Rota's Entropy Theorem

A Mathematically Precise And Universal Definition of Entropy

Essam Abadir

May 7, 2025

In memory of Gian-Carlo Rota, April 27, 1932 - April 18, 1999.

Introduction

"No one really knows what entropy is," according to John von Neumann. Perhaps we have not known till now, or rather, we missed the definition entirely circa 1978 when it was being handed out in a class at MIT. For decades, legendary MIT Professor Gian-Carlo Rota taught a class numbered 18.313 "Introduction to Probability" which largely covered the contents of a 400+ page manuscript laying out entropy's precise mathematical definition. It is a mystery as to how it could have been taught to generations of MIT students while still being basically unknown to the science community at large. Rota's Entropy Theorem (RET) provides the first and, to my knowledge, only formalizabley precise definition of entropy. The purpose of this paper is to rectify this historical blunder. Unifying entropy into a computable mathematical structure has potentially profound implications.

The deep link and mystery of entropy is that Von Neumann's version of it underlies the entire field of quantum information theory, while Claude Shannon's version of it underlies the entire field of information theory, and, arguably all of computer science by extension. This fact was not lost on either Von Neumann or Shannon, as the following quote from Shannon illustrates:

> My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.'
>
> —*Claude Shannon, 1956*

However, as the quote also illustrates, the link between Shannon's and Von Neumann's versions of entropy is not well understood. Rota's Entropy Theorem (RET) provides a formalization of this link and shows that it is not just a coincidence, but rather a deep and fundamental **equivalence**.

From physics to computer science, entropy is known to be at the core of almost everything in the natural sciences. Unfortunately, precisely defining what entropy actually is has been an elusive task and without that definition, it has been impossible to unify the many and varied descriptions of entropy.

That the proof of Rota's Entropy Theorem was available a half-century ago begs the question of what advances it might have enabled had it been widely known. Perhaps the next near future will bear out my suspicion that RET is quite possibly the most important result in the sciences of the last 50 years. It is an honor to share it here.

Theorem (Informal) All fundamental "continuous" physics distributions – notably the Maxwell–Boltzmann (MB), Fermi–Dirac (FD), and Bose–Einstein (BE) statistics – can be expressed as scaled versions of the discrete Shannon entropy functional.

In practical terms, RET asserts that for each of these statistical distributions, there exists some discrete probability distribution (a partition of a finite sample space with probabilities) such that the **Shannon entropy** of that partition, multiplied by an appropriate constant, reproduces the given physics distribution.

Symbolically, the theorem states: For every physical distribution D, there exists a set of probabilities $\{p_i\}$ and a constant C such that:

$$H(\{p_i\}) \cdot C = D$$

where $H(\{p_i\})$ is the Shannon entropy of the discrete probability distribution.

RET: Identification & Uniqueness of Entropy

RET is a two part theorem: 1. Properties Which Identify Probability Distributions As Entropy Functions; And 2. a proof that all entropy functions as defined in (1) are mathematically equivalent to Shannon Entropy by some constant scale factor.

The first part is the identification of entropy, which states a set of common properties that can be tested for on a probability distribution in order to give it a "label" of being an entropy function. The second part is the uniqueness of entropy, which proves that all.

The Entropy Indenfication Test

The following five properties, known as Rota's properties of entropy, define the mathematical criteria that entropy must satisfy. These properties form the foundation of the uniqueness of entropy.

1. **Definition on Probability Sets:** An entropy function H is defined on sets $\{p_1, p_2, \ldots, p_n\}$ of non-negative real numbers, which satisfy:

$$p_1 + p_2 + \ldots + p_n = 1.$$

2. **Dependence on Nonzero Probabilities:** If H is an entropy function, then for any set $\{p_1, p_2, \ldots, p_n, 0\}$, H satisfies:

$$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n).$$

   In other words, H depends only on the nonzero probabilities in a given set.

3. **Continuity:** The entropy function H is continuous with respect to the probabilities $\{p_1, p_2, \ldots, p_n\}$. Continuity is actually a byproduct of the other properties and the Law of Large Numbers, but it is useful to state it explicitly.

4. **Additivity via Conditional Entropy:** If π is a finer partition than σ, then the conditional entropy satisfies:

$$H(\pi|\sigma) = H(\pi) - H(\sigma).$$

5. **Maximum Entropy for Uniform Distribution:** Among all partitions with a given number of blocks, the partition with maximum entropy is the one where all blocks have equal probability:

$$H(p_1, p_2, \ldots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

These properties uniquely characterize the entropy function and ensure its consistency across various applications in probability and information theory.

Uniqueness: There is only one Entropy Function

As Rota himself warns in the text below regarding the proof of uniqueness "The proof is rather technical, so we suggest omitting it on the first reading." It will be easy for anyone to get lost in the mathematical details of the uniqueness proof - to do so is to completely miss the forest for the trees. The stunning implications of RET are that it provides a formal definition of entropy and *that this definition is the same not only for information theory's Shannon Entropy for also for "physics entropy" displayed by all physical systems*.

Achieving The Generalized Form: Key Techniques Used in the Proof

The reader will note three key techniques that are subtly used in the main proof discussing the Uniqueness of Entropy that allow the generalization of entropy to take place. five properties of entropy to be the universal test which defines whether a probability distribution is the result of an entropy function.

1. Rota uses a recursive argument about uniform distributions to show that any probability distribution can be expressed as a set of partitions each with a uniform distribution. This allows the extension of the proof to all distributions.

2. Filtering out non-zero probabilities allows any distribution to be transformed into a continuous distribution.

3. Conditional entropy and additivity of entropy allow for the mixing of different entropy distributions but still result in a single entropy distribution.

Formal Derivation of Rota's Entropy Theorem

The remainder of this section is excerpted from class text provided by Professor Gian-Carlo Rota. To my knowledge it is unpublished and uncopyrighted. "Introduction to Probability Theory, Second Preliminary Edition" manuscript circa 1993, authors are Kenneth Baclawski, Gian-Carlo Rota, & Sara Billis. It is similar to the one on the Internet Archive [3] where the same proof is present, but I have not found this particular version online.

Chapter VIII: Entropy and Information

Properties of Entropy

So far, we have discussed examples of the entropy of some random variables. Although these examples provide some motivation for our definition of entropy, they leave unanswered the more difficult question of why, out of all possible definitions, we use this one.

We will do this by finding five self-evident properties that ought to hold for any reasonable measure of information (or entropy). It then turns out that our definition of entropy is the only one that satisfies all these properties.

We begin with the most obvious of properties. As we have defined it, H is a function of partitions of the sample space. However, it should be clear that we want H to depend only on the set of probabilities of the blocks of the partition. In fact, we want H to depend only on the positive probabilities which occur. Moreover, we want H to be a continuous function of these probabilities. This is a convenience only. We could, with a great deal of effort, derive continuity from other more complex conditions; but we would rather concentrate on the important issues.

We summarize the conditions on H we have just described before going on to the difficult question of conditional entropy.

Entropy Property 1: An entropy is a function defined on sets $\{p_1, p_2, \ldots, p_n\}$ of non-negative real numbers, which satisfy $p_1 + p_2 + \ldots + p_n = 1$.

Entropy Property 2: If H is an entropy function, then for any set $\{p_1, p_2, \ldots, p_n, 0\}$ on which H is defined, H satisfies:

$$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n).$$

In other words, H depends only on the nonzero p_i's in a given set.

Entropy Property 3: An entropy function is continuous. The next property of entropy we consider requires the concept of conditional entropy. There are two ways to think of conditional entropy, and the fact that they are equivalent is our next property of entropy. To illustrate the ideas involved, we consider the following simple weighing problem:

We have three coins, some of which may be counterfeit (but not all). Counterfeit coins are distinguishable from normal coins by the fact that they are lighter. We are given a balance scale, and we wish to find out which, if any, of the coins are counterfeit. The sample space for this problem consists of seven sample points, one for each possible set of good coins. We denote them as follows:

$$\Omega = \{1, 2, 3, 12, 13, 23, 123\}.$$

Now what happens when we put the first two coins on each side of the scale? The sample space is partitioned into three blocks corresponding to the three possible outcomes of the weighing:

$$\sigma = \{\{12, 123, 3\}, \{2, 23\}, \{1, 13\}\}.$$

After recording the result of this weighing, we then place the second and third coins on the two sides of the scale. The result of this second weighing is to partition each of the blocks of the first weighing:

$$\{12, 123, 3\} \to \{\{12\}, \{123\}, \{3\}\}, \quad \{2, 23\} \to \{\{2\}, \{23\}\}, \quad \{1, 13\} \to \{\{1\}, \{13\}\}.$$

The combined information of the two weighings is represented by the partition into seven blocks, each with one sample point. Call this partition π. Conditional entropy is concerned with the effect of the second weighing, given that the first has occurred. One way to analyze this is to look at each block σ_i of the partition of the first weighing and to analyze the situation as if it were the whole sample space. In general, for an event A and a partition τ, we define the conditional entropy of π given A, written $H(\pi|A)$, to be the entropy of the partition $\tau_1 \cap A, \tau_2 \cap A, \ldots$ that τ induces on A.

Thus, in the above weighing problem, we have three conditional entropies, one for each possible outcome of the first weighing:

$$H(\pi|\sigma_1), \quad H(\pi|\sigma_2), \quad H(\pi|\sigma_3).$$

The conditional entropy of π given σ is then defined to be the average of these. More precisely, if π and σ are any two partitions of a sample space Ω such that π is finer than σ, we define the conditional entropy of π given σ to be the average value of $H(\pi|\sigma_i)$ over all blocks σ_i of σ:

$$H(\pi|\sigma) = \sum_i P(\sigma_i) H(\pi|\sigma_i).$$

On the other hand, we would like to think of information as a "quantity" that increases as we ask more and more questions about our experiment. Therefore, the conditional entropy of π given σ ought to be the net increase in entropy from σ to π. In other words, we require our entropy function to satisfy:

Entropy Property 4: If π is a finer partition than σ, then

$$H(\pi|\sigma) = H(\pi) - H(\sigma).$$

The last property we require is one that we have already discussed. The partition having maximum entropy among all partitions with a given number of blocks is the one for which all the blocks have the same probability.

Entropy Property 5: If H is an entropy function, then any set $\{p_1, p_2, \ldots, p_n\}$ on which H is defined satisfies:

$$H(p_1, p_2, \ldots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

We are now ready for the following remarkable fact: if H satisfies the above five properties, then H is given by the formula introduced earlier in this chapter, except for a possible scale change.

Uniqueness of Entropy

If H is a function satisfying the five properties of an entropy function, then there is a constant C such that H is given by:

$$H(p_1, p_2, \ldots, p_n) = C \sum_i p_i \log_2 \frac{1}{p_i}.$$

Proof: The proof is rather technical, so we suggest omitting it on the first reading. However, it is of interest to outline the main points. To show that H has the form given above, we use the following two facts:

1. The entropy of the partition consisting of just one block of probability 1 is zero, i.e., $H(\Omega) = 0$. By definition, $H(\Omega)$ is the same as $H(\{1\})$. Therefore, $H(\Omega) = H(\{1\}) = 0$.

2. We define a function $f(n)$ by $H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$. We have just shown that $f(1) = 0$ and we want to calculate $f(n)$ in general. Using properties 2 and 5, we show that $f(n)$ is increasing:

$$f(n) \leq f(n+1).$$

Next, we consider a partition σ consisting of n^k blocks, each of which has probability $\frac{1}{n^k}$. Then subdivide each of these into n parts, each of which has the same probability. Call the resulting partition π. The

conditional entropy $H(\pi|\sigma)$ for each block σ_i is clearly given by $f(n)$. Thus the conditional entropy $H(\pi|\sigma)$ is $f(k) - f(k-1)$. If we apply this fact k times, we obtain:

$$f(n^k) = kf(n).$$

Now fix two positive integers n and k. Since the exponential function is an increasing function, there is an integer b such that:

$$2b \leq n^k < 2b + 1.$$

We now apply the two facts about $f(n)$ obtained above to this relation:

$$f(2^b) \leq f(n^k) \leq f(2^{b+1}).$$

Since $f(n)$ is increasing, we know:

$$bf(2) \leq kf(n) \leq (b+1)f(2).$$

Now divide these inequalities by $kf(2)$:

$$\frac{b}{k} \leq \frac{f(n)}{f(2)} \leq \frac{b+1}{k}.$$

Now apply the increasing function $\log_2$ to the inequalities:

$$\frac{b}{k} \leq \log_2(n) \leq \frac{b+1}{k}.$$

It follows that both $f(n)/f(2)$ and $\log_2(n)$ are in the interval $[b/k, (b+1)/k]$. This implies that $f(n)/f(2)$ and $\log_2(n)$ can be no farther apart than $1/k$, the length of this interval. But n and k were arbitrary positive integers. So if we let k get very large, we are forced to conclude that:

$$f(n)/f(2) = \log_2(n).$$

Thus, for positive integers n, we have:

$$f(n) = f(2)\log_2(n).$$

We will define the constant C to be $-f(2)$. Since $f(2) \geq f(1) = 0$, we know that C is negative.

We next consider a set $\{p_1, p_2, \ldots, p_n\}$ of positive rational numbers such that $p_1 + p_2 + \ldots + p_n = 1$. Let N be their common denominator, i.e., $p_i = \frac{a_i}{N}$ for all i, where each a_i is an integer and $a_1 + a_2 + \ldots + a_n = N$. Let σ be a partition corresponding to the set of probabilities $\{p_1, p_2, \ldots, p_n\}$. Let π be a partition obtained by breaking up the i-th block of σ into a_i parts. Then every block of π has probability $\frac{1}{N}$. By definition of conditional entropy:

$$H(\pi|\sigma) = -\sum_i P(\sigma_i)H(\pi|\sigma_i) = -\sum_i f(a_i) - C\sum_i p_i \log_2(a_i).$$

By property 4, on the other hand, we have:

$$H(\pi|\sigma) = H(\pi) - H(\sigma) = f(N) - H(\sigma).$$

Combining the two expressions for $H(\pi|\sigma)$ gives us:

$$H(\sigma) = -C\log_2(N) + C\sum_i p_i \log_2(a_i).$$

By continuity (property 3), H must have this same formula for all sets $\{p_1, p_2, \ldots, p_n\}$ on which it is defined. This completes the proof.

We leave it as an exercise to show that the above formula for entropy actually satisfies the five postulated properties. We conclude by giving an interpretation of independence of partitions in terms of conditional entropy. Intuitively, if π and σ are independent, then their joint entropy $H(\pi \cap \sigma)$ is the sum of the individual entropies:

$$H(\pi \cap \sigma) = H(\pi) + H(\sigma).$$

In terms of conditional entropy, this says that $H(\pi \cap \sigma) = H(\pi)$.

The Shannon Coding Theorem

A consequence of Entropy Property 4 of the last section is that if we wish to answer a question X by means of a sequence of questions $S_1, S_2, \ldots, S_n$, the joint entropy of $S_1, S_2, \ldots, S_n$ must be at least as large as the entropy of X, and hence the sum of the entropies of the S_i's must be at least as large as the entropy of X. In particular, if the S_i's are yes-no questions, then $H_2(S_i) \leq 1$ and we get the crude inequality:

$$n \geq H_2(X).$$

The problem of finding a set of sufficient statistics for a random variable X is called the *coding problem* for X, and the sequence $S_1, S_2, \ldots, S_n$ is said to *code* X. As we will see in the exercises, the kinds of questions one may ask are usually restricted to some class of questions. Devising particular codes is a highly nontrivial task.

One of the reasons that coding is so nontrivial in general is that one is usually required to answer a whole sequence of questions $X_1, X_2, \ldots$, produced by some process, and as a result one would like to answer the questions in the most efficient way possible. Consider one example. Suppose that X takes values 1 through 200 each with probability 0.85, and takes values 0 with probability 7.5×10^{-4}. Then $H_2(X)$ is less than 1. Simply by counting one can see that at least 8 yes-no questions will be needed to achieve a sufficient statistic for X, even though the entropy suggests that one should be able to determine X with a single yes-no question.

Bibliography

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27(3), 1948.

[2] C. E. Shannon, "A Symbolic Analysis of Relay and Switching Circuits," *Master's thesis*, MIT, 1937. (Also *Transactions of the AIEE*, 57(12):713–723, 1938.)

[3] K. Baclawski, G.-C. Rota, and S. Billis, *Introduction to Probability Theory, Preliminary Edition*, MIT, circa 1979–1993 draft (unpublished).