International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V70I6P101
Volume 70 Issue 6, 1-12, June 2024

Original Article

Edgeworth Theoretic Characterization of Mutual Information Underneath Joint and Conditional Probability Distribution in Algorithmic Information Theory

Rohit Kumar Verma[1], JharanaChandrakar[2]

Department of Mathematics, BhartiVishwavidyalaya, Durg, C.G. India.

[1]*Corresponding Author : rohitkverma73@rediffmail.com*

Received: 06 April 2024 Revised: 20 May 2024 Accepted: 09 June 2024 Published: 29 June 2024

Abstract *- The science of communication started to attract a lot of attention in the early 1900s. Communication and the transmission of information was rapidly becoming a global phenomenon due to the ongoing growth of telegraph communication, the recent invention of the telephone [3], and, most importantly, the necessity for quick and dependable connections between military groups during the World War. There was an urgent need by the 1940s for a mathematical theory that would direct the development of more advanced communication technologies.*

Keywords *- Entropy, Information, Information improvements, Discrete information, Continuous information***.**

1. Introduction

The seminal work that gave rise to this theory was the groundbreaking publication A Mathematical Theory of Communication by electrical engineer Claude E. Shannon, which was published in the Bell System Technical Journal in 1948.

Shannon describes the "fundamental problem" of the field as frequently being "that of reproducing at one point either exactly or approximately a message selected at another point."and this study provided answers in addition to a thorough philosophy of communication [7].

Among these mechanisms were the communication system schematic that we covered in class and two functions of random variables—mutual information and entropy—that are now essential to any information study.Shannon's theorem, commonly referred to as the channel coding theorem, and the source coding theorem, are two crucial theorems that he proved at the top of his work. However, Shannon's work did not stop at one paper. To quote Robert E. McEliece [4]:
Based only on his 1948 publication, Shannon is recognised by all as the singular father of information theory. Furthermore, he is widely seen as the most significant post-1948 contributor to the field! Since "A mathematical theory of communication," nearly all of his publications have proven to be an invaluable source of research ideas for us mere humans.

Shannon's work was, of course, not without foundations, given how important it is to the communication sector today. Harry Nyquist and Ralph Hartley are acknowledged in his seminal study as early pioneers in the discipline. Specifically, Hartley is renowned for having proposed that the logarithmic function be used as the default function when measuring data. Shannon goes on to say later that Norbert Wiener's work serves as a major inspiration for much of the fundamental philosophy of information theory.

The key contribution Nyquist made to Shannon's research was published in a 1924 paper titled Certain Factors Affecting Telegraph Speed. While the focus of this work was mostly on telegraphs, it did calculate the maximum transmission rate for some noiseless channels [3]. Nyquist is primarily credited with realising that channels have maximum rates of transmission at all. He also proposed some possible causes of noise in telegraph lines that easily transfer to other communication systems [5].

Hartley, whose 1928 work Transmission of Information was an early attempt at developing a generic theory of information, is a more direct addition to the theory of information. Hartley maintained that by making "each [symbol] selection perfectly arbitrary," the psychological component of information—that is, the meaning of the message being conveyed—could be eliminated [2]. Despite many shortcomings, Hartley was able to develop a fully mathematical understanding of information because to this technique. The use of the logarithm to express the quantity of information in a message is one of Hartley's other significant contributions [7]. The logarithm is essential to any discussion of Verma's [11, 12] information formulation, as we shall see.

The 1942 publication of Wiener's note Extrapolation, Interpolation, and Smoothing of Stationary Time Series was the last stage in the process that led to Shannon's work. This book includes an introduction that draws comparisons between the fields of communication and time series, even though the latter is the primary focus of the former. More specifically, Wiener asserts that the definition of communication has far broader implications than was previously thought.

He proposes, for instance, that "the thickness of a roll of paper kept by a condenser working an automatic stop on a Fourdrinier machine" [9] is every bit a communication tool as any human-made message or Morse code, something Hartley utterly overlooked.

Wiener continues by suggesting that the statistical properties of time series apply to communication just as well, since any device intended for communication should be built to function on any possible message transmitted via it rather than just one specific message [9]. Hartley also held this opinion, but it wasn't yet the preeminent one at the time. Furthermore, Wiener made the implication that these probabilities were not always uniform [9], which served as the foundation for Hartley's hypothesis.

While Wiener's letter [9] goes on to describe filters and noise reduction in later sections, the discussion Wiener had above is very philosophically important to information theory in general.
"Much of the basic philosophy and theory of communication theory is owed to Wiener," adds Shannon.

The first accurate treatment of communication theory as a statistical problem can be found in [Time Series]." [7] It is actually rather evident from Shannon's work that communication engineering agrees with Wiener, as he states in the second paragraph, regardless of the material that needs to be sent.

Finally, in 1948, A Mathematical Theory of Communication was published, incorporating many of the most essential ideas of these predecessors and more. There are five sections to the piece. Discrete variables are covered in the first two portions because they are often a simpler topic.

2. Our Result

We shall model this decision in this study. Discrete entropy will be introduced and used to define mutual information first. Next, we shall derive both functions in a continuous context using these definitions as a foundation.

Let X be a discrete random variable taking values in an alphabet A = $\{x_1, x_2 \ldots\}$, called the range of X. Note that the range of X can be finite or countable. We define the probability that X takes the value x_i as

$$P\{X = x_i\} = p_i$$

According to McEliece [4], we will use the notation for any value x in A,

$$p(x) = P\{X = x\}$$

If X and Y are random variables (which may have separate ranges), then for x in the range of X and y in the range of Y, we write

$$p(x, y) = P\{X = x, Y = y\}$$

If the probability of the value of X depends on the value of Y, then X and Y are dependent and we write

$$p(x|y) = P\{X = x|Y = y\}$$

The overall probability of both events, x and y, is represented by p(x, y), therefore we have

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

In particular, if p(y|x) = p(y), then X, Y are said to be independent and

$$p(x, y) = p(x)p(y)$$

Of course, the probability of a value x for X must be the same whether there are other random variables or not. So we have

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

Lastly, we state that the expectation of a function f, denoted f(X), of a random variable is

$$E[f(x)] = \sum_{x} xp(x)$$

Using these notations, we define the entropy of a random variable X as [1]

$H(X) = E[-\log p(X)] = -\sum_x p(x) \log p(x) = \sum_x p(x) \log \frac{1}{p(x)}$

following the rule that 0 log 0 = 0 is an uncertain expression. Additionally, we define X's conditional entropy given Y as

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

With X and Y's combined entropy as

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

In this case, it makes sense to inquire about the logarithm's base.
Since entropy can be defined in terms of any base, this is intentionally left unclear [7]. This is as a result of the base formula changing.

$$\log_b x = \frac{\log_a x}{\log_a b}$$

The base a and base b logarithms therefore only differ by a constant. Then, a change in units can be the cause of this constant [7]. For instance, the bit is the unit of a base 2 entropy logarithm, but the nat is the unit of a base e entropy logarithm, which is frequently employed in continuous information. We've got

$$\log_2 x = \frac{\log_e x}{\log_e 2} \approx 1.44 \log_e x$$

Next, we define a bit as approximately 1.44 nats.

The entropy function has several important properties. First, if p(x) is 1 for some x_i and 0 for all other x_i , then

$$H(X) = -\sum_x p(x) \log p(x) = 1 \log 1 = 0$$

Also, because $p(x) \leq 1$ for any x, we have that $-\log p(x) \geq 0$, and so H(x) is a sum of positive numbers and $H(x) \geq 0$.

Contrarily, if p(x) = 1/n for every x_i, where n is the size of A, then

$$H(X) = -\sum_x \frac{1}{n} \log \frac{1}{n} = \log n$$

Then we have for any probability distribution over X,

$$H(X) = -X\ x\ p(x) \log p(x) \leq \log X\ x\ p(x)\ 1\ p(x) = \log n$$

The inequality here holds from Jensen's inequality, which can be stated as follows: given a convex function f(X) with a finite expectation,

$$E[f(X)] \leq f(E[X]),$$

If, informally, every secant line on the function's graph between two points (x, y) lies above the function's graph on the interval (x, y), then the function is said to be convex. Since − log x is convex, the condition above holds and
E[log p(X)] ≤ log E[p(X)].

Appendix B of McEliece's book The Theory of Information and Coding provides an extensive examination of Jensen's inequality [4]. According to the aforementioned derivation, which is also credited to McEliece [4], H(X) is a maximum if every p() = 1/n, which indicates that all potential values of X are equally likely, and a minimum if some p() = 1, or if the value of X is known. H(x) so characterises the "randomness" of X in this manner, or in other words, the amount of uncertainty we have about the value of X. Where a function is considered convex if, informally, every secant line between two points x, y on the graph of the function lies above the graph of the function on the interval (x, y). Then − log x is convex, so

$$E[\log p(X)] \leq \log E[p(X)],$$

and the inequality above holds. A thorough discussion of Jensen's inequality can be found in Appendix B of McEliece's book The Theory of Information and Coding [4]. The above derivation, also due to McEliece [4], shows that H(X) is a minimum if some $p(x_i) = 1$ - that is, if the value of X is certain - and is a maximum if every $p(x_i) = 1/n$ - if every possible value of X is equally likely. In this way, H(x) describes the "randomness" of X- or in other words, the amount of uncertainty we have about the value of X.
The following Shannon [7] derivation states a second significant property of H(X) about the conditional entropy H(X, Y), which we shall declare as a theorem.

Theorem 1. Any two random variables, X and Y,

$$H(Y) \geq H(Y|X).$$

Proof. (This proof is modeled after McEliece's proof that mutual information is nonnegative [4].) We have

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

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

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

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

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

and by Jensen's inequality,

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

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

$$= \log 1 = 0$$

The aforementioned theorem effectively states that knowing the value of X does not reduce our knowledge of the value of Y. This also makes natural sense. Let X be a symbol that a sender sends via a channel, perhaps with noise, and let Y be the symbol that a receiver on the other end receives. An outside observer's ability to estimate the value of Y would be restricted to crude guesswork. The basic probability of the values in the range of Y would be their only instrument for doing this. However, if the sender were to estimate the value of Y, they may make use of an extra bit of data, namely the value the amount that they sent in X. They could alter the probability distribution of Y by using this value to identify which values of Y are more likely—possibly by identifying which values are, in some way, more similar to that of X. It is reasonable to assume that this modification will only improve the sender's capacity to estimate the value of Y, and Theorem 1 confirms this.

This serves as another justification for the concept of error-correction, since fixing faults in a code simply involves doing the opposite: determining the value of X that is most likely given the value of Y. Since X and Y, a collection of codewords, have the same alphabets in this instance, we may calculate X's distance from Y using the Hamming distance. By doing this, we generate a new probability distribution, where the most likely values of X are those that have a small Hamming distance from Y. By using the reasonable assumption that most channel-induced mistakes are little, we can further enhance this distribution. Assuming a maximum distance of e for an e-error-correcting code, this increases the

possibility of codes with shorter Hamming distances from Y and, in fact, frequently reduces the set of potential codewords to exactly 1. Therefore, Theorem 1 suggests that the error correction procedure will always facilitate message decoding.

Arguably, the most important use of entropy is to find the reciprocal information between two random variables. We examine random variables in the context of Shannon's coding scheme [7] to clarify what is meant by this phrase. In a communication system, random variables can be either the system's input or output. As independent transmitters, the encoder, channel, and decoder accept a value from the alphabet of one random variable and transmit a value from the alphabet of another for every unit of time. Every random variable in the system has a unique probability distribution and entropy, and its component pieces could have states,or positions from which only specific values may be transferred; these positions may vary based on the values received and delivered at a given moment. The value that enters the section and its current status dictate the value that exits it.

This output value often becomes the input 4 value for the subsequent segment. Hence, the random variables that are transmitted between them as information link the many components of a communication system. One significant specific example of this system is the Discrete Memoryless Channel. Since a DMC only has one state, the output values are solely dependent on the input values.The next set of random variables in a chain of DMCs is known as a Markov chain, and its values are solely reliant on the value of the preceding variable [4].

We are inspired to investigate the information transported between random variables in a communication system given this link. Put differently, what is the information sharing capacity between two related random variables, X and Y? As we are well aware, H(Y) represents our level of uncertainty regarding the value of Y. H(Y | X) also indicates the degree of uncertainty that persists regarding Y even after we discover the value of X. Therefore, the amount of knowledge that X has contributed about Y is indicated by the quantity H(Y) − H(Y | X) [4]. Therefore, we define X and Y's mutual information as

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

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

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

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

Now, we extract a number of significant features for I(Y ; X).

First, note that I(Y ; X) is the quantity discussed in the proof of Theorem 1. Thus we have for free that
$I(Y ; X) \geq 0$.
This is to be expected because having insufficient information is illogical. Subsequently [4],

$$I(Y;X) = \sum_{x,y} p(x,y) \log \frac{p(y/x)}{p(y)}$$

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

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

$$= I(X;Y)$$

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

and, using the second line in the above derivation [4],

$$I(Y;X) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

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

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

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

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

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

Mutual information is symmetric, meaning that the quantity of 5 knowledge about the random variable on the other side is the same whether we begin on the transmitting orreceiving end of a channel, as demonstrated by the first derivation above.

The second derivation, which is the Principle of Inclusion and Exclusion, is important from a combinatorial perspective.

First, we tally the entire uncertainty in the system. We then counted the shared uncertainty between X and Y twice due to the overlap. In order to obtain the mutual information between the two random variables, we subtract H(X, Y) from the total.

We will now talk about three or more random variables. As previously mentioned, a communication system's encoder and decoder inputs and outputs, combined with a DMC acting as a channel, create a Markov chain that is represented as W → X → Y → Z. The encoder receives W as its input, the DMC receives X as its input, the decoder receives Y as its input, and the system output is Z. The next theorem is this one, which McEliece [4] presented:

Theorem 2. I(Z; X, Y) ≤I(Z; Y), with equality if and only if p(z|y, x) = p(z|y), where p(z|y, x) = P{Z = z|Y = y, X = x}.

Proof.Before we prove this theorem, we will define the terms used in stating it. First, I(X, Y; Z) refers to the information shared between Z and the pair X, Y. It is defined as [4]:

$$I(Z; X, Y) = E\left[\log \frac{p(z/x,y)}{p(z)}\right]$$

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

Second, we define the probability of Z given X, Y as

p(z|x, y) = P{Z = z|X = x, Y = y}.

This probability acts much in the same way as the conditional probabilities between two variables. That is, we have

$$p(z/x,y) \leq p(z/x)$$
$$p(z/x,y) \leq p(z/y)$$
$$p(x,y)p(z/x,y) = p(x,y,z)$$

Because expectation is linear,

$$I(Z;Y) - I(Z;X;Y) = E[\log \frac{p(z/y)}{p(z)} - \log \frac{p(z/x,y)}{p(z)}]$$
$$= E[\log \frac{p(z/y)}{p(z/x,y)}]$$

By Jensen's Inequality,

$$I(Z;Y) - I(Z;X,Y) \geq \log E[\frac{p(z/y)}{p(z/x,y)}]$$
$$= \log \sum_{x,y,z} p(x,y,z) \frac{p(z/y)}{p(z/x,y)}$$
$$= \log \sum_{x,y,z} p(z/y)p(x,y)$$
$$= \log \sum_{y,z} p(z/y) \sum_{x} p(x,y)$$
$$= \log \sum_{y,z} p(z/y)p(y)$$
$$= \log \sum_{y,z} p(y,z)$$
$$= \log 1 = 0$$

If p(z/x,y)=p(z/y), then

$$I(Z;Y) - I(Z;X,Y) = E[\log\frac{p(z/y)}{p(z)} - \log\frac{p(z/x,y)}{p(z)}]$$

$$= E[\log\frac{p(z/y)}{p(z)} - \log\frac{p(z/y)}{p(z)}] = 0,$$

and if $p(z/x,y) \neq p(z/y)$, then $p(z/x,y) > p(z/y)$ and we have

$$I(Z;Y) - I(Z;X,Y) = E[\log\frac{p(z/y)}{p(z)} - \log\frac{p(z/x,y)}{p(z)}]$$

$$> E[\log\frac{p(z/y)}{p(z)} - \log\frac{p(z/y)}{p(z)}] = 0,$$

and so

$$I(Z;Y) - I(Z;X,Y) \neq 0$$

This theorem leads to a very interesting result for any communication system which can be modeled by a Markov chain. By symmetry of the above theorem, we have

$$I(Z; X, Y) \leq I(Z; X),$$

and $I(Z; X, Y) = I(Z; Y)$ since $X \rightarrow Y \rightarrow Z$ is a Markov chain.
Thus,

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

This reasoning can be used again to determine that, in a Markov chain [4],

$$I(X;Z) \geq \begin{cases} I(X;Y) \\ I(Y;Z) \end{cases}$$

We now apply the aforementioned theorem to the Markov chain, $W \rightarrow X \rightarrow Y \rightarrow Z$, that was previously discussed. We examine the $W \rightarrow X \rightarrow Z$ sub-Markov chain. As $I(Z; W, X) = I(Z; X)$, we are left with

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

To put it succinctly, less information is shared between the sender and received messages during a conversation than there is during the channel itself. This statement, which is also known as the data-processing theorem [4], asserts that information can only be preserved or destroyed during the data processing phases of encoding and decoding. Information cannot be created during these steps. It is logical to anticipate this. Since an encoder's only job is to convert a message into a format that can be sent across a channel, it is unlikely that it will add additional information to a message. However, because it lacks access to X, the decoder would find it difficult to add any information from X that did not pass through the channel.It is logical to anticipate this. Since an encoder's only job is to convert a message into a format that can be sent across a channel, it is unlikely that it will add additional information to a message. However, since it does not have access to the information, the decoder would find it difficult to add any information from X that did not pass over the channel. Regarding the channel, its primary concern is obliterating information, if any at all. The message can only be returned to its original state by error repair, which for the sake of this Markov chain discussion happens as a reduction of the information destruction inherent in the channel.

Not even it can generate new knowledge. Thus, logically, the data-processing theorem is also valid. In information theory, the previously discussed discrete random variables are frequently the main emphasis. There are several reasons for this.

First, as we have seen in class, finite fields and discrete alphabets are crucial to many of the applications of information theory, including computer science and error correction. But these are by no means the only situations in which information theory can be used. As was previously indicated, Wiener [9] made evident how frequently information theory is used. That some of these locations rely on continuous probability distributions over the real numbers is not surprising. Speaking in human, for example, starts as an idea, which is the source.It is delivered through the air channel as a frequency that can take on any value within a continuous range of conceivable frequencies after passing through a 7 encoder in the form of vocal chords. Then, a human's ear "decodes" it into a series of sounds that the recipient can understand. All other noises that are not produced digitally travel along a very similar route. Even sounds that are produced digitally frequently require continuous information since noise can distort and destroy the discrete quality of delivered frequencies. Because the human ear is incapable of quantizing sounds again, they remain continuous while being decoded. Transmissions on television and radio are other instances of applicability.

As with discrete information, the first formulation of continuous information was given in Shannon's original paper [7]:

The entropy of a discrete set of probabilities $(p_1 p_2 \ldots\ldots p_n)$ is defined as follows.

:

$$H = -\sum p_i \log p_i.$$

In an analogous manner we define the entropy of a continuous distribution with the density distribution function p(x) by:

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x) dx.$$

But in this particular instance, Shannon's work falls short. Shannon only defined this continuous entropy function analogously and expected it to be valid, without actually deriving it. But entropy cannot be described by this function! Take the probability distribution U(x) with [1] as an example.

$$U(x) = \begin{cases} 2 & x \in [0,1/2] \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$H(x) = -\int_{-\infty}^{\infty} U(x) \log U(x)\, dx$$

$$= -\int_{0}^{1/2} 2 \log 2\, dx$$

$$= -\log 2$$

Nonetheless, this defies the condition that $H(X) \geq 0$.

Further, it can be shown that H(x) is coordinate variant - that is, it changes depending on the coordinate system chosen [7]. Clearly this function does not satisfy the necessary properties of entropy.

Differential entropy (h(X), which is the modern term for Shannon's notion of continuous entropy, has several applications, as we shall see later. It does not, however, in any way convey the uncertainty or randomness of a continuous random variable. We already know that discrete entropy describes these aspects of a signal, therefore we need to create a continuous entropy definition that does the same. Since measuring and integrating over probability distributions is a key component of continuous probability theory, let us first give a brief overview of measure theory. An Introduction to Measure Theory by Terence Tao contains a more in-depth treatment [8].

Assume that $\{ B_n$, n = 1, 2, · · · } is a countable collection of intervals over R that have the form[a, b]. Then for any set E of real numbers, we define the Lebesgue measure, or Lebesgue outer measure, of E as

$$m^*(E) = \min_{\cup_{n=1}^{\infty} B_n \supseteq E} \sum_n |B_n|$$

Next, we declare that E is Lebesgue measurable if there is an open set U such that for any $\epsilon > 0$.

$$m^*(U/E) \leq \epsilon$$

Assume that F is a σ-algebra on **Ω**, meaning that it is a collection of subsets that contain ᴓ and are closed under countable crossings, countable unions, and complement.

The space (**Ω**, F, μ) is referred to as a measure space if μ is a measure on **Ω**. In our case, these relatively abstract notions resolve to more concrete ideas. First, **Ω** = R since we are only working with real numbers. Furthermore, we define F as the collection of all real number Lebesgue-measurable subsets, which is proved to form a σ-algebra. Lastly, a probability measure P will take the place of our abstract measure μ. P(**Ω**) = 1 is the abstract definition of a probability measure, which is a measure over **Ω**. Usually, though, this just indicates that P is the real numbers' probability distribution. It is usual practice to take into account numerous measure spaces and probability measures in this manner, particularly when working with several random variables.

Now let Q = $\{S_i$, i = 1, 2, · · · } be an at most countable partition of R into Lebesgue-measurable sets. Then we can define the quantization of X by Q as a discrete random variable $[X]_P$ taking values in the set {1, 2, · · · }, with the probability distribution

p(i) = P{ $[X]_Q$ = i}
= P{X ∈ S_i }
= ∫ dP(x)alongS_i

If the measure P(x) is Lebesgue measurable, then the above integral can be evaluated as follows [8]:

p(i) = ∫ f(x) d$m^*(x)$ along S_i = ∫ f(x) dxalong S_i,

Where dm^* denotes that the integral is taken with respect to the Lebesgue measure of x, $m^*(x)$. In this way, we reduce X to a discrete random variable $[X]_Q$ taking values in the set {1, 2, · · ·}, to which our discussion of discrete information applies directly.

Of course,$[X]_Q$ provides little more than an approximation of the actual probabilities of X by way of grouping sets of real numbers together and considering the total of their probabilities. However, this approximation can be improved arbitrarily [2]. Given two partitions Q_1 and Q_2 of R, we say thatQ_1 is a refinement ofQ_2 if every interval S_j inQ_1 is contained in some interval$S_i$in Q_1. Then, because the sets in Q_1 are subsets of those inQ_2, the probabilities p(j) forQ_1 give a better approximation of the actual probabilities of each real number than the probabilities p(i) forQ_1. Now, given two continuous random variables X and Y , we define the mutual information of X and Y as

$$I(X; Y) = \binom{sup}{Q,R} I([X]_Q;[Y]_R)$$

Where Q and R are R partitions, and all such partitions are taken over by the supremum.

Note that, unlike entropy, the mutual information of two continuous random variables is straightforward to define. In fact, it also obeys all the properties one might expect from mutual information. For instance, because I(X; Y) is a supremum of nonnegative values, I(X; Y) itself must also be non-negative. Also, consider random variables X, Y taking values in discrete subsets of R, denoted $\{x_1, x_2, \cdots\}$ and $\{y_1, y_2, \cdots\}$ Then there exist partitions Q, R of R such that every x_i is contained S_i in , no S_i contains more than one x_i , and the same is true for S_j and y_j.Then we have

p(i) = P{X ∈ S_i} = P{X =x_i }
q(j) = P{Y ∈ S_j } = P{Y=y_j }.

Because there is only one nonzero point in each S_i ,S_j, no refinement of Q or R can increase the mutual information I($[X]_Q$;$[Y]_R$)
Thus,

$$\binom{sup}{Q,R} I([X]_Q;[Y]_R) = I(X; Y)$$

and our definition holds for a discrete variable.

Finally, if Q_1 is a refinement of Q_2, a partition of the real numbers for X, then the value of $[X]_{Q_2}$depends on the value of $[X]_{Q_1}$. Consequently, the data-processing theorem gives us

$$I([X]_{Q_2}; Y) \geq I([X]_{Q_1} ; Y)$$

This allows us to think of the supremum in our mutual information definition as a kind of limit as Q and R are continuously improved upon [2]. Though our definition makes sense intellectually, it is hard to assess in practice. There is no universal form for an arbitrary probability distribution [4]. On the other hand, for random variables with continuous, nonegative probability distributions, a definition that is simpler to compute exists. The following definition can be found in McEliece's [4] theorem:

Theorem 3. Let X and Y be continuous random variables each taking values in R with probability distributions$P_1(x)$ and $P_2(y)$ respectively. Then let P(x,y) be the probability distribution over R^2 which is the joint probability of X and Y - that is,

P{X ∈ A, Y ∈ B} = ∬ $dP(x, y)$ along A and B
= ∬ $P(x, y)dxdy$ along A and B

Proof. Assume that the distribution P(x,y) is continuous and nonnegative. Next, let's

$$h(X) = -\int_{-\infty}^{\infty} \log p(x) dP_1(x)$$

be the differential entropy of X as defined by Shannon, and let the conditional differential entropy of X given Y be denoted, according to Shannon's definition [7], as

$$h(X/Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log P(x/y) dP(x,y)$$

with

$$P(x/y) = \frac{P(x,y)}{P_2(y)}$$

Then

$$I(X;Y) = h(X) - h(X/Y).$$

We choose $\in_1, \in_2$ arbitrarily, with $0 < \in_1 < \in_2$. Then let Q = $\{Q_i, i = 1,2,\ldots\ldots..\}$ be a partition of R such that the interval Q_i has supremum x_i and infimum x_{i-1} with

$$\in_1 < \Delta x_i < \in_2$$

With $\Delta x_i = x_i - x_{i-1}$ Similarly, we let R= $\{R_j j = 1,2,\ldots\ldots..\}$ be a partition of R such that R_j has supremum y_j with

$$\in_1 < \Delta y_j < \in_2$$

Then, by the mean value theorem, there exists some q_i in Q_i and some r_j in R_j such that

$$p(i) = P\{[X]_{Q=i}\} = \int_{x_{i-i}}^{x_i} dP_1(x) = \Delta x_i P_1(q_i)$$

$$q(j) = P\{[Y]_{R=j}\} = \int_{y_{j-i}}^{y_j} dP_2(y) = \Delta y_j P_2(r_j)$$

and with the help of a two dimensional mean value theorem, we have that there exists a 2- tuple (q_{ij}, r_{ij}) in R^2 such that

$$p(i,j) = P\{[X]_Q = i, [Y]_R = j\}$$

$= \int_{y_{j-i}}^{y_j}\int_{x_{i-i}}^{x_i} P(x/y) P_2(y) dy dx$

$= P(q_{ij}/r_{ij}) \int_{y_{j-i}}^{y_j}\int_{x_{i-i}}^{x_i} P_2(y) dy dx$

$= P(q_{ij,}/r_{ij}) \quad \Delta x_i \int_{y_{j-i}}^{y_j} P_2(y) dy$

$= P(q_{ij,}/r_{ij}) \quad \Delta x_i \, q(j)$

$\Rightarrow p(i/j) = \frac{p(i,j)}{q(j)} = \Delta x_i P(q_{ij,}/r_{ij})$

Now we have

$I([X]_Q ; [Y]_R) = H([X]_Q) - H([X]_Q/[Y]_R)$

$$= -\sum_i p(i) \log p(i) + \sum_{i,j} p(i,j) \log P(i/j)$$

Which we can write

$$-\sum_i \Delta x_i P_1(q_i) \log \Delta x_i P_1(q_i) + \sum_{i,j} \Delta x_i \Delta y_j P(q_{ij,}/r_{ij}) P(r_j) \log \Delta x_i P(q_{ij,}/r_{ij})$$

$$= \sum_i \Delta x_i P_1(q_i) \log \frac{1}{\Delta x_i} + \sum_i \Delta x_i P_1(q_i) \log \frac{1}{P_1(q_i)} - \sum_{i,j} \Delta x_i \Delta y_j P\left(\frac{q_{ij,}}{r_{ij}}\right) P(r_j) \log \frac{1}{\Delta x_i} - \sum_{i,j} \Delta x_i \Delta y_j P(q_{ij,}/r_{ij}) P(r_j) \log \frac{1}{P(q_{ij,}/r_{ij})}$$

Note that the first and second sums were derived from $H([X]_Q)$ while the third and fourth were derived from $H([X]_Q/[Y]_R)$. The third sum reduces as follows:

$$\sum_{\mathrm{i,j}} \Delta x_i \, \Delta y_j \, \mathrm{P}(q_{ij,}/r_{ij}) \, \mathrm{P}(r_j) \log \frac{1}{\Delta x_i} = \sum_i \sum_j p(i,j) \log \frac{1}{\Delta x_i}$$

$$= \sum_{\mathrm{i}} \mathrm{P(i)} \log \frac{1}{\Delta x_i}$$

$$= \sum_{\mathrm{i}} \Delta x_i \; \mathrm{P}_1 (q_i) \; \log \frac{1}{\Delta x_i}$$

and so

$\sum_{\mathrm{i}} \Delta x_i \, \mathrm{P}_1 (q_i) \log \frac{1}{\Delta x_i} - \sum_{\mathrm{i,j}} \Delta x_i \, \Delta y_j \, \mathrm{P}(q_{ij,}/r_{ij}) \, \mathrm{P}(r_j) \log \frac{1}{\Delta x_i} = 0$

and

$\mathrm{I}([\mathrm{X}]_{\mathrm{Q}} ; [\mathrm{Y}]_{\mathrm{R}} = \sum_{\mathrm{i}} \Delta x_i \, \mathrm{P}_1 (q_i) \log \frac{1}{\mathrm{P}_1 (q_i)} - \sum_{\mathrm{i,j}} \Delta x_i \, \Delta y_j \, \mathrm{P}(q_{ij,}/r_{ij}) \, \mathrm{P}(r_j) \log \frac{1}{\mathrm{P}(q_{ij,}/r_{ij})}$

Now, we recall that

$$\mathrm{I(X; Y)} = \binom{sup}{Q,R} \mathrm{I}([X]_Q ; [Y]_R)$$

As stated previously, the supremum in this definition is a limit of sorts as P and Q are refined. This, of course, corresponds to a limit as $\Delta x_i \, \Delta y_j$ approach 0. (It would also correspond to an increase in the amount of values which i an j can take, if they were not already countable.) Thus, we can evaluate the continuous case as

$$\mathrm{I(X; Y)} = \lim_{(\Delta x_i \, \Delta y_j) \to (0,0)} \mathrm{I}([X]_Q ; [Y]_R)$$

Since the first sum is simply an approximation of h(X), so as ε_2 approaches 0, this sum converges to h(X). Similarly, the second sum is a step approximation of h(X|Y), and so it converges to h(X|Y).

In this case, we can observe that differential entropy is useful for determining the mutual information between two random variables. As one reads through A Mathematical Theory of Communication, Theorem 3 actually suggests a number of significant points. But first, we look at the two sums that cancelled out in the Theorem 3 proof before talking about this. For a random variable X with a continuous, nonnegative probability distribution, $P_1 (x)$, we have, if we just take into account the first pair of sums,

$$\mathrm{H}([X]_Q) = \sum_{\mathrm{i}} \Delta x_i \, \mathrm{P}_1 (q_i) \log \frac{1}{\Delta x_i} + \sum_{\mathrm{i}} \Delta x_i \, \mathrm{P}_1 (q_i) \log \frac{1}{\mathrm{P}_1 (q_i)}$$

As we've seen, when the discrete variable becomes closer to the continuous variable, the supremum of all partitions Q of the range of X functions as a kind of limit. Thus, it is reasonable to assume that these sums will converge to X's entropy, and the second sum does in fact converge to h(X). Nevertheless, we discover that the initial sum actually does not converge as Δx_i gets closer to 0. Rather, we discover that

$$\lim_{\Delta x_i \to 0} \sum_{\mathrm{i}} \Delta x_i \, \mathrm{P}_1 (q_i) \log \frac{1}{\Delta x_i} = \lim_{\Delta x_i \to 0} \log \frac{1}{\Delta x_i} \sum_{\mathrm{i}} \mathrm{p(i)}$$

$= \lim_{\Delta x_i \to 0} \log \frac{1}{\Delta x_i}$.

Thus, as Δx_i approaches 0, this term approaches ∞, and we have

$\lim_{\Delta x_i \to 0}$ H(X) = h(X) + ∞ = ∞,

H(X) always diverges, in other words. Actually, it has been demonstrated by M. S. Pinsker [6] that a probability distribution whose entropy does prove to be finite is, in reality, discrete itself since it is concentrated at discrete points. (We will not provide his derivation here as it entails considerably more measure theory than we are ready to employ.) It is obvious that continuous entropy is a far more challenging topic than discrete entropy. Even though Edwin Thompson Jaymes' limiting density of discretepoints was one attempt to redefine continuous entropy as a more useful function, and even though that one in particular has achieved some success by establishing a function known as relative entropy [4], the entropy itself still turns out to be finite.

3. Conclusion

Shannon's assumption that differential entropy is the accurate continuous equivalent of discrete entropy is fortunate, as otherwise, the field of continuous entropy may potentially be extremely static. It has several uses even though it does not describe uncertainty in the way we would prefer. In fact, some of these uses are defined in A Mathematical Theory of

Communication, indicating that Shannon's discussion of continuous entropy was not entirely lost. Shannon actually specifies [7]

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

Where H(X) is the actual continuous entropy and H(X|Y) is the actual conditional entropy, for a continuous channel between random variables X and Y. But naturally, as we observed in Theorem 3

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

Hence the rate is accurately defined by the functions Shannon employed. He then defined the channel capacity as the highest transmission rate that can be achieved over the channel, in line with his discrete definitions of capacity. Naturally, this merely entails determining R's maximum [7], and since this is the maximum of a function that is known to be valid, the definition must be accurate. A straightforward switch in notation from entropy to differential entropy leads to the next conclusion, conclusion 16, which validates the entirety of Part IV of A Mathematical Theory of Communication. This presents a thought-provoking viewpoint of the distinctions between differential and continuous entropy.Although having a function that satisfies every discrete entropy property while remaining finite would be ideal in theory, in actuality it doesn't really matter if it doesn't. Entropy is really just a function, which explains why. In contrast to mutual information, it does not address a central query in the subject or clarify a necessary component of information theory. The fact that the differential entropy does not convey our level of uncertainty about a variable, while the entropy of a random variable does, does not appear to matter in 12 fact. Even while these characteristics were initially employed to specify the function [7], in actuality they are not so important that they can be ignored. Thus, despite its theoretical shortcomings, differential entropy is a perfectly valid operational substitute for continuous entropy.

References

[1] Differentialentropy. [Online]. Available: https://en.wikipedia.org/wiki/Differentialentropy

[2] R.V.L. Hartley, “Transmission of Information,” *Bell System Technical Journal*, vol. 7, no. 3, pp. 535-563, 1928. [CrossRef] [Google Scholar] [Publisher Link]

[3] George Markowsky, Information Theory, 2024. [Online]. Available: https://www.britannica.com/science/information-theory

[4] Robert J. McEliece, *The Theory of Information and Coding*, Cambridge: New York, 2004. [Google Scholar] [Publisher Link]

[5] H. Nyquist, “Certain Factors Affecting Telegraph Speed,” *Journal of the A.I.E.E.*, vol. 43, no. 2, pp. 124-130, 1924. [CrossRef] [Google Scholar] [Publisher Link]

[6] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes*, *San Francisco: Holden-Day*, pp. 1-243, 1964. [Google Scholar] [Publisher Link]

[7] C.E. Shannon, and W. Weaver, “The Mathematical Theory of Communication,” *Urbana: University of Illinois Press*, 1972.

[8] Terence Tao, *An Introduction to Measure Theory*, Providence, RI: American Mathematical Society, 2013. [Google Scholar] [Publisher Link]

[9] Rocco Tenaglia, Entropy and Information in Discrete and Continuous Information Theory, 2017. [Google Scholar]

[10] Norbert Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*, The MIT Press, 1949. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rohit Kumar Verma, and BabitaVerma, *A New Approach in Mathematical Theory of Communication (A New Entropy with its Application)*, Lambert Academic Publishing, 2013. [Google Scholar]

[12] R.K. Verma, *Family of Measures of Information with their Applications in Coding Theory and Channel Capacity*, Lambert Academic Publishing, 2023.