log(x) = 0$

- Uniform distributions have maximum uncertainty

- Uniform distributions with more outcomes have more uncertainty.

- The best reason for this definition of entropy is that it can be used to quantify the resources needed to store information.

More concretely,

Suppose there is some source (perhaps a radio antenna) which is producing information of some sort, say in the form of a bit string. Let's consider a very simple model for a source — we model it as producing a string $X_1, X_2, \ldots$ of independent, identically distributed random variables.

Most information sources don't behave quite this way, but it's often a good approximation to reality.

Shannon asked what minimal physical resources are required to store the information being produced by the source, in such a way that at a later time the information can be reconstructed?

ANS : Entropy

ie., $H(X)$ bits are required per source symbol.

where $H(X) = H(X_1) = H(X_2) = \ldots$ is the entropy of each random variable modeling the source. This result is known as Shannon's noiseless coding theorem.

Example

Suppose an information source produces one of the 4 symbols, 1, 2, 3 or 4. Without compression 2 bits of storage space corresponding to the 4 possible outputs are consumed for each use of the source.

Suppose that the symbol 1 is produced by the source with probability $1/2$, the symbol 2 with probability $1/4$, and the symbols 3 and 4 both with probability $1/8$.

We can make use of the bias b/w the source outputs to compress the source, using fewer bits to store commonly occurring symbols such as 1, and more bits to store rarely occurring symbols like 3 and 4.

One possible compression scheme is to encode 1 as the bit string 0, 2 as the bit string 10, 3 as the bit string 110, and 4 as the bit string 111.

The avg. length of the compressed string is,

$$\frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3 = \frac{7}{4} \text{ bits of}$$

information per use of the source.

This is less than is required in the naive approach to store this source.

This matches the entropy of the source,

$$H(X) = -\sum P_x \log(P_x)$$

$$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right)$$

$$= 7/4$$

Moreover, it turns out that any attempt to compress the source further results in data being irretrievably lost; the entropy quantifies the optimal compression that may be achieved.

Ex: 11.1 What's the entropy associated with the toss of a fair coin? With the roll of a fair die? How would the entropy behave if the coin or die were unfair?

Ans: $H(X) = -\sum_x P_x \log_2(P_x)$

(a) $H(X) = -\frac{1}{2}\log(1/2) - \frac{1}{2}\log(1/2) = 2 \times \frac{1}{2}\log 2 = 1$

(b) $H(X) = -\frac{1}{6}\log 1/6 - \frac{1}{6}\log 1/6 - \dots$

$= \log(6) \approx 2.58496 = 2.585$ bits.

(c) Unfair coin: $P(X = \text{heads}) = \frac{1}{4}$

$P(X = \text{tails}) = 3/4$

$H(X) = -\frac{1}{4}\log 1/4 - 3/4 \log 3/4$

$= \frac{1}{2} + \frac{3}{4}\log\frac{4}{3} = \frac{1}{2} + \frac{3}{4}\left[\log 4 - \log 3\right]$

$= \frac{1}{2} + \frac{3}{2} - \frac{3}{4}\log(3) = \frac{1}{2} + \frac{3}{2} - \frac{3}{4}(1.58496)$

$= 0.81128$

(d) Unfair die: $P(1) = \frac{1}{2}$,

$$P(2) = \frac{1}{30}, P(3) = \frac{2}{30}, \ldots, P(6) = \frac{5}{30}$$

1

$$H(X) = \frac{1}{2}\log(2) + \frac{1}{30}\log 30 + \frac{1}{15}\log 15 + \frac{1}{10}\log 10 + \frac{2}{15}\log\frac{15}{2} + \frac{1}{6}\log 6$$

$=$

2 $P(1) = 8\%$, $P(2) = 17\%$, $P(3) = P(4) = 25\%$, $P(5) = 17\%$, $P(6) = 8\%$

$H(X) = 2.454$ Bits

⇒ Entropy has the value based on probability and is maximum when the coin is fair. But when the coin is fair we have the highest uncertainty about the outcome of the experiment. Hence, we could state that entropy is the amount of uncertainty in a random variable.

Basic properties of Entropy

(A) The binary entropy

The entropy of a two-outcome random variable is called the binary entropy, which is defined as,

$$H_{bin}(p) \equiv -p \log p - (1-p) \log (1-p)$$

where p and $1-p$ are the probabilities of the two outcomes.

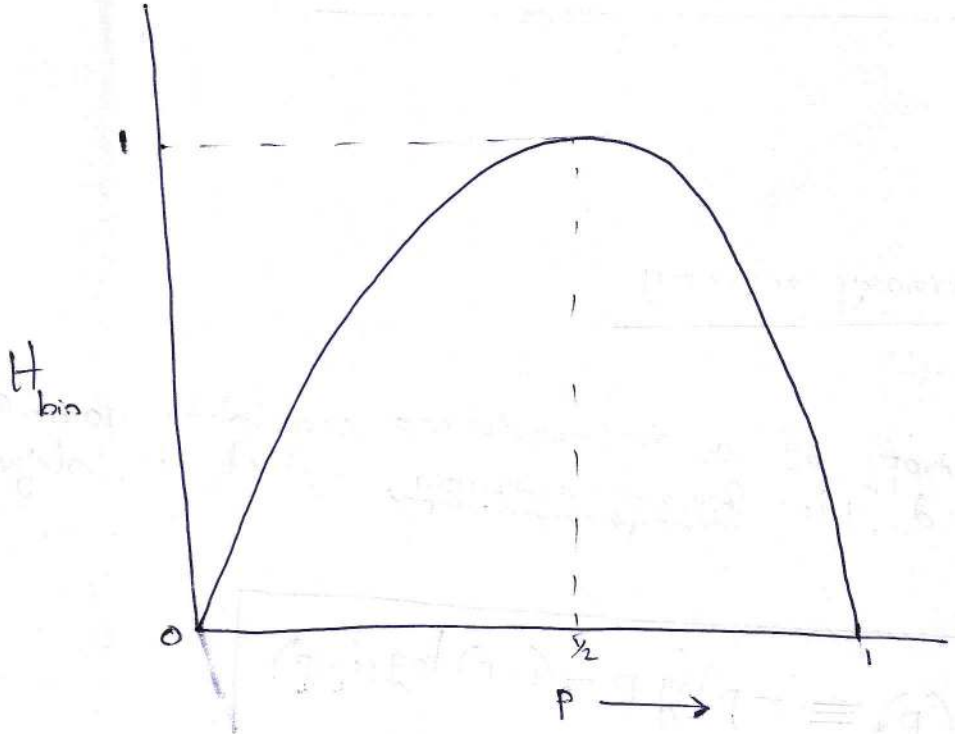
- $H(p) = H(1-p)$
- $H_{bin}^{max}(p) = 1 \quad$ at $p = 1/2$

Proof

$$H'_{bin}(p) = -1 - \log p + 1 + \log(1-p) = \log\left(\frac{1-p}{p}\right) = 0$$

$$\Rightarrow \frac{1-p}{p} = 1 \iff 1-p = p \Rightarrow 1 = 2p \Rightarrow p = 1/2$$

* Binary entropy function, $H(p)$



How the entropy behaves when we mix 2 or more probability distributions ?

Example

Alice has in her possession 2 coins, one a quarter from the US, the other a dollar coin from Australia. Both coins have been altered to exhibit bices with

P_U : probability of heads on the US coin

P_A : probability of heads on the Australian coin.

Suppose,

Alice flips the US coin with probability q and the Australian coin with probability $1-q$, telling Bob whether the result was heads or tails.

How much information does Bob gain on average ?

Intuitively, it is clear that Bob should gain at least as much information as the average of the information he should have gained from a US coin flip or an Australian coin flip.

This can be expressed as an equation,

$$H(qP_U + (1-q)P_A) \geq qH(P_U) + (1-q)H(P_A)$$

where,

P_U / P_A : probability of heads on the US/Australian coin.

$q\ (1-q)$: probability of flipping the US/Australian coin.

Note: Sometimes the inequality can be strict, because Bob gains information not only about the value (heads or tails) of the coin, but also some additional information about the identity of the coin.

Consider the probability vector,

$$P(i) = q P_U(i) + (1-q) P_A(i)$$

This corresponds to a process involving two sequencial choices: 1st a coin is tossed to decide b/w P_U and P_A, with the former picked up with probability q, and then another or the later with probability $(1-q)$, to decide b/w the possible outcome i according to the probabilities given by P_U or P_A.

① What's the information content corresponding to the possible outcomes?

In other words, completely disregarding the internal structure of P (meaning it is being actually composed of 2 sequencial choices), what is its entropy?

$\Rightarrow H(P) = H(qP_U + (1-q)P_A)$

(2) If you already know the outcome of the 1st choice (ie., you know whether you are in the P_U or in the P_A branch) what's the information content corresponding to the possible outcomes?

This is $H(P_U)$ if we are in the P_U branch or $H(P_A)$ if we are in the other. If we don't want to assume we already know the outcome of the 1st coin toss, we ought to consider the (weighted) average of such entropy, which is

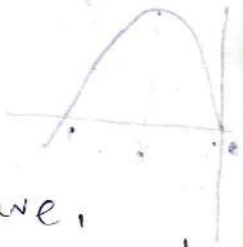
$$q\,H(P_U) + (1-q)\,H(P_A)$$

* The binary entropy is a concave function.

$$H_{bin}(px_1 + (1-p)x_2) \geq p\,H_{bin}(x_1) + (1-p)\,H_{bin}(x_2)$$

where $0 \leq p, x_1, x_2 \leq 1$.

⇒ Mixing 2 probability distributions increases the entropy.



The binary entropy is strictly concave,
ie., the above inequality is an equality for the trivial cases

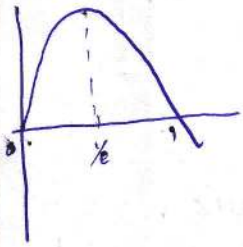
Proof.

$$H_{bin}\left(p x_1 + (1-p) x_2\right) = -\sum_i \left(p x_1(i) + (1-p) x_2(i)\right) \log\left(p x_1(i) + (1-p) x_2(i)\right)$$

$$\geq \sum_i \left(-p x_1(i) \log\left(p x_1(i)\right) - (1-p) x_2(i) \log\left((1-p) x_2(i)\right)\right)$$

$-x \log x$ is concave →

$$\geq \sum_i \left(-p x_1(i) \log\left(x_1(i)\right) - (1-p) x_2(i) \log\left(x_2(i)\right)\right)$$



$$= -p \sum_i x_1(i) \log\left(x_1(i)\right) - (1-p) \sum_i x_2(i) \log\left(x_2(i)\right)$$

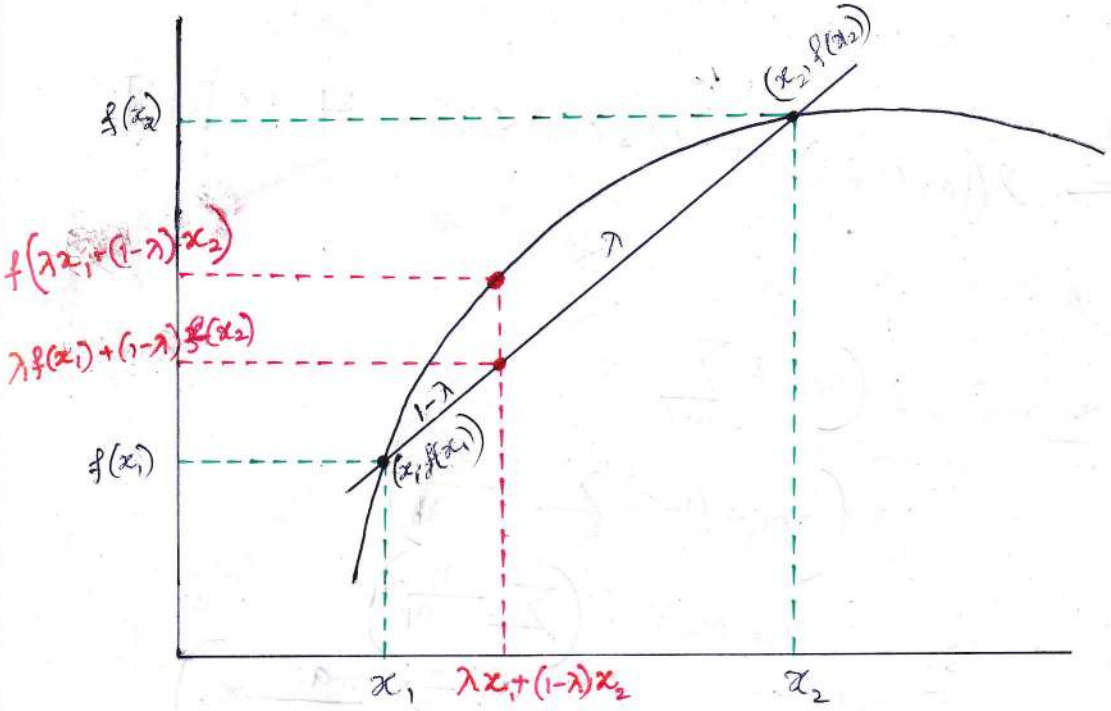
$$= p\, H_{bin}(x_1) + (1-p)\, H_{bin}(x_2)$$

Concave function

A real valued function f on an interval (a,b) is called concave, provided for each pair of points $x_1, x_2 \in (a,b)$ and each λ with $0 \le \lambda \le 1$, we have

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall \lambda \in [0,1]$$

ie., a line connecting any 2 points on the function lies below or on the function.



* If f is a concave function of one real variable. Let $x_1, \dots, x_n \in \mathbb{R}$ and $a_1, \dots, a_n \geq 0$ satisfy $a_1 + \cdots + a_n = 1$. Then,

$$f(a_1x_1 + \cdots + a_nx_n) \geq a_1f(x_1) + \cdots + a_nf(x_n)$$

$$\Updownarrow$$

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \geq \sum_{i=1}^{n} a_i f(x_i)$$

Proof $f(ta + (1-t)b) \geq tf(a) + (1-t)f(b) \quad \forall\, t \in [0,1]$

Stark 1/1/2023.

$$f\left(\sum_{i=1}^{n} a_i x_i\right) = f\left(a_1x_1 + \sum_{i=2}^{n} a_i x_i\right)$$

$$= f\left(a_1x_1 + (1-a_1)\sum_{i=2}^{n} \frac{a_i x_i}{1-a_1}\right)$$

$$\geq a_1 f(x_1) + (1-a_1) f\left(\sum_{i=2}^{n} \frac{a_i}{1-a_1} x_i\right)$$

$$= a_1 f(x_1) + (1-a_1) f\left(\frac{a_2}{1-a_1}x_2 + \sum_{i=3}^{n} \frac{a_i}{1-a_1} x_i\right)$$

$$= a_1 f(x_1) + (1-a_1) f\left(\frac{a_2}{1-a_1}x_2 + \frac{1-a_1-a_2}{1-a_1}\sum_{i=3}^{n} \frac{1-a_1}{1-a_1-a_2}\cdot\frac{a_i}{1-a_1} x_i\right)$$

$$\geq a_1 f(x_1) + (1-a_1)\left[\frac{a_2}{1-a_1} f(x_2) + \frac{1-a_1-a_2}{1-a_1} f\left(\sum_{i=3}^{n} \frac{a_i}{1-a_1-a_2} x_i\right)\right]$$

$$= a_1 f(x_1) + a_2 f(x_2) + (1-a_1-a_2) f\left(\sum_{i=3}^{n} \frac{a_i}{1-a_1-a_2} x_i\right)$$

$$\geq a_1 f(x_1) + \cdots + a_{n-1} f(x_{n-1}) +$$

$$(1-a_1-\cdots-a_{n-1}) f\left(\sum_{i=n}^{n} \frac{a_i}{1-a_1-\cdots a_{n-1}} x_i\right)$$

$$= a_1 f(x_1) + \cdots + a_{n-1} f(x_{n-1}) + a_n f(x_n) = \sum_{i=1}^{n} a_i f(x_i)$$

---

* Concavity of f on (a,b) implies that the slope of the chord from $(x_1, f(x_1))$ to $(x, f(x))$ is greater than or equal to the slope of the chord from $(x, f(x))$ to $(x_2, f(x_2))$, ie.,

$$\frac{f(x)-f(x_1)}{x-x_1} \geq \frac{f(x_2)-f(x)}{x_2-x} \quad \forall \quad x_1 < x < x_2 \text{ in } (a,b)$$

Proof

(With $x_1 < x_2$ in (a,b) and $x \in (x_1, x_2)$, we have $x = \lambda x_1 + (1-\lambda) x_2$ for $\lambda = \frac{x_2 - x}{x_2 - x_1}$

$$f(\lambda x_1 + (1-\lambda) x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$

$$f(x) \geq \left(\frac{x_2 - x}{x_2 - x_1}\right) f(x_1) + \left(\frac{x - x_1}{x_2 - x_1}\right) f(x_2)$$

$$(x_2 - x_1) f(x) \geq (x_2 - x) f(x_1) + (x - x_1) f(x_2)$$

$$(x_2 - x + x - x_1) f(x) \geq (x_2 - x_1) f(x_1) + (x - x_1) f(x_2)$$

$$(x_2 - x) f(x) + (x - x_1) f(x) \geq (x_2 - x_1) f(x_1) + (x - x_1) f(x_2)$$

$$(x_2 - x)(f(x) - f(x_1)) \geq (x - x_1)(f(x_2) - f(x))$$

$$\frac{f(x) - f(x_1)}{x - x_1} \geq \frac{f(x_2) - f(x)}{x_2 - x} \quad \forall \; x_1 < x < x_2 \text{ in } (a,b).$$

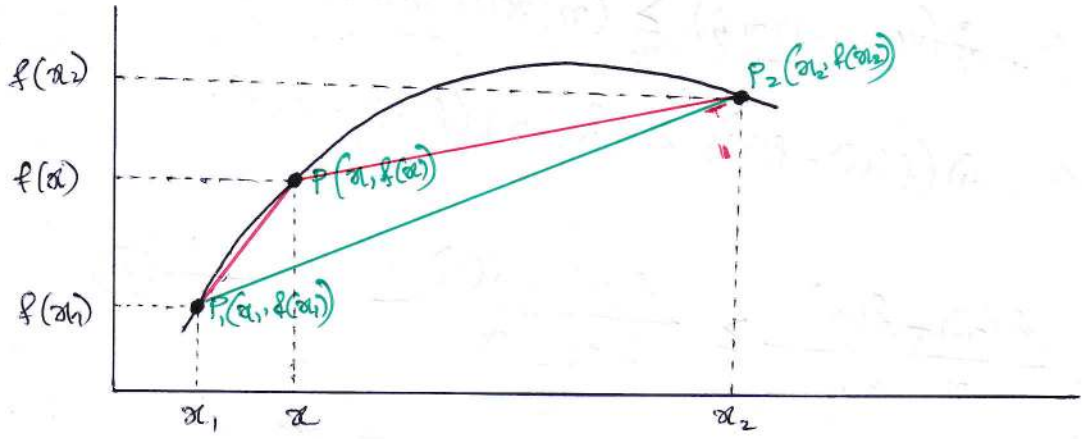
---

* The Chordal Slope lemma

Let f be concave on (a,b). If $x_1 < x < x_2$ are in (a,b) then for points $P_1 = (x_1, f(x_1))$, $P = (x, f(x))$ and $P_2 = (x_2, f(x_2))$ we have

$$\frac{f(x)-f(x_1)}{x-x_1} \geq \frac{f(x_2)-f(x_1)}{x_2-x_1} \geq \frac{f(x_2)-f(x)}{x_2-x} \quad \forall \; x_1 < x < x_2 \text{ in } (a,b)$$

$\Rightarrow$ Slope of $\overline{P_1P} \geq$ Slope of $\overline{P_1P_2} \geq$ Slope of $\overline{PP_2}$



Proof

$$f(x) \geq \left(\frac{x_2-x}{x_2-x_1}\right) f(x_1) + \left(\frac{x-x_1}{x_2-x_1}\right) f(x_2)$$

$$(x_2-x_1) f(x) \geq (x_2-x) f(x_1) + (x-x_1) f(x_2)$$

$$(x_2-x_1) f(x) \geq (x_2-x_1+x_1-x) f(x_1) + (x-x_1) f(x_2)$$

$$= (x_2-x_1) f(x_1) + (x-x_1)\left(f(x_2)-f(x_1)\right)$$

$$(x_2 - x_1)(f(x) - f(x_1)) \geq (x - x_1)(f(x_2) - f(x_1))$$

$$\frac{f(x) - f(x_1)}{x - x_1} \geq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \quad \text{---- (1)}$$

Similarly,

$$(x_2 - x_1) f(x) \geq (x_2 - x) f(x_1) + (x - x_1) f(x_2)$$

$$= (x_2 - x) f(x_1) + (x - x_2 + x_2 - x_1) f(x_2)$$

$$= (x_2 - x) f(x_1) + (x - x_2) f(x_2) + (x_2 - x_1) f(x_2)$$

$$(x_2 - x_1)(f(x) - f(x_2)) \geq (x_2 - x)(f(x_1) - f(x_2))$$

$$(x_2 - x_1)(f(x_2) - f(x)) \leq (x_2 - x)(f(x_2) - f(x_1))$$

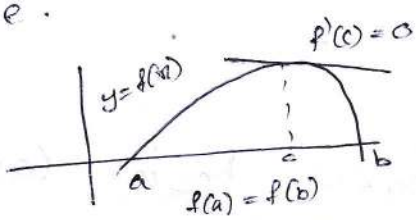
$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq \frac{f(x_2) - f(x)}{x_2 - x} \quad \text{---- (2)}$$

(1) & (2) $\Rightarrow$

$$\frac{f(x) - f(x_1)}{x - x_1} \geq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq \frac{f(x_2) - f(x)}{x_2 - x}$$

* If f is differentiable on (a,b) and its derivative f' is decreasing, then f is concave.

If f'' exists on (a,b) and $f'' \leq 0$ on (a,b) then f is concave.

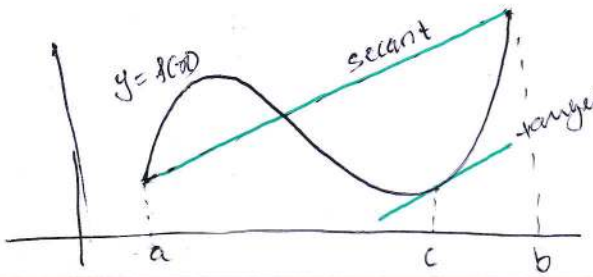


Proof

Rolle's theorem: If a real valued function f is continuous on a proper closed interval $[a,b]$, differentiable on the open interval (a,b) and $f(a) = f(b)$, then there exists at least one c in (a,b) such that $f'(c) = 0$.

ie., any real valued differentiable function that attains equal values at 2 distinct points must have at least one stationary point somewhere between them.

Mean value theorem: If f is a continuous function on the closed interval $[a,b]$ and differentiable on the open interval (a,b), then there exists a point $c \in (a,b)$ such that the tangent at c is ||el to the secant line thro' the end points $(a, f(a))$ and $(b, f(b))$, ie., $f'(c) = \frac{f(b) - f(a)}{b - a}$



Let $x_1 < x_2$ be in (a,b) and let $x \in (x_1, x_2)$.

f is differentiable $\Longrightarrow$ Mean value theorem applies to f on intervals $[x_1, x]$ and $[x_1, x_2]$.

$\therefore$

Choose $c_1 \in (x_1, x)$ and $c_2 \in (x_1, x_2)$ for which

$$f'(c_1) = \frac{f(x) - f(x_1)}{x - x_1} \quad \text{and} \quad f'(c_2) = \frac{f(x_2) - f(x)}{x_2 - x}.$$

Since f' is decreasing then,

$$\frac{f(x) - f(x_1)}{x - x_1} = f'(c_1) \geq f'(c_2) = \frac{f(x_2) - f(x)}{x_2 - x}$$

$\Longrightarrow$ f is concave.

(B) The relative entropy (Kullback-Leibler divergence)

Suppose $p(x)$ and $q(x)$ are 2 probability distributions on the same index set, x. The relative entropy of $p(x)$ to $q(x)$ is defined as,

$$H(p(x) \| q(x)) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\equiv \sum_x p(x) \left(\log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right)$$

$$\equiv -\sum_x p(x) \log q(x) + \sum_x p(x) \log p(x)$$

$$\equiv -\sum_x p(x) \log q(x) - H(X)$$

* An outcome x that has a probability $p(x)$, we can encode it in any event using $i = \log \frac{1}{p(x)}$ bits.

• We use the convention that $-0 \log 0 \equiv 0$ & $-p(x) \log 0 = +\infty$ if $p(x) > 0$

• $0 \log \frac{0}{0} = 0$ and based on continuity arguments that: $0 \log \frac{0}{q} = 0$ & $p \log \frac{p}{0} = \infty$.

$\Rightarrow H(p(x)\|q(x)) = \infty$ if there exists any $x \in X$ such that $p(x) > 0$ and $q(x) = 0$

Since,

$$H(p(x)\|q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log \frac{q(x)}{p(x)}$$

and $-p(x) \log 0 = \infty$ if $p(x) > 0$.

* $H(P(x)||Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = \sum_x P_x \left(\log \frac{1}{Q(x)} - \log \frac{1}{P(x)} \right)$

An outcome x that has a probability $P(x)$, we can encode it in any event using $i = \log_2 \frac{1}{P(x)}$ bits.

⇒ Relative entropy $H(P(x)||Q(x))$ is the # of extra bits required per letter on average to encode a source with a distribution $Q(x)$ when the true underlying distribution is $P(x)$.

ie.,
expected extra-message length per datum that must be communicated if a code that is optimal for a given (wrong) distribution $Q(x)$ is used, compared to using a code based on the true distribution $P(x)$. : it is the excess entropy.

ie.,
If we knew the true distribution of the random variable, then we could construct a code with average description length $H(P)$.
If instead, we used the code for a distribution Q, we would need $\sum_x P(x) \log \frac{1}{Q(x)} = H(P) + H(P||Q)$ bits on the average to describe the random variable.

- The relative entropy / Kullback-Leibler divergence is a very useful entropy-like measure of the closeness of 2 probability distributions, $p(x)$ and $q(x)$, over the same index set x.

Relative entropy is always non-negative and is zero if and only if $P = Q$.

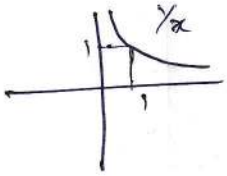
However, it is not a true distance b/w distributions since it is not symmetric and does not satisfy the triangle inequality — it is not a metric.

* The relative entropy is non-negative

$\boxed{H(p(x)\|q(x)) \geq 0}$, with equality iff $p(x)=q(x)$ for all x. ②

Proof.

Lemma: $\ln(x) \leq x-1$ for all $x>0$ with $\ln(x)=x-1$ iff $x=1$



?

Proof

case 1: $0<x<1 \rightarrow \frac{1}{x}>1$ ③

$$\Rightarrow \int_x^1 \frac{dx}{x} \geq \int_x^1 dx \Rightarrow \left[\ln(x)\right]_x^1 = \left[x\right]_x^1$$

$$\Rightarrow \ln(1)-\ln(x) \geq 1-x$$

$$\Rightarrow -\ln(x) \geq 1-x \Rightarrow \underline{\ln(x) \leq x-1}.$$

case 2: $x \geq 1 \Rightarrow \frac{1}{x} \leq 1$

$$\Rightarrow \int_1^x \frac{dx}{x} \leq \int_1^x dx \Rightarrow \left[\ln(x)\right]_1^x \leq \left[x\right]_1^x$$

$$\Rightarrow \ln(x)-\ln(1) = \underline{\ln(x) \leq x-1}$$

$\therefore \quad \ln(x) \leq x-1 \quad \forall \quad x>0$

② For $x>0$,

$f(x) = \ln(x) - x + 1 \Rightarrow f'(x) = \frac{1}{x} - 1$

$f'(x) = 0 \Rightarrow x = 1$

$f''(x) = \frac{-1}{x^2} < 0 \quad \forall \quad x > 0$

$\therefore \quad x = 1$ is a maximum.

? $\lim_{x \to 0^+} f(x) = -\infty = \lim_{x \to \infty} f(x)$

$x=1$ is global maximum

$\forall \; x>0, \quad f(x) \leq f(1) = 0$

$\therefore \quad \ln(x) - x + 1 \leq 0 \Rightarrow \ln(x) \leq x-1 \quad \forall \; x>0.$

③. $\ln(x) \leq x-1 \Rightarrow x \leq e^x - 1 \Rightarrow e^x \geq x+1$ By putting e^x for x.
$\forall x$

$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$

Bernoulli's inequality

$(1+t)^n \geq 1+nt$ if $t > -1$ and $n > 0$.

$\frac{x}{n} > 0 \Rightarrow \left(1 + \frac{x}{n}\right)^n \geq 1 + x \Rightarrow e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n \geq 1 + x.$

$\Rightarrow \ln(x) \leq x-1 \quad \forall \quad x>0$

$$\frac{\ln(x)}{\ln(2)} = \log_2(x)$$

$$\Rightarrow \log(x)\ln(2) = \ln(x) \leq x-1 \quad \forall\, x>0$$

with equality iff $x=1$.

$$-\log(x)\ln(2) \geq 1-x \Rightarrow -\log(x) \geq \frac{1-x}{\ln(2)}$$

$\therefore$

$$H(p(x)\|q(x)) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

$$= -\sum_x p(x)\log\left(\frac{q(x)}{p(x)}\right)$$

$$\geq \frac{1}{\ln(2)}\sum_x p(x)\left(1-\frac{q(x)}{p(x)}\right)$$

$$= \frac{1}{\ln(2)}\sum_x \left(p(x)-q(x)\right)$$

$$H(p(x)\|q(x)) \geq \frac{1}{\ln(2)}\left[\sum_x p(x) - \sum_x q(x)\right]$$

$$= \frac{1}{\ln(2)}(1-1) = 0$$

$$\Longrightarrow H(p(x)\|q(x)) \geq 0$$

$$\Longrightarrow H(p(x)\|q(x)) = 0 \quad \text{iff} \quad q(x)/p(x) = 1 \quad \forall x$$

$$p(x) = q(x) \quad \forall x$$

* A divergence is a function (binary) that takes 2 probability distributions as input, and returns a # that measures how much they differ. The # returned must be non-negative, and equal to zero iff the 2 distributions are identical.

Bigger numbers indicate greater dissimilarity.

In addition to the requirements above for the divergence, a distance metric must also be symmetric: $D(a,b) = D(b,a)$. And, it must satisfy the triangle inequality: $D(a,c) \leq D(a,b) + D(b,c)$

Divergences are defined specifically on probability distributions, whereas distance metric can be defined on other types of objects too.

All distance metrics b/w probability distributions are also divergences, but the converse is not true.

→ divergences measure the dissimilarity b/w distributions.

The simplest divergence is squared Euclidean distance (SED) and divergences can be viewed as generalizations of SED.

The relative entropy is often useful, not in itself but because other entropic quantities can be regarded as special cases of the relative entropy.

* Suppose X is a random variable with d outcomes. Then, $\boxed{H(X) \leq \log(d)}$ with equality iff X is uniformly distributed over those d outcomes.

Proof

Suppose $p(x)$ is a probability distribution for X, over d outcomes.

Let $q(x) \equiv 1/d$ be the uniform probability distribution over those outcomes. Then,

$$H(p(x)\|q(x)) = H(p(x)\|1/d) = -\sum_x p(x)\log q(x) - H(X)$$

$$= -\sum_x p(x)\log(1/d) - H(X)$$

$$= -\log(1/d)\sum_x p(x) - H(X)$$

$$= \log d - H(X) \geq 0$$

$$\Rightarrow \underline{H(X) \leq \log d}$$

Ex: 11.5 Subadditivity of the Shannon entropy

$$\boxed{H(X,Y) \leq H(X) + H(Y)}$$

with equality iff X and Y are independent random variables.

Proof

$$H(X)+H(Y)-H(X,Y) = -\sum_x p(x)\log p(x) - \sum_y p(y)\log p(y) + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= -\sum_{x,y} p(x,y)\log p(x) - \sum_{x,y} p(x,y)\log p(y) + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)\,p(y)} = H\left(p(x,y)\,||\,p(x)\,p(y)\right) \geq 0$$

Equality is achieved iff $p(x,y) = p(x)\,p(y)$ $\forall\, x,y$

$\Rightarrow$ the subadditivity inequality is saturated if and only if X and Y are independent.

© Conditional entropy & mutual information

Suppose X and Y are two random variables. How is the information content of X related to the information content of Y ?

The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

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

The joint entropy measures our total uncertainty about the pair (X,Y).

There is nothing really new in this definition because (X, Y) can be considered to be a single vector-valued random variable.

Suppose we know the value of Y, so we have acquired $H(Y)$ bits of information about the pair, (X,Y). The remaining uncertainty about the pair (X,Y), is associated with our remaining lack of knowledge about X, given that we know Y.

The entropy of X conditional on knowing Y is defined as,

$$H(X|Y) = \sum_y p(y)\, H(X|Y=y)$$

$$= \sum_y p(y) \left[-\sum_x p(x|y) \log p(x|y) \right]$$

$$= -\sum_{x,y} p(y)\, p(x|y) \log p(x|y)$$

$$= -\sum_{x,y} p(x,y) \log p(x|y)$$

⇒ The conditional entropy $H(X|Y)$ is a measure of how uncertain we are, on average, about the value of X, given that we know the value of Y.

* $$H(x,y) = H(x) + H(y|x)$$
$$= H(y) + H(x|y)$$

Proof

$$H(x,y) = -\sum_{x,y} p(x,y) \log p(x,y)$$

$$= -\sum_{x,y} p(x,y) \log p(y)\, p(x|y)$$

$$= -\sum_{x,y} p(x,y) \log p(y) - \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{y} \left(\sum_{x} p(x,y) \right) \log p(y) - \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{y} p(y) \log p(y) - \sum p(x,y) \log p(x|y)$$

$$= H(y) + H(x|y)$$

The mutual information $H(X:Y)$ content of X and Y, measures how much information X and Y have in common.

ie.,
is a measure of the amount of information that one random variable contains about another random variable.

Suppose we add the information content of X, $H(X)$, to the information content of Y. Information which is common to X and Y will have been counted twice in this sum, while information which is not common will have been counted exactly once.

Subtracting off the joint information of (X,Y), $H(X,Y)$, we obtain the common or mutual information of X and Y:

$$H(X:Y) \equiv H(X) + H(Y) - H(X,Y)$$

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

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

$\therefore$

$$\boxed{\begin{aligned} H(X:Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}}$$

$\Rightarrow$ The mutual information is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Ex: 11.5

Ex: 11.7

Ans:

QE Stalk
23/12/2022

* Consider 2 random variables X and Y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $q(x)$. The mutual information $H(X:Y)$ can be defined as the relative entropy b/w the joint distribution and the product distribution $p(x)p(y)$,

$$H(X:Y) = H\big(p(x,y) \,\|\, p(x)\,p(y)\big)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

Ex: 11.5

Ex: 11.7 Find an expression for the conditional entropy $H(Y|X)$ as a relative entropy b/w 2 probability distributions. Use this to deduce that $H(Y|X) \geq 0$, and find the equality conditions.

Ans:

$$H(Y|X) = \sum_x p(x)\, H(Y|X=x)$$

$$= \sum_x p(x) \times -\sum_y p(y|x) \log p(y|x)$$

$$= -\sum_{x,y} p(x)\,p(y|x) \log p(y|x)$$

$$= -\sum_{x,y} p(x,y) \log p(y|x)$$

QE Statlk 23/12/2022

$$P(y|x) = \frac{P(x,y)}{P(x)} \quad \Longleftarrow \quad P(x,y) = P(x)\,P(y|x)$$

$$H(Y|X) = -\sum_{x,y} p(x,y) \log \frac{P(x,y)}{P(x)}$$

$$= -\sum_{x,y} p(x,y) \log \frac{P(x,y)}{P(x) \times \frac{1}{d_y}} \times \frac{1}{d_y}.$$

$$H(P(x) \| q(x)) = \sum_x P(x) \log \frac{P(x)}{q(x)}$$

$P(x)$ & $q(x)$ must be defined over the same set.

where d_y: # of outcomes for the random variable Y.

$$= -\sum_{x,y} p(x,y) \log \frac{P(x,y)}{P(x) \times \frac{1}{d_y}} + \sum_{x,y} p(x,y) \log(d_y)$$

We have the joint distribution $p(x,y)$ b/w 2 random variables X and Y, and consider a distribution q which is defined as

$$q(x,y) := P(x) \times \frac{1}{d_y} = \frac{P(x)}{d_y}.$$

ie., q is a product distribution of the marginal distribution of X together with a uniform distribution over the outcome space of Y.

$\therefore$

$$H(Y|X) = -H\left(p(x,y) \,\|\, p(x)/d_y\right) + \log(d_y)$$

Basic properties of Shannon entropy

1. $H(X,Y) = H(Y,X)$, $H(X:Y) = H(Y:X)$

2. $H(Y|X) \geq 0 \longrightarrow H(X:Y) \leq H(Y)$
with equality iff Y is a function of X, $Y = f(X)$

3. $H(X) \leq H(X,Y)$
with equality iff Y is a function of X

4. Subadditivity : $\boxed{H(X,Y) \leq H(X) + H(Y)}$
with equality iff X and Y are independent random variables. Ex: 11.5

5. $H(Y|X) \leq H(Y)$ and thus $H(X:Y) \geq 0$.
with equality in each iff X and Y are independent random variables.

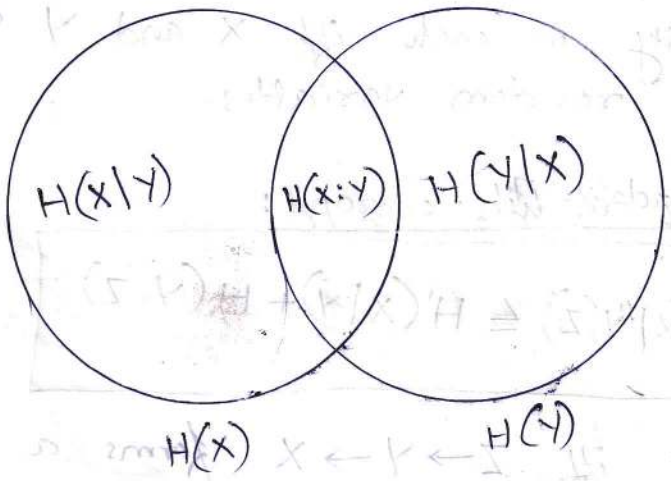
6. Strong subadditivity :

$$\boxed{H(X,Y,Z) + H(Y) \leq H(X,Y) + H(Y,Z)}$$

with equality iff $Z \to Y \to X$ forms a markov chain.

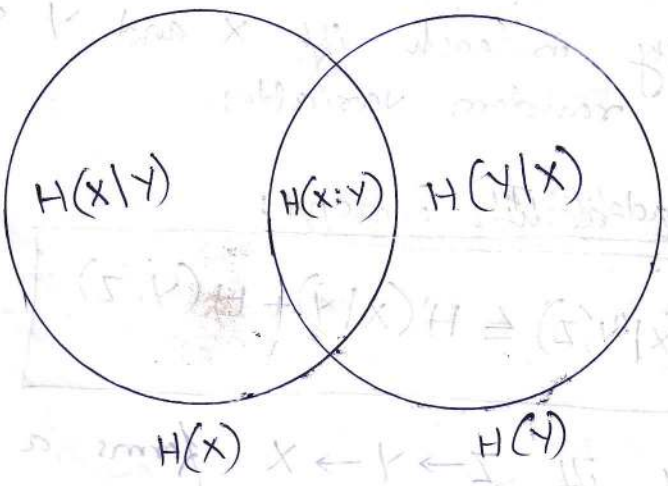
reduces entropy.

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



7. Conditioning reduces entropy:

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



Proof

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

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

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

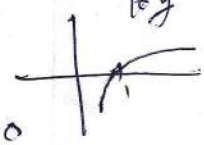
$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$

2. $H(X:Y) = H(Y) - H(Y|X)$

$$H(Y|X) = \sum_x p(x)\, H(Y|X=x)$$

$$= \sum_x p(x) \cdot x - \sum_y p(y|x) \log p(y|x)$$

$$= -\sum_{x,y} p(x,y) \log p(y|x)$$



$0 \le p(y|x) \le 1 \Rightarrow \log p(y|x) \le 0 \Rightarrow -\log p(y|x) \ge 0$

$\therefore H(Y|X) \ge 0$

$H(X:Y) = H(Y) - H(Y|X) \Longrightarrow H(Y) - H(X:Y) \ge 0$

$\Rightarrow H(X:Y) \le H(Y)$

$H(Y|X)=0$ Equality iff Y is a deterministic function of X.

i.e., knowing X determines the value of Y and vice versa.

$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = 0 \Longrightarrow$ whenever $p(X=x, Y=y) > 0$ we have $p(Y=y|X=x) = 1$

$H(Y|X) = 0 \Longrightarrow H(X:Y) = H(Y)$

$\therefore$ The mutual information is the same as the uncertainty contained in Y alone, namely the entropy of Y.

- $H(X) \geq 0$

  $H(X) = 0$ iff X is deterministic.

3. $H(X,Y) = -\sum_{x,y} P(x,y) \log P(x,y)$

$= -\sum_{x,y} P(x,y) \log P(x) P(y|x)$

$= -\sum_{x,y} P(x,y) \log P(x) - \sum_{x,y} P(x,y) \log P(y|x)$

$= -\sum_{x} \left(\sum_{y} P(x,y)\right) \log P(x) - \sum_{x,y} P(x,y) \log P(y|x)$

$= -\sum_{x} P(x) \log P(x) - \sum_{x,y} P(x,y) \log P(y|x)$

$= H(X) - \sum_{x,y} P(x,y) \log P(y|x)$

$P(y|x) \leq 1 \Rightarrow \log P(y|x) \leq 0 \Rightarrow -\log P(y|x) \geq 0$

$\therefore H(Y|X) = -\sum_{x,y} P(x,y) \log P(y|x) \geq 0$

$\therefore H(X,Y) - H(X) \geq 0 \Rightarrow \underline{H(X) \leq H(X,Y)}$

5. $H(X,Y) = H(X) + H(Y) - H(X:Y)$

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

$\Rightarrow \quad H(X:Y) \geq 0$

$H(X:Y) = H(Y) - H(Y|X) \geq 0$

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

$H(X:Y) = H\left(P(x,y) \,||\, P(x)\,P(y)\right) = 0$

$\Rightarrow \quad P(x,y) = P(x)\,P(y)$

$\therefore$ Equality iff X & Y are independent random variables.

6. $H(X,Y,Z) + H(Y) \leq H(X,Y) + H(Y,Z)$

Proof

$$H(X,Y,Z) = -\sum_{x,y,z} P(x,y,z) \log P(x,y,z)$$

Chain rule: $P(x,y) = P(y)\,P(x|y)$

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

$\therefore$

$$H(X,Y,Z) = -\sum_{x,y,z} p(x,y,z) \log p(x|y,z)\, p(y,z)$$

$$= -\sum_{x,y,z} p(x,y,z) \Big\{ \log p(x|y,z) + \log p(y,z) \Big\}$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y,z) - \sum_{y,z} \sum_{x} p(x,y,z) \log p(y,z)$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y,z) - \sum_{y,z} p(y,z) \log p(y,z)$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y,z) + H(Y,Z)$$

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

$$= -\sum_{x,y} p(x,y) \log p(y)\, p(x|y)$$

$$= -\sum_{x,y} \Big(\sum_{z} p(x,y,z) \Big) \log p(x|y) - \sum_{y} \Big\{ \sum_{x} p(x,y) \Big\} \log p(y)$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y) - \sum_{y} p(y) \log p(y)$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y) + H(Y)$$

$$\Big\{H(X,Y,Z) - H(Y,Z)\Big\} + \Big\{H(Y) - H(X,Y)\Big\} =$$

$$= -\sum_{x,y,z} p(x,y,z) \log p(x|y,z) + \sum_{x,y,z} p(x,y,z) \log p(x|y)$$

$$= \sum_{x,y,z} p(x,y,z) \log \frac{p(x|y)}{p(x|y,z)}$$

Check. $H(p(x)\|q(x)) \geq 0$ for proof

$$\log(x)\ln(2) = \ln(x) \leq x-1 \quad \forall\ x > 0$$

$$\leq \frac{1}{\ln(2)} \sum_{x,y,z} p(x,y,z) \left\{ \frac{p(x|y)}{p(x|y,z)} - 1 \right\}.$$

$$= \frac{1}{\ln(2)} \sum_{x,y,z} \left\{ p(x|y,z)\, p(y,z)\, \frac{p(x|y)}{p(x|y,z)} - p(x,y,z) \right\}$$

$$= \frac{1}{\ln(2)} \sum_{x,y} p(x|y) \sum_{z} p(y,z) - \frac{1}{\ln(2)} \sum_{x,y,z} p(x,y,z)$$

$$= \frac{1}{\ln(2)} \sum_{x,y} p(x|y)\, p(y) - \frac{1}{\ln(2)} * 1$$

$$= \frac{1}{\ln(2)} \left\{ \sum_{x,y} p(x,y) - 1 \right\} = 0.$$

$$= \frac{1}{\ln(2)} \left\{ \sum_{x,y} p(x,y) - 1 \right\}$$

$\therefore \quad H(X,Y,Z) + H(Y) \leq H(X,Y) + H(Y,Z)$

$\ln(x) = x-1 \quad$ when $x=1 \cdot \left\{ \frac{p(x|y)}{p(x|y,z)} = 1 \right.$

$\therefore$ Equality is achieved iff $p(x|y) = p(x|y,z)$.

i.e., $Z \rightarrow Y \rightarrow X$ is a Markov chain.

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

Proof

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

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

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

$$H(X,Y,Z) - H(Y,Z) \leq H(X,Y) - H(Y)$$

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

Intuitively, we expect that the uncertainty about X, given that we know the value of Y and Z is less than our uncertainty about X, given that we only know Y.

Chaining rule for conditional entropies

Theorem 11.4 : Let $X_1, \ldots, X_n$ and Y be any set of random variables. Then,

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

$$H(X_1, X_2 | Y) = H(X_1 | Y) + H(X_2 | Y, X_1)$$

$$H(X_1, X_2, X_3 | Y) = H(X_1 | Y) + H(X_2 | Y, X_1) + H(X_3 | Y, X_1, X_2)$$

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

Proof

For $n = 2$,

$$H(X_1, X_2 | Y) = H(X_1, X_2, Y) - H(Y)$$
$$= H(X_1, X_2, Y) - H(X_1, Y) + H(X_1, Y) - H(Y)$$
$$= H(X_2 | Y, X_1) + H(X_1 | Y)$$

which is the required result for $n = 2$.

Assume that the result holds for general n,

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

Using the $n=2$ case,

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

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

Applying the inductive hypothesis to the 1st term on the right hand side gives,

$$H(X_2, \ldots, X_{n+1}) = \sum_{i=2}^{n+1} H(X_i|Y, X_1, X_2, \ldots, X_{i-1})$$

$$\therefore H(X_1, \ldots, X_{n+1}|Y) = H(X_1|Y) + \sum_{i=2}^{n+1} H(X_i|Y, X_1, X_2, \ldots, X_{i-1})$$

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

Ex: 11.8 **Mutual information is not always subadditive**

Let X & Y be independent identically distributed random variables taking the values 0 or 1, with probability 1/2. Let $Z = X \oplus Y$, where $\oplus$ denotes addition modulo 2. Show that the mutual information in this case is not subadditive.

$$H(X,Y:Z) \nleq H(X:Z) + H(Y:Z)$$

Ans:

X	Y	Z	P
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

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

$$H(X) = H(Y) = -\sum p(x) \log p(x) = 2 \times -\frac{1}{2} \log \frac{1}{2} = 1 = H(Z)$$

$$H(X,Y,Z) = -\sum_{x,y,z} p(x,y,z) \log p(x,y,z) = 4 \times -\frac{1}{4} \log \frac{1}{4} = 4 \times \frac{1}{2} = 2$$

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) = 4 \times -\frac{1}{4} \log \frac{1}{4} = 2 = H(Y,Z) = H(X,Z)$$

$$H(X:Z) = H(X) + H(Z) - H(X,Z) = 1 + 1 - 2 = 0 = H(Y:Z)$$

$$H(X,Y:Z) = H(X,Y) + H(Z) - H(X,Y,Z) = 2 + 1 - 2 = 1$$

$$\therefore H(X,Y:Z) = 1 \neq 0 = H(X:Z) + H(Y:Z)$$

Ex: 11.9 Mutual information is not always superadditive

Let X_1 be a random variable taking values 0 or 1 with respective probabilities of $1/2$, and $X_2 \equiv Y_1 \equiv Y_2 \equiv X_1$.
Show that the mutual information in this case is not superadditive.

$$H(X_1 : Y_1) + H(X_2 : Y_2) \neq H(X_1, X_2 : Y_1, Y_2)$$

Ans:

X_1	X_2	Y_1	Y_2	P
0	0	0	0	$1/2$
1	1	1	1	$1/2$

$$H(X_1 : Y_1) = H(X_1) + H(Y_1) - H(X_1, Y_1)$$

$$= 1+1 - \left(- \sum_{x,y} p(x,y) \log p(x,y) \right)$$

$$= 1+1 + 2 \times \frac{1}{4} \log \frac{1}{4} = 1+1 - 2 \times \frac{1}{4} \times 2$$

$$= 1+1-1 = 1$$

$$H(X_2 : Y_2) = H(X_2) + H(Y_2) - H(X_2, Y_2)$$

$$= 1$$

$$H(X_1, X_2 : Y_1, Y_2) = H(X_1, X_2) + H(Y_1, Y_2) - H(X_1, X_2, Y_1, Y_2)$$

$$= 1+1-1 = 1$$

$\therefore$

$$H(X_1 : Y_1) + H(X_2 : Y_2) = 1+1 = 2 \neq 1 = H(X_1, X_2 : Y_1, Y_2)$$

The data processing inequality

- In information theory, the data processing inequality states that information about the o/p of a source can only decrease with time: once information has been lost, it is gone forever.

The intuitive notion of information processing is captured in the idea of a Markov chain of random variables.

- A Markov chain is a sequence $X_1 \to X_2 \to \cdots$ of random variables such that X_{n+1} is independent of $X_1, \ldots, X_{n-1}$, given X_n.

$$P\left(X_{n+1} = x_{n+1} \middle| X_n = x_n, \ldots, X_1 = x_1\right) = P\left(X_{n+1} = x_{n+1} \middle| X_n = x_n\right)$$

Random variables X, Y, Z are said to form a Markov chain in the order, denoted by, $X \to Y \to Z$, if the conditional distribution of Z depends only on Y and is conditionally independent of X.

ie.,

X, Y and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as.

$$P(x,y,z) = P(x)\,P(y|x)\,P(z|x,y)$$

$$= P(x)\,P(y|x)\,P(z|y)$$

Under what conditions does a Markov chain lose information about its early values, as time progresses ?

• **Data processing inequality**

Suppose $X \to Y \to Z$ is a Markov chain.
Then,

$$H(X) \geq H(X:Y) \geq H(X:Z)$$

The 1st inequality is saturated ($H(X) = H(X:Y)$) iff, given Y, it is possible to reconstruct X.

⇒ If a random variable X is subject to noise, producing Y, then further actions on our part (data processing) cannot be used to increase the amount of mutual information b/w the output of the process and the original information X.

Proof

Part 1 : $H(X) \geq H(X:Y)$

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

$$H(X|Y) = \sum_y p(y)\, H(X|Y=y)$$

$$= \sum_y p(y) * - \sum_x p(x|y) \log p(x|y)$$

$$= -\sum_{x,y} p(y)p(x|y) \log p(x|y)$$

$$= -\sum_{x,y} p(x,y) \log p(x|y)$$

$0 \leq p(x|y) \leq 1 \Rightarrow \log p(x|y) \leq 0$

$\Rightarrow - \log p(x|y) \geq 0$

$\therefore H(X|Y) \geq 0$

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

$H(X|Y) = H(X) - H(X:Y) \geq 0$

$\Rightarrow H(X) \geq H(X:Y)$

Part 2: $H(X:Y) \geq H(X:Z)$

$$\cancel{H(X)} - H(X|Y) \geq \cancel{H(X)} - H(X|Z)$$

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

- If $X \to Y \to Z$ is a Markov chain then $Z \to Y \to X$ is also a Markov chain.

Proof

$$P(z|x,y) = P(z|y) \implies \frac{P(x,y,z)}{P(x,y)} = \frac{P(y,z)}{P(y)}$$

$$\implies \frac{P(x,y,z)}{P(y,z)} = \frac{P(x,y)}{P(y)}$$

$$\implies P(x|y,z) = P(x|y)$$

$$\therefore H(X|Y) = H(X|Y,Z) = H(X,Y,Z) - H(Y,Z)$$

$$H(X|Z) = H(X,Z) - H(Z)$$

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

$$\therefore H(X|Y) \leq H(X|Z) \iff H(X,Y,Z) - H(Y,Z) \leq H(X,Z) - H(Z)$$

$$\iff H(X,Y,Z) + H(Z) \leq H(X,Z) + H(Y,Z)$$

which is the strong subadditivity inequality.

If $H(X:Y) < H(X)$,

if Z is the attempted reconstruction based only on knowledge of Y, then $X \to Y \to Z$ must be a Markov chain. $\Leftarrow$

Data processing inequality : $H(X) \geq H(X:Y) \geq H(X:Z)$

$\Rightarrow H(X) > H(X:Z)$

$\Rightarrow Z \neq X$

∴ it is not possible to reconstruct X from Y.

If $H(X:Y) = H(X)$,

$H(X:Y) = H(X) - H(X|Y) \Longrightarrow H(X|Y) = 0$

$H(X|Y) = \sum_y P(y)\, H(X|Y=y) = \sum_y P(y) * -\sum_x P(x|y) \log P(x|y)$

$= -\sum_{x,y} P(y) P(x|y) \log P(x|y) = -\sum_{x,y} P(x,y) \log P(x|y)$

$H(X|Y) = 0 \Rightarrow X$ is a deterministic function of Y

ie., knowing Y determines the value of X.

Whenever $P(X=x, Y=y) > 0$ we have

$$P(X=x \mid Y=y) = 1$$

$\Rightarrow$ If $Y=y$ then we can infer with certainty that X was equal to x, allowing us to reconstruct X.

If $X \to Y \to Z$ is a Markov chain, then so is $Z \to Y \to X$.

Data processing inequality: $H(X) \geq H(X:Y) \geq H(X:Z)$

Data pipelining inequality: $\boxed{H(Z:Y) \geq H(Z:X)}$

$\Rightarrow$ any information Z shares with X must be information which Z also shares with Y; the information is 'pipelined' from X thro' Y to Z.

At Classmate, **INDIA'S NO. 1 NOTEBOOK BRAND***, we are committed to providing high quality stationery products that are a result of a deep understanding of our consumers, thoughtful ideation, innovative designs and superior craftsmanship, that have, in turn, helped us become one of India's leading stationery brands!

classmate

Our Classmate range of products include: **NOTEBOOKS,** Writing Instruments - **PENS** (ball, gel and roller), **PENCILS** (mechanical), **MATHEMATICAL DRAWING INSTRUMENTS, ERASERS, SHARPENERS** and **ART STATIONERY** (wax crayons, colour pencils, sketch pens and oil pastels).

*Survey conducted by IMRB in Aug, 2020

NO TREE FROM NATURAL FOREST WAS CUT TO MAKE THIS NOTEBOOK
FOREST SAFE

Forest Safe, is not just a badge or an icon that this exercise book sports. To us it is a label of merit that ensures this Classmate exercise book of ours makes the world and environment a better place. Wood harvested to make paper and board of this Classmate exercise book is not cut from natural forests, but is ethically and responsibly sourced from sustainably managed plantations, showing our commitment towards building an inclusive and secure future for our stakeholders, and the society at large. Join us in our efforts to create a better future, page by page.

WHITE & BRIGHT PAPER | OZONE TREATED | Green Purse

Classmate elemental uses eco-friendly and chlorine free paper

TY

Scan with your smartphone.
Visit us at
www.classmateshop.com

FEEDBACK ? SUGGESTIONS ?

Quality Manager, ITC Ltd.- ESPB,
ITC Centre, 5th Floor,
760, Anna Salai, Chennai. 600 002.
classmate@itc.in 18004253242
(Toll-Free from MTNL/BSNL lines)

A quality product marketed by

ITC
Enduring Value

Exercise Book

172 Pages
(Total Pages Include Index & Printed Information)

Size : 24 x 18 cm

MRP Rs. 48.00
Inclusive of all taxes

Batch: B/AE FT

02000222

8 902519 002222

©ITC Limited

Type of Ruling :

Unruled