TJTMI Volume 1, Issue 1 (Spring 2022), pp. 35–65

A New Perspective of Entropy

Dr. Tai-Danae Bradley
PhD, Mathematics, CUNY Graduate Center
Research Mathematician at Sandbox@Alphabet
Visiting Research Professor of Mathematics at The Master's University

Abstract: This article describes a new connection between two seemingly disparate topics in science, namely entropy and higher mathematics. It does not assume prior knowledge of either subject and begins with a brief introduction to information theory and a concept known as Shannon entropy, which we simply refer to as entropy. We then survey the vast landscape of higher mathematics, giving special attention to advanced analogues of high-school algebra and geometry known as abstract algebra and topology, respectively. Our goal is then to show that entropy, abstract algebra, and topology are inextricably linked through a version of a well-known formula from calculus known as the Leibniz rule. This result is given in the author's recent work in [Bra21], and this present article is intended to give an overview of the ideas by gently introducing them from the ground up.

Introduction

In 2009 mathematician and theoretical physicist Freeman Dyson wrote an article for the American Mathematical Society in which he surveyed the works of notable mathematicians of the past few centuries. The scientific landscape is exceedingly broad, and yet as Dyson observed, mathematicians often fall into two categories [Dys09]:

> Some mathematicians are birds, others are frogs. Birds fly high in the air and survey broad vistas of mathematics out to the far horizon. They delight in concepts that unify our thinking and bring together diverse problems from different parts of the landscape. Frogs live in the mud below and see only the flowers that grow nearby. They delight in the details of particular objects, and they solve problems one at a time.

Lest we conclude that birds are better than frogs or vice versa, Dyson quickly adds:

> Mathematics is rich and beautiful because birds give it broad visions and frogs give it intricate details.... It is stupid to claim that birds are better than frogs because they see farther, or that frogs are better than birds because they see deeper. The world of mathematics is both broad and deep, and we need birds and frogs working together to explore it.

The bird's-eye view of the landscape is a valuable perspective, and discoveries of unexpected connections between different parts of it are fascinating. But making those connections precise and rigorous often requires a frog's attention to detail. The subject of this present article has a similar bird-and-frog feel to it. It is a new connection between information theory and parts of higher mathematics related to algebra and geometry, and my recent technical article [Bra21] contains all the "froggy" details. In this article, however, we will begin by flying high in the air and surveying the ideas from a bird's vantage point, occasionally landing on the ground when necessary.

To begin, information theory fits broadly under the purview of science, technology, and engineering, while more advanced versions of algebra and geometry (called *abstract algebra* and *topology*, respectively) fit under higher mathematics. Information theory has a very applied flavor, whereas higher mathematics has a very pure flavor. The two thus reside within separate regions of the scientific landscape, and historically neither has had much much to say to the other. But in recent years a few mathematicians have unearthed parts of what seems to be an interesting bridge connecting them. Our present discussion is one of those small parts. It is a new way to understand entropy from the perspective of higher mathematics. But what is entropy? What does entropy have to do with information? What is the connection to higher mathematics? And what is meant by "higher mathematics" anyway? We will answer these questions one at a time.

A First Look at Information and Entropy

The study of information and communication finds its home in a branch of science called **information theory**. At first glance it may be surprising to learn that information has its own field of academic study, but a few moments of thought should dispel the surprise. After all, what are the basic ingredients of communication? There must be an information source (something that produces information), a channel (the medium through which information is sent), and a destination (the person or object intended to receive the information) at least. These simple ingredients quickly turn into a feast of questions. What if the channel has a limited capacity? If some information is lost along the way, can it be recovered? If so, how and to what extent? How is information stored, encoded, and decoded to produce meaningful messages? Is it possible to quantify something as general as "information" in the first place? Rephrasing these ideas in the precise language of mathematics allows such questions to be asked and answered in more useful, quantitative ways. That is precisely what mathematician and computer scientist Claude Shannon did in a seminal 1948 paper that launched the field of information theory [Sha48].

To see how mathematics can help quantify information, consider the following two statements: "The sun was shining in Los Angeles today," and "There was a blizzard in Los Angeles today." Which of those two sentences conveys more information? Readers familiar with US geography will know that it is almost always sunny in Southern California, so it is not surprising to learn that today was also sunny. Little information is conveyed in that first statement. On the other hand, it would be extraordinarily surprising—and somewhat distressing—to learn that Los Angeles was experiencing blizzard conditions. That would be a surprising scenario, and so a great deal of information is conveyed in the second statement.

These examples illustrate the intuitive idea that information and probability are inversely proportional. An event with high probability ("The sun was shining in Los Angeles today.") seems to carry little information, whereas an event with low probability ("There was a blizzard in Los Angeles today.") seems to carry lots of information. We can express this inverse relationship as a simple fraction. If an event occurs with probability p, then we might say the amount of information conveyed is $1/p$ because if p is small, then $1/p$ is large and vice versa. This is almost the same quantity Shannon used in his 1948 paper, though he instead used the logarithm of $1/p$, which is a little more convenient to work with.[1] This is a minor, technical detail for us, but let us briefly digress to explain what is meant by "convenient." Think of an event that is 100% guaranteed to happen, that is, an event that occurs with probability $p = 1$. Since the event is certain to occur, it would not be surprising to learn that it did indeed happen. Intuitively, such a lack of surprise corresponds to the fact that no information has been conveyed. Zero surprise. Zero information. And yet since $p = 1$ the fraction $1/p = 1/1 = 1$ is not zero, which goes against that intuition. On the other hand, if $p = 1$, then $\log(1/p) = \log(1) = 0$ as desired. This is one reason why logarithms are more convenient. So, with this intuition in hand, we *define* the amount of information conveyed in a single event with probability p to be the number $\log(1/p)$.

Thinking back to the weather, there are a range of possibilities—sunny, snowy, windy, cloudy, and so on. Each may occur in Los Angeles with a particular probability, so we may also compute the average (or "mean" or "expected") value of information contained in a statement describing today's weather. Generally speaking, this average amount of information has a name: entropy. More precisely it is called **Shannon entropy** to distinguish it from other notions of entropy that arise in science. (The precise formula for Shannon entropy will be given later on.) Perhaps the most familiar kind of entropy is that which appears in the Second Law of Thermodynamics, which says that the total entropy in a physical system never decreases. This kind of entropy is a measure of the amount of disorder or randomness in a system, and it

[1] The logarithm of some number x is another number that we will denote by $\log(x)$ (taken to be the natural logarithm in this paper). In particular, a useful fact to know in this article is that $\log(1) = 0$.

is conceptually the same as Shannon's version of entropy. Instead of asking about the weather in Los Angeles, we may instead ask about the speed or position of a molecule of gas, for instance. Other notions of entropy include von Neumann entropy, Tsallis entropy, Rényi entropy, and more. Our present discussion will not concern these, and so there will be no confusion if we simply refer to Shannon entropy as *entropy*. And notice that entropy, being the average of some numbers, is itself a number. It is not a vague notion or an intangible concept. It is a concrete mathematical object that has rich mathematical properties, as we will see in the page to come.

Entropy and information thus go hand-in-hand. Shannon's entropy was introduced roughly 70 years ago in a quest for a mathematical theory of communication. Thermodynamic entropy has been studied since the 1870s. Both are still of great interest to scientists today. American theoretical physicist Lee Smolin once reflected on the role that entropy has played in the past and future directions of physics, from the discovery of atoms to modern-day research on black holes [Smo01]:

> The search for the meaning of temperature and entropy of matter led to the discovery of atoms. The search for the meaning of the temperature and entropy of radiation led to the discovery of quanta. In just the same way, the search for the meaning of the temperature and entropy of a black hole is now leading to the discovery of the atomic structure of space time.

American-Israeli theoretical physicist Jacob Bekenstein, who died in 2015 and is known for his work on black hole thermodynamics, has also observed the fundamental relationship between information and the natural world [Bek03]:

> Ask anybody what the physical world is made of, and you are likely to be told matter and energy. Yet if we have learned anything from engineering, biology and physics, information is just as crucial an ingredient.... Indeed, a current trend, initiated by John A. Wheeler of Princeton University, is to regard the physical world as made of information, with energy and matter as incidentals.

So entropy and information naturally arise in investigations of the physical world. One goal of this article is to show that entropy also naturally arises in higher mathematics—that is, in the sophisticated analogues of algebra and geometry alluded to above. It is natural, however, to wonder who might find these ideas interesting. Why is such a connection worth writing about?

Discovering the Unexpected

Nothing in this discussion so far suggests entropy may be related to the world of higher mathematics. Indeed, the only math used so far has involved fractions and probabilities. Higher, abstract mathematics is nowhere in sight. And yet, as we will see below, it is in fact inevitable. Entropy is therefore a link between two things that seem very different, which may suggest that deeper connections are waiting to be discovered. In a recent interview, Edward

Witten, one of the world's premier mathematical physicists, reflected on his nearly 50 years of work in the field. When asked which current developments he is most excited about, his reply included entropy and related phenomena that may uncover some of the hidden mysteries surrounding the quantum world and Einstein's theory of general relativity [Cha21]. It is intriguing to think about the mathematics that may underlie such developments. Speculations aside, this is simply meant to whet the reader's appetite. Why would anyone devote time and attention to these technical ideas, that is, to a new way to understand entropy through abstract algebra and topology? Perhaps one day it may shed light on a new corner of science and mathematics not yet seen before. This is one reason why the ideas in the pages below are worth sharing. Consider what German theoretical physicist Max Plank, who won the 1918 Nobel prize in physics for his work in quantum theory, once said towards the end of his life [Pla48] (quoted in [Nic01, p. 201], emphasis added):

> What has led me to science and made me since youth enthusiastic for it is not the at all obvious fact that the laws of our thoughts coincide with the regularity of the flow of impressions which we receive from the external world, [and] that it is therefore possible for man to reach conclusions through pure speculation about those regularities. Here it is of essential significance that *the external world represents something independent of us, something absolute which we confront, and the search for the laws valid for this absolute appeared to me the most beautiful scientific task in life.*

We will revisit this line of thought at the end of the article, but it is now time to turn to the mathematics. The next section will open with a brief introduction to the vast landscape of higher mathematics. Our attention will then shift to two regions within that landscape: abstract algebra and topology. We will give a short explanation of each, learning just enough to see how these kinds of advanced mathematical ideas are a natural part of entropy. The discussion will then climax into the main result of my technical article [Bra21], which describes a specific link between information theory and higher mathematics. We will then close with a final brief remark on the intrigue of such discoveries.

The Landscape of Higher Mathematics

The word "mathematics" may bring to mind the procedural material we once learned in school: word problems, long division, timed multiplication worksheets, and the like. But in reality the world of mathematics extends far beyond—and is very different from—the subject we are taught at a young age. So it is natural to wonder, "What does it mean to discover *new* mathematics?" Far from being a static subject, there is a sweeping, flourishing landscape of higher mathematics, and what is taught in school occupies only a tiny fraction of that land. It is elsewhere in this terrain that we will spend most of our time in this article. Afterwards, we will have seen an example of what it means to discover new mathematics. To that end, let us give a better description of our terrain.

The phrase "higher mathematics" becomes clearer when drawing an analogy with athletics. Upon learning that someone is an athlete, we may be curious to know which sport he or she plays. The word athlete is a broad term, and merely knowing that someone *is* an athlete does not tell us much about what that person *does*. The athletic landscape is comprised of a variety of sports, and any given athlete typically specializes in one or two of them: basketball, baseball, soccer, track and field, and so on. A professional athlete may go a step further and devote decades of his or her life to excelling in a specific role within a single sport. So, athletes generally differ greatly from one another even though they share a common profession. The same is true for mathematicians.

Like the athletic landscape, the mathematical landscape is also comprised of many different realms, and a professional mathematician may spend decades of his or her life in a particular locale within one of those realms. But unlike the names of our favorite sports, the names of these mathematical realms may sound less familiar: *abstract algebra, topology, category theory, differential geometry, complex analysis*, and more. And unlike the familiar features of the athletic landscape, the rolling hills and towering landmarks of the higher mathematical landscape go largely unnoticed despite their being foundational to science and technology. They are truly hidden in plain sight. There have, however, been some attempts at making this invisible landscape visible. One excellent example is the beautifully detailed map created by mathematician Martin Kuppe in the 2018 article [Bra18, p. 23]. The vintage-colored map entitled "Mathematistian" takes the voyager on a quest through magical mathematical lands with cleverly devised names, such as "Probabilistan" and "Statistigrad," the "Plains of Analysis," and the "Ocean of Logic." Two such regions within Kuppe's map are most relevant to our present discussion on entropy, namely the "Califate of Al-Gebra" and the "Tundra of Topology." The first is a play on words and refers to a branch of math called abstract algebra, which (as the name suggests) is a more sophisticated, abstract version of the subject we learn in high school. The second is topology, which is a more sophisticated, abstract version of geometry. Entropy is inherently algebraic and topological, so it will be helpful to first take a brief journal through both.

Abstract Algebra: Math Beyond Numbers

The word algebra originates from the Arabic word *al-jabr*, which means "reunion of broken parts." It appears in the title of a ninth century book on the subject written by Persian scholar Mohammad ibn Mûsâ al-Khowârizmî [Gan26] and brings to mind the basic concept of combining things to form something new. If we have two numbers, for instance, then we can combine them, say by multiplication, to form a new number: $2 \times 3 = 6$. Of course there is nothing special about the numbers 2 and 3 in the previous sentence. If we have *any* two numbers, say x and y, then we can multiply them to obtain a new number $x \times y$, also denoted by

xy. This simple idea of combining things (whether by multiplication or addition or something else) quickly leads to the kind of algebra we learn in high school, where we are tasked with assignments such as "Simplify the expression $(x^3)^2y^4x^{-1}$ using laws of exponents" and "Factor the quadratic polynomial $x^2 + 7x + 12$" and the like. But this kind of algebra is nothing like the algebra studied at the graduate and research levels, which is called **abstract algebra** or **modern algebra**. To get a feel for the difference, it will help to think like a bird and not a frog. Forget the details. Forget the symbols. Forget words like "exponents" and "polynomial." Instead, think back to the simple idea mentioned above: *combining things to form something new.* Whenever things can be combined, whether they are numbers or something else, there is likely algebraic structure behind the scenes.

Multiplication of numbers is just one example. Consider human language, for instance, where words combine to form longer expressions. *Yellow* is a word, and *banana* is a word, and we can "multiply" them to form the new expression *yellow banana*. The technical term for stacking words side-by-side is concatenation, as opposed to multiplication. But the term is not so important. The concept is. Concatenation and multiplication are conceptually similar. They both allow us to combine things to form something new. There is, however, a notable difference. The order in which we multiply numbers does not matter, whereas the order in which we concatenate words certainly matters. The product 2×3 is the same as 3×2, but *yellow banana* is not the same as *banana yellow*. This property is called commutativity. Multiplication of numbers is said to be commutative, whereas concatenation of English words is not commutative. Associativity, on the other hand, is a property shared by both multiplication and concatenation. When multiplying three numbers, it does not matter which two are multiplied first: $(2 \times 3) \times 5$ is the same as $2 \times (3 \times 5)$, and the analogous holds for concatenation of English words.

This gives a taste of abstract algebra, where concrete details are abstracted away. In this branch of mathematics, the particulars of *what* is being combined, whether they are numbers, or words, or something else, is not the main focus. More important is the abstract structure, that is, the operation itself (multiplication or concatenation or...) and the properties it possesses. Informally speaking, any collection of things that can be combined—that is, where some notion of "multiplication" makes sense—is called **an algebra**, and when the multiplication possesses certain properties, the algebra is usually given a descriptive name. Examples include commutative algebras, associative algebras, Lie algebras, A_∞-algebras, and more.[2] In fact, an algebra is just *one* kind of algebraic structure. There are many more. Mathematicians also study groups, rings, fields, and vector spaces, to name a few. Each of these algebraic

[2] More formally, an algebra is defined to be a *vector space* equipped with a way to multiply vectors. (These words will be familiar to students of linear algebra.) Lie algebras are named after Norwegian mathematician Marius Sophus Lie (1842–1899) and are used widely in physics. An A_∞-algebra is one where the multiplication is not associative on the nose. Instead, it is only associative up to some wiggle room. This kind of structure appears in topology, the subject of the next section.

objects has a different (and sometimes multiple) notion(s) of combining things, and each plays a different role on the mathematical stage.

What, then, is the point of abstracting away details? Why pursue this line of thinking? One advantage is that it is clarifying. It helps us see relationships between things that initially seem unrelated. Karen H. Parshall, an American historian of mathematics, summarizes this nicely in an article on the history of abstract algebra in the *The Princeton Companion to Mathematics* [GBGL08, Section II.3]:

> One objective of this new type of algebra is to understand the underlying structure of the objects and, in doing so, to build entire theories of groups or rings or fields. These abstract theories may then be applied in diverse settings where the basic axioms are satisfied but where it may not be at all apparent a priori that a group or ring or field may be lurking. This, in fact, is one of modern algebra's great strengths: once we have proved a general fact about an algebraic structure, there is no need to prove that fact separately each time we come across an instance of that structure. This abstract approach allows us to recognize that contexts that may look quite different are in fact importantly similar.

With that, it appears we have progressed from high-school algebra to advanced mathematics rather quickly. Here is the bottom line. First, the claim that entropy can be understood in terms of algebra refers to *abstract* algebra and not to high-school algebra. Second, know that abstract algebra encompasses a large zoo of advanced mathematical structures. It is not necessary to know about any of them in detail, but it is good to be aware of their existence. This will help us make the connection to entropy later on. Here is a quick preview: it turns out that probabilities exhibit both algebraic and topological structure, and entropy interacts very nicely with both. More to the point, the *way* in which entropy interacts with algebra and topology is its defining characteristic—its fingerprint, so to speak. To understand this claim, however, we must first understand what topology is.

Topology: Geometry's Flexible Cousin

Like geometry, **topology** is a branch of mathematics that involves the study of shapes. But unlike geometry, where angles, areas, lengths, and size take center stage, topology focuses on something else. What else can be said about shapes if not these features? *A lot*. Consider, for instance, the notion of sameness. What does it mean for two shapes to be equivalent? This may seem to be an innocent question, but as mathematician Barry Mazur once astutely observed [Maz07],

> One can't do mathematics for more than ten minutes without grappling, in some way or other, with the slippery notion of *equality*. Slippery, because the way in which objects are presented to us hardly ever, perhaps never, immediately tells us—without further commentary—when two of them are to be considered equal. We even see this, for example, if we try to define real numbers as decimals, and then have to mention aliases like $20 = 19.999\ldots$, a fact not unknown to the

merchants who price their items $19.99. The heart and soul of much mathematics consists of the fact that the "same" object can be presented to us in different ways.

A key difference between geometry and topology is how each answers the question, "When are two shapes considered the same?" In topology, shapes are thought of as malleable and pliable—made of something like Play-Doh—and two shapes are considered to be the same if one can be molded and deformed into the other without ever tearing or ripping the shape. As an example, the familiar shapes shown in Figure 1 are all considered to be equivalent. There is no difference between triangle, a square, a hexagon, or a circle in the eyes of topology. A triangle made of Play-Doh, for instance, can be deformed into a circle by smoothing the corners and rounding out the edges. So, rather than focusing on their differences—a triangle has three straight sides and a circle does not—topology instead embraces what they have in common: both shapes enclose a region on the page. On the other hand, a circle and a line are fundamentally different from this perspective. A line is not "closed" in the way that a circle is.

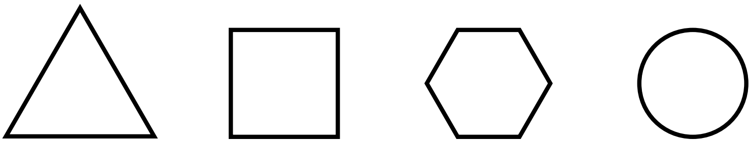


Figure 1: In topology a triangle, square, hexagon, and circle are considered to be equivalent shapes.

This notion of sameness is neither arbitrary nor meaningless mathematics, but rather is a consequence of formalizing another fundamental idea, namely that of closeness. In Mazur's quote above, we understand that the number 19.99 is relatively close to 20, and that 19.999 is closer, and that 19.9999 is closer still. Said differently, the closeness of the numbers is measured by the distance between them. We also have the intuition that two items whose prices are close together will have values that are close together, as well. We could purchase a new book for $19.99, for instance, but would not expect that an extra penny could afford us a new luxury yacht for $20. The concept of mapping between objects (money and items in this case) in a way that preserves closeness is at the heart of topology. It is called *continuity*. More precisely, topology allows us to generalize the notion of distance in settings beyond numbers, and it does so in a way that also formalizes the idea of continuity. Simply put, any set—that is, any collection of things or elements—can be equipped with extra structure known as **a topology**. Very roughly speaking, "a topology on a set X" is a mathematician's way of declaring which elements in X are close to each other. When considered together as a pair, both the set and its topology are referred to as a **topological space**. The formal definition is quite abstract and will not be given here, but it encompasses many familiar shapes. Lines, circles, triangles, squares, spheres, pyramids, cubes, and a plethora of more exotic shapes are all examples of topological spaces.

So, the field of topology generalizes distance or closeness, and this informs how one should address the question of sameness. Two topological spaces are—again, roughly speaking—considered to be the "same" if they can be transformed into each other while preserving their respective versions of closeness. Such transformations include the deformations of Play-Doh triangles and circles mentioned above. While these ideas may sound quite different from high-school geometry, topology is a deeply rich and useful branch of mathematics, with applications ranging from explorations of the shape of space [Wee20] to DNA modeling [Ada04] to data analysis [CVJ21] and much more [Ghr14].

Entropy + Algebra + Topology = ?

Now that we have some familiarity with the higher mathematical landscape—algebra and topology in particular—we are ready to see how they are inextricably related to entropy. Our introductory discussion began with the observation that information and probability are inversely proportional. We also said that Shannon entropy, or entropy for short, is the average amount of information contained in a collection of probabilities. In other words, entropy is a number associated to a list of probabilities, and we interpret that number as a measure of information. That number is computed by a particular formula that will be shared below. The remainder of this article will rely heavily on that formula, and it will help to first establish some terminology in the next section. The subsequent pages aim to give a bird's-eye view of the mathematics while occasionally providing frog-level details for interested readers. Such technical paragraphs are decorated with a triangle and labeled "▶ *In more detail*." These paragraphs are included for the enjoyment of those who wish to dig deeper into the mathematics and may be safely skipped, if desired.

Entropy is a Number

We begin by introducing basic terminology. To start, a list of probabilities is called a probability distribution. That is, a **probability distribution** is a finite list of numbers between 0 and 1 whose sum is 1. For example, $\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\left(\frac{2}{5}, \frac{1}{2}, \frac{1}{10}\right)$ are both probability distributions whereas $(7, 3)$ and $\left(-\frac{2}{5}, -\frac{1}{2}, -\frac{1}{10}\right)$ are not. In the first example, $\frac{1}{2}$ and $\frac{1}{2}$ may represent the respective probabilities of landing a heads or tails on a fair coin toss, while in the second example $\frac{2}{5}, \frac{1}{2}$, and $\frac{1}{10}$ may represent the respective probabilities of choosing cereal, oatmeal, or fruit for breakfast. In both cases, it helps to think of probabilities as numbers associated to a finite set of options: the first option (choosing cereal), the second option (choosing oatmeal), the third option (choosing fruit), and so on.

Said more formally, given a **natural number** n (that is, a whole number $1, 2, 3, \ldots$), a probability distribution on a set $\{1, 2, \ldots, n\}$ is a list of nonnegative real numbers $p = (p_1, p_2, \ldots, p_n)$

satisfying $p_1 + p_2 + \cdots + p_n = 1$. Elements in the set $\{1, 2, \ldots, n\}$ may be thought of as enumerating the different options or outcomes, each of which is assigned a particular probability. The letter p is being used as shorthand to represent the full list of numbers $(p_1, p_2, \ldots, p_n)$. Teasing this out with $n = 2$, suppose we have a set of two elements $\{1, 2\}$ corresponding to the two faces of a coin, namely heads (option #1) or tails (option #2). If we let $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2}$, then the list $p = (p_1, p_2)$ is the first example of a probability distribution given above.

In addition to viewing probability distributions as lists, they can also be visualized with pictures. Figure 2 gives an example. There, the probability distribution $p = \left(\frac{1}{2}, \frac{1}{2}\right)$ associated to a coin toss is represented by a stick-like **tree** with one **root** and two **leaves**, each representing a face of the coin. The words "tree" and "root" and "leaf" are technical terms used in a branch of mathematics called graph theory. (Even more formally, the pictures in Figure 2 are known as **planar rooted trees**.) The right-hand side of the figure shows a tidier version of this by labeling each leaf with its corresponding probability. We can likewise depict an arbitrary probability distribution $p = (p_1, p_2, \ldots, p_n)$ as a tree with one root and n leaves that are labeled by the individual probabilities as in Figure 3. Visualizing probability distributions in this way will be a worthwhile adjustment for us, as we will see later on. A picture is worth a thousand words in mathematics, too.

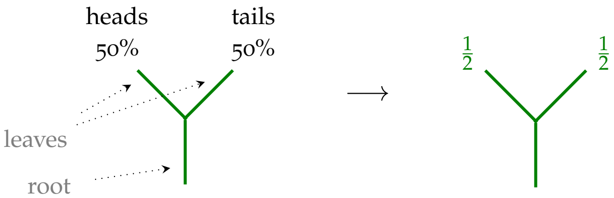


Figure 2: The probability distribution $p = \left(\frac{1}{2}, \frac{1}{2}\right)$ can be visualized as a tree with a root and one leaf for each outcome of a coin toss, labeled with the respective probabilities.

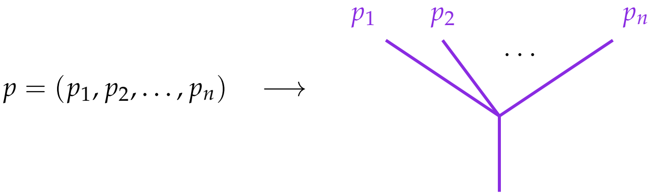


Figure 3: Any probability distribution on n elements can be illustrated as a tree with n leaves labeled by the individual probabilities.

What's more, every probability distribution p has a number associated to it called entropy. This number, which we will denote by $H(p)$, is given explicitly in the following formula.

Definition. *The **entropy** of a probability distribution $p = (p_1, p_2, \dots, p_n)$ is defined to be*

$$H(p) = -p_1 \log(p_1) - p_2 \log(p_2) - \cdots - p_n \log(p_n). \tag{1}$$

It will be helpful to gain intuition for this expression, and we may start by comparing it to our opening remarks on entropy at the beginning of this article. There we defined the information contained in a single event with probability p to be the number $\log(1/p)$. By basic properties of logarithms, this number is the same as $\log(1) - \log(p)$ which is equal to $-\log(p)$ because $\log(1) = 0$. In other words, the information in an event with probability p is the number $-\log(p)$. But notice the formula for entropy in Equation (1) does not merely consider one event. It considers *multiple* events, each of which has its own probability. So the formula first computes the information associated to each event, that is, $-\log(p_1)$ and $-\log(p_2)$ and so on. Then it computes the average of those numbers by multiplying each by its respective probability and adding them. That is what is displayed in Equation (1), where the minus signs are important. The natural logarithm $\log(p)$ is negative whenever its input p is between 0 and 1, so $-\log(p)$ is nonnegative. In other words, the number $H(p)$ is always either positive or zero.

Examples are helpful. Think of an event that is guaranteed to happen. An avid coffee drinker, for instance, will look forward to a cup of coffee each day. Suppose there is a 100% chance they will have coffee each day and a 0% chance they will not. This scenario corresponds to the probability distribution $p = (1, 0)$. If we were to learn that this person indeed drank coffee today, then we would not be surprised. The behavior is expected, so no information has been conveyed. This lack of surprise is mathematically represented by the entropy of p in this example, which is zero:

$$H(p) = -1 \log(1) - 0 \log(0) = -\log(1) = 0.$$

Zero entropy corresponds to zero uncertainty. Now think about the other extreme, that is, an event with maximal uncertainty. Consider the outcome of tossing a coin. The result is either heads or tails, and neither is more likely assuming it is a fair coin. We expect the entropy of the corresponding probability distribution $p = \left(\frac{1}{2}, \frac{1}{2}\right)$ to be positive since the outcome is totally uncertain. This is indeed the case:

$$H(p) = -\tfrac{1}{2} \log\left(\tfrac{1}{2}\right) - \tfrac{1}{2} \log\left(\tfrac{1}{2}\right) = -\log\left(\tfrac{1}{2}\right) = -\log(1) - (-\log(2)) = \log(2).$$

This computation is a special instance of a more general pattern. Whenever there are n outcomes each with equal probability $\frac{1}{n}$, the entropy of the resulting probability distribution $\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$ will always be $\log(n)$, and this turns out to be the maximum possible value. In other words, one can show that $0 \leq H(p) \leq \log(n)$ for any probability distribution p on a set with n elements. So, entropy is akin to a measure of surprise or uncertainty, it has a formula, and we have now seen two extremal examples. With these basics in hand, it is now time for a slight shift in perspective. We have seen that entropy is number, but in actuality, that number is a shadow of something more. Entropy is not merely a number. It is a *function.*

Entropy is a Function

Recall that every probability distribution p on n elements corresponds to a number $H(p)$. The quantifier "every" suggests there is a function lurking behind the scenes. Indeed, the process of assigning a real number to a probability distribution is precisely what it means to have a function from the set of *all* probability distributions on n things to the set of real numbers. New mathematical notation allows us to restate this idea more conveniently:

$$\textit{For each natural number } n\textit{, there is a function } H\colon \Delta_n \to \mathbb{R}. \tag{2}$$

We have used the letter H as the name of the function that assigns to a probability distribution p its entropy $H(p)$. We will also use Δ_n to denote the set of all possible probability distributions on the set $\{1, 2, \ldots, n\}$. For example,

$$\left(\tfrac{1}{2}, \tfrac{1}{2}\right) \text{ and } (1, 0) \text{ are elements of the set } \Delta_2,$$

$$\left(\tfrac{4}{10}, \tfrac{5}{10}, \tfrac{1}{10}\right) \text{ and } \left(\tfrac{1}{5}, \tfrac{3}{5}, \tfrac{1}{5}\right) \text{ are elements of the set } \Delta_3,$$

$$\left(\tfrac{2}{7}, \tfrac{1}{7}, 0, \tfrac{4}{7}\right) \text{ and } \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}\right) \text{ are elements of the set } \Delta_4,$$

and so on. Mathematicians usually prefer to lower the index by 1 and write Δ^{n-1} instead, but constantly "being off by 1" would be inconvenient in the work to come, so we will use Δ_n instead.[3] In our menagerie of mathematical notation, we have also used the symbol $\mathbb{R}$ to denote the set of all real numbers. Moreover, the letter H paired with an arrow $\to$ represents the assignment of a probability distribution p to the number $H(p)$. The notation $H\colon \Delta_n \to \mathbb{R}$ is very convenient, so we will always use analogous notation $f\colon X \to Y$ to mean "a function f from a set X to a set Y that assigns to each input x in X one output $f(x)$ in Y."

In summary, entropy defines a function H, and we will refer to both H and its values $H(p)$ by the same word: entropy. There is a subtlety, however. The claim that "entropy defines a function" is not the full truth. Entropy does not merely define one function. It defines *infinitely many* functions. The number n provides a clue.

[3] The notation Δ_n is pronounced "delta n" and is a clever choice. We will see later that the set of all probability distributions on three elements can be visualized as a triangle $\triangle$.

Entropy is a Collection of Functions

Look back to the sentence displayed in (2): "For each natural number n, there is a function H." The word order of that sentence implies that H depends on the value of n. Since there are infinitely many natural numbers $n = 1, 2, 3, \ldots$, there are necessarily infinitely many Hs, as well. It would be helpful to indicate this dependency in our notation and write H_n instead of H. Doing so would allow us to refine the sentence in (2) as follows:

$$\textit{Entropy defines a collection of functions } \{H_n \colon \Delta_n \to \mathbb{R}\}. \tag{3}$$

This is better. Even so, the subscripts are rather cumbersome to carry around, so we will continue to omit them and write H instead of H_n. Simply remember that H depends on n. Notation aside, here is the essential idea: we must *first* choose a natural number n to specify the length of a probability distribution p, and *then* we may compute its entropy $H(p)$. In this way, entropy defines infinitely many functions, one for each natural number. There is $H \colon \Delta_1 \to \mathbb{R}$ and $H \colon \Delta_2 \to \mathbb{R}$ and $H \colon \Delta_3 \to \mathbb{R}$, and so on.

LET US NOW pause and consider our changes in perspective. We began with the idea that entropy is a number. Then we observed that entropy defines a function. Now we see it defines a collection of infinitely many functions. There are many layers to entropy with still more to come. The functions $H \colon \Delta_n \to \mathbb{R}$ turn out to possess nice mathematical properties, both individually and collectively. For instance, if we were to change the probabilities of a probability distribution p by a small amount, then it can be shown that the entropy $H(p)$ would change by a small amount, as well. Conceptually this invokes the idea of "closeness." Two probability distributions that are close or similar will have entropies that are likewise close or similar. This intuitive property has a name, which we briefly mentioned earlier—**continuity**. In other words, each of the functions $H \colon \Delta_n \to \mathbb{R}$ are not *merely* functions; they are said to be *continuous* functions. We now find ourselves in the world of topology.

Entropy is a Collection of Continuous Functions

Recall that the symbol Δ_n represents the set of all probability distributions on n things. Importantly, each of these sets $\Delta_1, \Delta_2, \Delta_3, \ldots$ is not merely a set. There is a standard way in which they are in fact topological spaces. The first few are familiar and simple shapes. It can be shown that Δ_1 is a point, Δ_2 is a line segment, Δ_3 is a triangle, and Δ_4 is a pyramid, as displayed in Figure 4.

► *In more detail.* Consider the case when $n = 1$. The set Δ_1 consists of all probability distributions on a single element. Such a probability distribution is a list consisting of a single number, and moreover that number must be equal to 1. So Δ_1 is the singleton number 1, which can be viewed as a point on the number line, as in Figure 4. A point is not a very interesting shape, but it is a shape nonetheless. Consider the more interesting case when $n = 2$. The set Δ_2 of probability distributions on two elements is the set of all pairs of nonnegative numbers (x, y) whose sum is 1; that is, $x + y = 1$. We can rearrange this equation to see it is equivalent to $y = 1 - x$, which is the equation of a line. Since we have further restricted x and y to be nonnegative, we may conclude that Δ_2 is a line segment in the positive quadrant of the two-dimensional plane as in Figure 4. So, a point on that line segment is a probability distribution on two elements. Moving on to the case when $n = 3$, the set Δ_3 of probability distributions on three elements is the set of all triples of nonnegative numbers (x, y, z) whose sum is 1; that is $x + y + z = 1$ or equivalently $z = 1 - x - y$, which is the equation of a plane in three-dimensional space. Restricting to nonnegative coordinates gives rise to the triangular slice of the plane as in Figure 4. So Δ_3 is likewise a very simple shape. It is a triangle, and a point on that triangle is a probability distribution on three elements. Understanding the case when $n = 4$ requires more work, but it can be shown that the set Δ_4 of all probability distributions on four elements is a tetrahedron, or triangular pyramid, as shown in Figure 4. A point on that tetrahedron is a probability distribution on four elements.

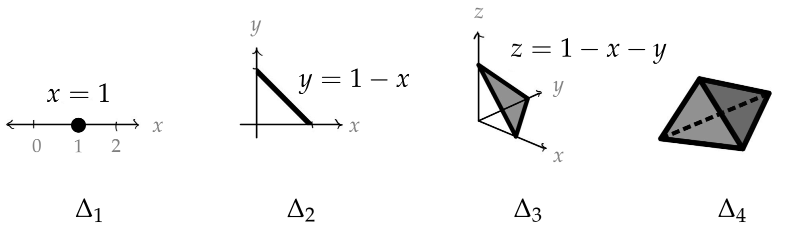


Figure 4: For small values of n, the topological spaces Δ_n are easy to visualize. For example, Δ_1 is a point, Δ_2 is a line segment, Δ_3 is a triangle, and Δ_4 is a tetrahedron.

The picture for Δ_n becomes harder to visualize when n is greater than four, but each Δ_n is indeed a topological space. The set of real numbers $\mathbb{R}$ is *also* a topological space, and the upshot is that the functions $H\colon \Delta_n \to \mathbb{R}$ defined by entropy are continuous with respect to the topologies. The Δ_n further play an especially fundamental role in topology, where for each n the topological space Δ_n is called an $n - 1$**-simplex**.[4] For small values of n we have seen that simplices coincide with familiar objects: a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. In general, simplices are often used by topologists as building blocks for more complicated topological spaces, somewhat like Lego pieces. A hierarchy is already apparent: a tetrahedron is built up from triangles,

[4] Recall that our Δ_n are traditionally denoted by Δ^{n-1}, which is why we have called Δ_n an $n - 1$-simplex rather than an n-simplex.

a triangle is built up from line segments, and a line segment consists of points. The popular online periodical *Quanta Magazine* recently expounded upon this idea [HE21]:

> Many topological shapes can be built by gluing together pieces of different dimensions.... Individual pieces of the shape are grouped by dimension and then arranged hierarchically: The first level contains all the points, the next level contains all the lines, and so on. (There's also an empty zeroth level, which simply serves as a foundation.) Each level is connected to the one below it by arrows, which indicate how they are glued together. For example, a solid triangle is linked to the three edges that form its boundary.

That "each level" is connected by arrows is an additional part of the mathematical theory that is not the focus of this article. But the quote above alludes to the fact that topological simplices are part of a larger framework that enables mathematicians to translate difficult topological problems into a language that is easier to work with. This framework is called homology, and the *Quanta* article reference above is a good place to learn more. It all starts by breaking up complicated topological spaces into little pieces or simplices. By definition, these same simplices have a probabilistic interpretation, and whenever there are probabilities, entropy is not far behind. Topology and entropy are thus inextricably linked. As we will soon see, algebra is just as inevitable.

LET US AGAIN pause to take inventory of our progress so far. We understand that entropy is not just a number but is rather a collection of infinitely many functions. Those functions moreover behave well from the perspective of topology because they are continuous. There now remains one final property of entropy to know about. One more layer to peel back. It is the fulcrum of this article and manifests itself whenever an event or outcome can be viewed as a composite process. This concept is best explained with an example.

Composing Probabilities: Where Algebra and Topology Meet

This section contains an example of "composing" probability distributions, an operation that will take center stage in our understanding of entropy. Both the example and ensuing discussion are heavily inspired by a 2011 informal article written by mathematical physicist John Baez [Bae11], as well as a talk given by mathematician Tom Leinster at the Centre International de Rencontres Mathématiques in 2017 [Lei]. The work and masterful expositions of both Baez and Leinster served as a primary source of motivation for the main result in [Bra21] that we are en route to unveiling, as well as the narrative we are sharing along the way.

Consider the following example. Suppose we flip a fair coin and then decide what to eat for breakfast or dinner depending on which face the coin lands. There is a 50-50 chance the coin will land on heads or tails, which corresponds to the probability distribution $p = \left(\frac{1}{2}, \frac{1}{2}\right)$. As shown previously in Figure 2, we may further represent p as a (green) tree with two leaves labeled by the probabilities. Now, if the coin lands on heads, suppose we will choose what to have for breakfast. Say there is a 40% chance we choose cereal, a 50% chance we choose oatmeal, and a 10% chance we choose fruit. This probability distribution on three breakfast items will be denoted by $q = \left(\frac{2}{5}, \frac{1}{2}, \frac{1}{10}\right)$, which is a point in Δ_3. The picture for this probability distribution is on the left-hand side of Figure 5. If the coin instead lands on tails, then suppose we will decide what to have for dinner. Say there is a 30% chance we choose pizza and a 70% chance we choose stir fry. Denote this probability distribution on two dinner options by $r = (\frac{3}{10}, \frac{7}{10})$, which is a point in Δ_2. The picture of this probability distribution is shown on the right-hand side of Figure 5.

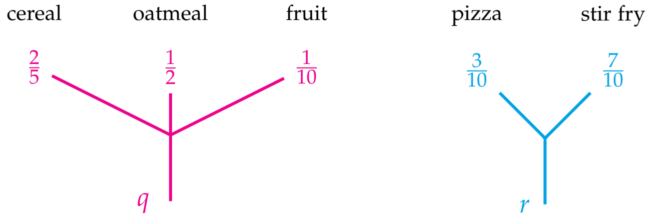


Figure 5: The probability distribution q on three elements can be represented as a tree with three leaves labeled by the probabilities, and similarly for r.

Notice there are five possible outcomes of this two-step process: cereal, oatmeal, fruit, pizza, or stir fry, depending on the coin toss. Importantly, each of those five outcomes has a probability associated to it, and those probabilities are easy to calculate. For example, the probability of flipping heads (a 50% chance) and then choosing cereal (a 40% chance) is the product of the probabilities of each individual outcome, namely $\frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$ or 20%. Similarly, the probability of flipping heads (a 50% chance) and then choosing oatmeal (a 50% chance) is equal to $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ or 25%, and the probability of flipping heads and choosing fruit is equal to $\frac{1}{2} \times \frac{1}{10} = \frac{1}{20}$ or 5%. This, too, has a visual counterpart. One can imagine grafting the root of the pink tree representing q onto the first leaf of p, since the first leaf corresponds to flipping heads, and moreover letting the green probability $\frac{1}{2}$ for "heads" propagate up through the three leaves of q. Figure 6 shows the picture. Similar calculations show that the probability of flipping tails and choosing pizza is $\frac{3}{20}$, and the probability of flipping tails and choosing stir fry is $\frac{7}{20}$. The corresponding picture is analogous. To summarize this example, the process of flipping a coin and then choosing a meal gives rise to a new probability distribution on five things—cereal, oatmeal, fruit, pizza, stir fry—and based on our calculations above, that new probability distribution is equal to $\left(\frac{1}{5}, \frac{1}{4}, \frac{1}{20}, \frac{3}{20}, \frac{7}{20}\right)$. It is a point in Δ_5, and its picture is shown in Figure 7.

This process of combining p and q and r to obtain a new probability distribution is not a standard operation. It would not be found in the table of contents of a typical textbook

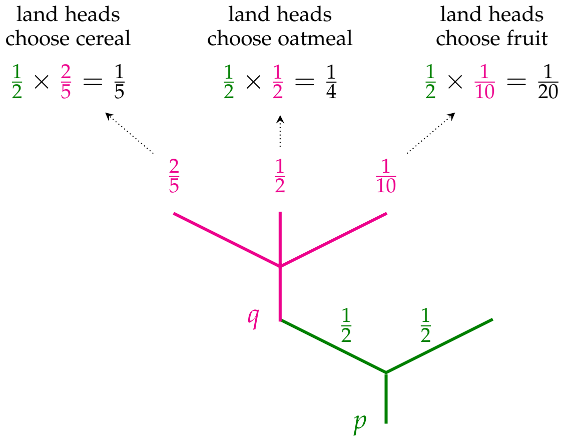


Figure 6: To compute the probability of flipping heads and choosing cereal, oatmeal, or fruit, simply multiply the probability for each food item by $\frac{1}{2}$. In our graphical notation, this is represented by grafting the root of the pink tree into the first leaf of the green tree.

on probability and statistics, for instance. Rather, it is an example of something new, and it is an essential ingredient in our discussion on entropy. To see how, it will be helpful to first introduce new notation for such composite probability distributions. Since this is a new mathematical construction, we are at liberty to invent our own notation for it. What should we choose? That is, what notation should we use to denote the probability distribution in Figure 7? We obtained the probabilities by multiplication, so it might be instructive to incorporate the "times symbol" × somehow, which would also invoke the essence of abstract algebra introduced earlier, namely the notion of *combining things to form something new*. That is indeed what we have done here. A probability distribution p has been combined with probability distributions q and r to obtain a new probability distribution. We may therefore wish to denote this new distribution by, say, "$p \times (q,r)$" to remind us that the probabilities in q and r were multiplied by the probabilities in p to give rise to new probabilities. This would be a reasonable choice. Nevertheless, we will replace the × with a circle ∘ and write the following instead:

$$p \circ (q,r) = \left(\frac{1}{5}, \frac{1}{4}, \frac{1}{20}, \frac{3}{20}, \frac{7}{20}\right).$$

Why a circle? In a remarkable turn of events, this new mathematical operation was recently found to be *not* new at all. That is, our way of "multiplying" probability distributions—or rather, our way of *composing* them—is just one example of what can be represented by the tree grafting shown above. Many other kinds of composite mathematical objects fit neatly into the same template. That template is called an **operad**, and in the operadic literature a circle ∘ is standard notation for the kind of composition seen in the example above. Moreover, our ability to compose probabilities is summarized in the fact that *topological simplices* $\Delta_1, \Delta_2, \Delta_3, \ldots$ *form*

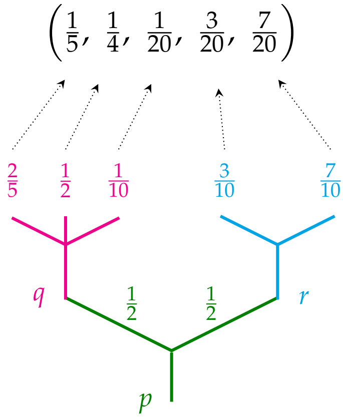


Figure 7: There are five possible outcomes from flipping a coin and then choosing a meal. Each of those outcomes has a certain probability, which is obtained by multiplying individual probabilities from p by those from q or r. This composite process can be illustrated by grafting the respective trees as shown.

an operad.

So, what is an operad? The formal definition strikes a balance between the concrete and the abstract. It is concrete enough to be useful; it is abstract enough to subsume many examples. It is also rather technical and thus beyond the scope of this discussion. But some intuition will be helpful. Loosely speaking, an operad is an abstract mathematical tool that keeps track of certain properties of operations such as commutativity and associativity. One might think of these properties as "flavors" of multiplication. In the culinary world, there are many types of foods, and those foods come in a variety of flavors. Similarly, in the mathematical world, there are many types of operations and each may have a different flavor. Multiplication of numbers is just one example, but there are many more. Mathematicians frequently "multiply" things that are not numbers and then ask whether the flavors of those operations are interesting. We noted earlier that these more elaborate algebraic objects are well-known in mathematics with names such as commutative algebras, associative algebras, and Lie algebras, among many others. The language of operads distills these objects down to their core, leaving only the bare essentials that distinguish one algebra from another. For this to be useful, however, mathematicians had to first pin down an appropriate definition—one that was both rigorous and abstract, and this was accomplished in the early 1970s. An **operad** is defined to be a collection of "abstract operations" that accept n inputs for each natural number $n = 1, 2, 3, \ldots$ together with a notion of composing them, and moreover that composition must satisfy a list of reasonable axioms. For an accessible introduction to the formal definition, see [Sta04, Bra17a, Bra17b] as well as [Lei21, Chapter 12.1].

▶ *In more detail.* Though a mouthful, the formal definition is conceptually simple. Pictures are especially helpful. An abstract operation generalizes the concept of a function that accepts n inputs and combines them to produce one output. That output can then be used as one of the inputs for a different function. Multiplication, for instance, is a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ that accepts a pair of numbers (x, y) as input and computes their product $f(x, y) = xy$ as output, which can be visualized by the cartoon in Figure 8. If we wish, we can then use the output xy, which is just a number, as one of the inputs for some *other* function. This the basic idea behind function composition, a concept usually taught in high-school or college algebra. More generally, abstract operations can likewise be illustrated as cartoon-like trees with leaves that enumerate inputs and a single root to represent the output, exactly like those appearing in Figures 2, 3, 5, 6, and 7. In this way, we begin to see how probabilities start to fit into the theory of operads. But there is a mental exercise here. Probability distributions are now on par with abstract operations, which may seem confusing. How is a list of numbers an operation? It is not, really. It would be better to represent a probability distribution $p = (p_1, p_2, \ldots, p_n)$ by an actual (continuous) function $f\colon \mathbb{R}^n \to \mathbb{R}$ that accepts a list of numbers $x = (x_1, x_2, \ldots, x_n)$ as input and that outputs a single number $f(x)$. In such a passage from *probability distributions* to *functions*, the formula for f may involve the probabilities $p_1, p_2, \ldots, p_n$ somehow. As an example, given a probability distribution p and a list of arbitrary numbers $x = (x_1, x_2, \ldots, x_n)$ as input, we could define $f(x)$ to be equal to the sum $p_1x_1 + p_2x_2 + \cdots + p_nx_n$, a number known as the "dot product" between p and x. In general, the passage from probability distributions to functions is a standard part of the theory of operads. Traditionally, such passages are (quite confusingly) called "algebras over the operad," although one might prefer to call them **representations of the operad** [Bra21].

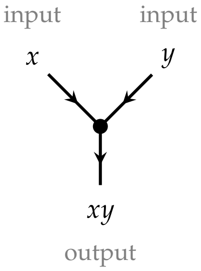


Figure 8: Multiplication can be visualized as a tree with two leaves and one root. The arrows indicate the input-output flow.

Needless to say, this topic is a specialized one. Not all mathematicians work with operads or are familiar with them, and yet the prevalence of operads throughout the higher mathematical landscape is quite astounding. Here is an overview given in [MSS02], a book written in 2002 for graduate students, research mathematicians, and mathematical physicists:

> Significant examples [of operads] first appeared in the 1960's though the formal definition and appropriate generality waited for the 1970's. These early occurrences were in algebraic topology in the study of (iterated) loop spaces and their chain algebras. In the 1990's there was a renaissance and further development of the theory inspired by the discovery of new relationships with graph cohomology, representation theory, algebraic geometry, derived categories, Morse theory, symplectic and contact geometry, combinatorics, knot theory, moduli spaces, cohomology and, not least, theoretical physics, especially string field theory and deformation quantization.

While not all the terms may sound familiar, the variety is unmistakable. One more can now be added to the list: information theory. Around 2010, Leinster observed that the composition of probabilities described above is precisely what is needed to have an operad. That is, Leinster showed that the collection of topological simplices $\Delta_1, \Delta_2, \Delta_3, \ldots$ admits the structure of an operad [Bae11, Lei21]. The upshot is that the way we have composed probability distributions p, q and r to obtain $p \circ (q, r)$ in our coin-food example is neither homeless nor isolated in the land of mathematics. It finds a natural home in the established theory of operads. The composition of probabilities is moreover a collision between the worlds of algebra and topology. It is algebraic because we are combining probability distributions. It is topological because those probability distributions are elements of topological simplices. And as we will now see, entropy is not far behind.

The Chain Rule for Entropy

Recall that in its most basic sense, entropy is a number associated to a probability distribution. Our coin-food example involved four probability distributions—namely, p (a coin toss) and q (breakfast choices) and r (dinner choices) and their composition $p \circ (q, r)$—and each has an entropy associated to it as in Equation (1). Because the composite probability distribution $p \circ (q, r)$ is built up from three individual distributions, it is natural to wonder whether the entropy of $p \circ (q, r)$ can *likewise* be built up from the entropies of the three individual distributions. In other words, can the formula for $H(p \circ (q, r))$ be reexpressed in terms of $H(p)$ and $H(q)$ and $H(r)$? Perhaps, for instance, the entropy of the composite distribution is equal to the sum of the entropies of the individual distributions: $H(p \circ (q, r)) \overset{?}{=} H(p) + H(q) + H(r)$. While a good guess, this is not the case. But one can show the equality *does* hold *if* the entropies of q and r are multiplied by the probabilities of p, as follows:

$$H(p \circ (q, r)) = H(p) + \tfrac{1}{2} H(q) + \tfrac{1}{2} H(r). \tag{4}$$

It may not be obvious why this modified equality is indeed true, but it can be easily verified using basic arithmetic and high school algebra. Simply apply the formula in Equation (1) to $p \circ (q, r)$ and recall basic properties of the logarithm function.

Quite crucially, Equation (4) is just one example of a more general rule. An analogous equation holds whenever *any* probability distribution $p = (p_1, p_2, \dots, p_n)$ is combined with n other probability distributions, as in Figure 9, which will now be denoted with superscripts: $q^1, q^2, \dots, q^n$. (Pay careful attention to the difference between subscripts and superscripts: p_1 is a number, whereas q^1 is a list of numbers.) As in our motivating example above, the probability distributions that are composed with p may be of different lengths. For example, q^1 might be an element of Δ_3, while q^2 might be an element of Δ_{17}, while q^3 might be an element of Δ_5, and so on. In general, it can be shown that the entropy of the composite distribution $p \circ (q^1, q^2, \dots, q^n)$ satisfies the following important equation, which is sometimes called the **chain rule for entropy** [Lei21, Proposition 2.2.8]:

$$H(p \circ (q^1, q^2, \dots, q^n)) = H(p) + p_1 H(q^1) + p_2 H(q^2) + \cdots + p_n H(q^n). \tag{5}$$

The chain rule is important because entropy is essentially the *only* collection of continuous functions that satisfies it. In other words, continuity and the chain rule are at the heart of entropy; they are enough to distinguish it from all other functions that assign real numbers to probability distributions. In mathematical parlance, entropy is said to be "uniquely characterized" by the chain rule. This is a theorem, and a proof of it may be found in a recent book by Leinster [Lei21, Theorem 2.5.1].

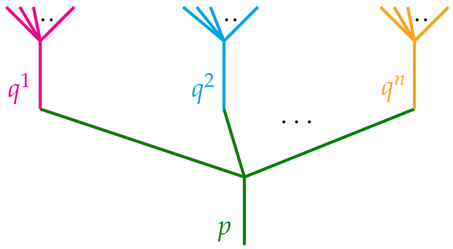


Figure 9: A picture for the composite probability distribution $p \circ (q^1, q^2, \dots, q^n)$.

▶ *In more detail.* This characterization may be stated more formally as follows: if $\{F \colon \Delta_n \to \mathbb{R}\}$ is any other collection of continuous functions satisfying Equation (5), then F must in fact be *equal* to H or to some multiple of it. In symbols this means $F = cH$ for some real number c. As mentioned above, this important theorem appears in [Lei21] and is actually a slight variation of a closely related theorem about entropy that Russian mathematician Dmitry Faddeev proved more than six decades ago [Fad56]. Here is the precise statement of Leinster's version of Faddeev's theorem:

Theorem 1. *Let $\{F\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ be a sequence of functions. The following are equivalent:*

i. The functions F are continuous and satisfy the chain rule.

ii. $F = cH$ for some real number c.

By now it might seem we are drowning in a sea of symbols. Perhaps our mathematical lungs are gasping for air. "Why does the chain rule hold? What is going on here?" These are reasonable questions. The first is easily resolved. Why is Equation (5) true? As claimed above, it can be verified using simple math. Walking through those calculations here, however, would cause us to sink deeper into the symbolic sea. But the details are found in [Lei21, Proposition 2.2.8] whose surrounding text also contains a more formal discussion of the mathematics in this section. Our second question is much more interesting. What is going on here? What does the chain rule in Equation (5) *mean*? To start, it tells us that entropy interacts with the operadic composition $\circ$ of probability distributions in a very principled manner. Curiosity drives us to wonder whether a similar kind of interaction arises elsewhere in the mathematical landscape and, if so, *where*. On the surface this may appear to be a grandiose task. Where would we even begin to explore such a question?

A good place to begin is at the beginning. Suppose $n = 1$. Then the chain rule takes on the form $H(p \circ q) = H(p) + H(q)$ for probability distributions p and q. In words, this equation seems[5] to say that entropy H behaves nicely with composition $\circ$ of probability distributions and with addition $+$ of real numbers. Functions with this kind of behavior may be familiar to those who have taken a course in abstract algebra, where such functions are called homomorphisms. To elaborate, in abstract algebra we often begin with a set X together with some way to combine or "multiply" elements in that set. This idea was introduced previously in this article, but let us unwind it further. Suppose 🟢 and 🔴 are any two elements of X. These dots may represent numbers such as 2 and 3, or they may represent English words such as *yellow* and *banana*. We will not specify what the elements of X are because it does not matter. Supposing we can combine or multiply elements of X, we will use juxtaposition 🟢🔴 to denote this operation. This could represent multiplication of real numbers, or concatenation of English words, or something entirely different. The abstraction allows us to represent all possibilities simultaneously. Similarly, suppose Y is another set with its own operation, which we will also denote by juxtaposition. Then a function $f\colon X \to Y$ is called a **homomorphism** if combining elements in X and then applying f is the same as first applying f and then combining the elements in Y; that is, if the function satisfies the equation $f($🟢🔴$) = f($🟢$)f($🔴$)$ for all elements 🟢 and 🔴 in X. When comparing this with the chain rule for entropy when $n = 1$, it may seem that entropy is a homomorphism. The equation $H(p \circ q) = H(p) + H(q)$ is indeed analogous to the newly

[5] When $n = 1$ the only possible choice for p is the trivial probability distribution (1), which necessarily implies that $p \circ q = q$ and $H(p) = 0$ and so the chain rule reduces to $H(q) = H(q)$, which is uninteresting. But we will momentarily ignore this for the sake of exposition.

introduced equation $f(🟢🔴) = f(🟢)f(🔴)$. So if we were asked to decipher the mathematical message in the chain rule, may we thus confidently assert, "Entropy is a homomorphism"? Alas, the answer is no. Baez shared this conundrum in his 2011 article [Bae11]:

> While [the chain rule] is cute, it's a bit tricky to find its proper place in the world of abstract algebra.... Shannon entropy gives a map H from probability distributions to numbers. So, if you're algebraically inclined, you would want H to be a homomorphism.... We see laws of this sort all over math. But the true law has an extra term. What's going on?

To see where the problem lies, let us progress from $n = 1$ to $n = 2$. In that case, a probability distribution $p = (p_1, p_2)$ may be combined with two other probability distributions q^1 and q^2, and the chain rule then becomes[6]

$$H(p \circ (q^1, q^2)) = H(p) + p_1 H(q^1) + p_2 H(q^2). \tag{6}$$

This makes it clear that entropy is not a homomorphism, as the ps have intermingled with the qs on the right-hand side. Alternatively, one might wish that the right-hand side of Equation (6) did *not* include the term $H(p)$, for in its absence a small trick could be applied to make the equation look more like a homomorphism. (This is what Baez meant by "the true law has an extra term." The details of the trick are given in his article.) So, it seems we are back to where we started. If entropy is not a homomorphism, then what is it? To gain the clarity we seek, we must strip away the details. It will help to squint our eyes while looking at the chain rule in Equation (6) and ignore most of the symbols. Forget the subscripts and superscripts. Forget the ps and qs. To see what is *really* going on with entropy, imagine that p is a green dot 🟢 and the qs collectively are a pink dot 🔴. Then Equation (6) roughly looks like something of the following form:

$$H(🟢🔴) = H(🟢) + 🟢H(🔴). \tag{7}$$

In words, this says that the entropy of two things 🟢🔴 is equal to the entropy of the first thing 🟢 *plus* the first thing 🟢 *mulitplied by* the entropy of the second thing 🔴. Does this sound familiar? Perhaps not. Regardless, Equation (7) is undeniably asymmetric. It looks off, visually speaking. On the right-hand side of the equals sign there are two green dots but only *one* pink dot. That asymmetry is somewhat irksome, like having a pebble stuck in one's shoe. The formula would appear more balanced if there were an extra pink dot on the right, like so:

$$H(🟢🔴) = H(🟢)🔴 + 🟢H(🔴). \tag{8}$$

Now, does *this* look more familiar? Students of calculus may indeed recognize the equation above. It is reminiscent of a famous formula known as the "product rule" or the **Leibniz rule**, which is standard material in a first course on calculus.

[6] Notice this equation coincides with Equation (4) when $p = \left(\frac{1}{2}, \frac{1}{2}\right)$ corresponds to a coin toss and when $q^1 = q$ and $q^2 = r$ correspond to our breakfast and dinner choices in the example in the previous section.

▶ *In more detail.* In calculus, the product rule is a formula for the derivative of a product of functions. To elaborate, the derivative of a differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ (that is, a function whose derivative exists) is another function often denoted by f' or by $\frac{df}{dx}$ or by $d(f)$. Given two differentiable functions f and g, it is natural to ask about the derivative of their product. Can the derivative $(fg)'$ be expressed in terms of the individual functions f' and g'? The affirmative answer is given by the famous product rule, which says that the derivative $(fg)'$ is the function whose value at a point x is given by $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$. In words, the derivative of fg is obtained by multiplying g by the derivative of f, then multiplying f by the derivative of g, and then taking the sum of these two functions. This may be written more succinctly as $d(fg) = d(f)g + fd(g)$, which should be compared with Equation (8).

But calculus is not the only place in mathematics where a version of the Leibniz rule appears. Many other functions may satisfy an analogous equation,[7] and such functions are given a name: *derivations*. Elaborating, suppose we have any set of elements ${\color{green}\bullet}, {\color{magenta}\bullet}, \ldots$ together with some notion of "multiplication" between them so that we may make sense of expressions such as ${\color{green}\bullet}{\color{magenta}\bullet}$. Perhaps these dots are numbers, but perhaps they are something else. The abstraction is once again intentional. Then, informally speaking, a **derivation** is defined to be any function d on this set that satisfies the Leibniz rule $d({\color{green}\bullet}{\color{magenta}\bullet}) = d({\color{green}\bullet}){\color{magenta}\bullet} + {\color{green}\bullet}d({\color{magenta}\bullet})$ for all elements ${\color{green}\bullet}$ and ${\color{magenta}\bullet}$. As one might expect, this is a loose explanation of a much more formal definition, but the takeaway is that derivations are a staple in the world of advanced mathematics. A small digression may help to illuminate this claim. In our opening discussion on topological simplices, we briefly mentioned "homology," which is a mathematical framework that assigns algebraic objects called *homology groups* to a topological space. Those groups encode valuable information about the topological space and are often easier to work with than the space itself. A similar story holds if the words "topological space" are replaced with "associative algebra." There, the analogous construction is called the *Hochschild cohomology* of the algebra, and derivations of the algebra play a vital role: they are what are known as "cocycles of degree 1" [Wit19, Chapter 1]. Lingo aside, the takeaway is that derivations are functions that generalize the Leibniz rule from calculus, and Equations (7) and (8) hint at a tantalizing connection between derivations and entropy.

▶ *In more detail.* As suggested above, derivations are functions defined with respect to some algebraic structure. That is, one must work in some kind of setting where "multiplication" makes sense. Given an algebra A, for instance, a derivation on A is formally defined to be a

[7] As an example, here is a simple exercise. Let $[0,1]$ denote the set of all numbers between and including 0 and 1, and define a function $d\colon [0,1] \to \mathbb{R}$ by declaring $d(x) = -x\log(x)$ if $x > 0$ and $d(x) = 0$ if $x = 0$. Show that d satisfies the Leibniz rule. That is, show that $d(xy) = d(x)y + xd(y)$ for all numbers x and y in $[0,1]$. This computation only requires arithmetic and basic properties of the logarithm function.

linear function $d\colon A \to A$ satisfying the Leibniz rule $d(ab) = d(a)b + ad(b)$ for all elements a and b in A. This can be generalized slightly by replacing the target A with another object M called a "bimodule over A" and instead considering functions $d\colon A \to M$ satisfying the same equation. These, too, are called derivations. The only difference now is that $d(a)$ is an element of M, which is allowed to be different than A. So, some care is needed here. If M and A are genuinely different from one another, then it is not at all obvious how to make sense of the expressions $d(a)b$ and $ad(b)$. Here, $d(a)$ and b are elements of different sets, namely M and A, respectively, and we have not said what it means to multiply elements that are not members of the same set. By way of analogy, multiplication of two numbers $0.5 \times 8 = 4$ makes sense, but what would it mean to multiply a number with an English word, $0.5 \times \textit{yellow} = ?$ This is the nature of the question we are faced with here, and a similar thought holds for a and $d(b)$. Bimodules are the answer to such questions. A **bimodule over** A is a mathematical object M that is equipped with a way to "multiply" its elements by elements from A. Readers familiar with linear algebra have seen this idea before. It generalizes what is known as scalar multiplication. The ability to multiply a vector $\mathbf{v}$ by a real number k to obtain a new vector $k\mathbf{v}$ is precisely the statement that a *real vector space is a bimodule over the real numbers.* In this analogy a and b are "scalars" and $d(a)$ and $d(b)$ are "vectors."

In short, derivations are functions that interact with algebraic structure in a precise way known as the Leibniz rule. Moreover, the rough form of the chain rule in Equation (7) suggests that entropy behaves somewhat like a derivation. The similarity is indeed hard to miss:

$$\begin{array}{cc} \text{Leibniz rule:} & \text{entropy:} \\ d({\color{green}\bullet}{\color{magenta}\bullet}) = d({\color{green}\bullet}){\color{magenta}\bullet} + {\color{green}\bullet}d({\color{magenta}\bullet}) & H({\color{green}\bullet}{\color{magenta}\bullet}) = H({\color{green}\bullet}) + {\color{green}\bullet}H({\color{magenta}\bullet}) \end{array} \tag{9}$$

We are now in a position to ask the obvious question: *Is there a real sense in which entropy is a derivation?* Baez posed this very puzzle in his 2011 article, where he wondered how the similarity between entropy and derivations might be reconciled with some of Leinster's work on the operad of topological simplices: "So an interesting question presents itself: How does the 'derivation' way of thinking about the [chain rule] relate to Tom Leinster's interpretation of it...?" [Bae11] Ten years later, my work in [Bra21] gave an answer to this question.

THE ANSWER IS that there is a correspondence—that is, a way to go back and forth—between Shannon entropy and **derivations of the operad of topological simplices**. The latter expression is a brand new generalization of the Leibniz rule in the context of operads, and a formal definition is one of the contributions of [Bra21]. The definition draws inspiration from the familiar concept of a derivation from abstract algebra, yet it is notably different. A derivation of an operad turns out to consist not of a *single* function, but rather of *infinitely* many functions. Happily, we have made a similar shift in perspective before. It is analogous to our understanding that entropy is not merely a number, nor is it merely a function, but rather it is a collection of infinitely many continuous functions $\{H\colon \Delta_n \to \mathbb{R}\}$ satisfying a certain equation—the chain rule. Analogously, a derivation of the operad of simplices is defined to be a collection of infinitely many continuous functions $\{d\colon \Delta_n \to \blacksquare\}$ satisfying a certain equation—the Leibniz rule. The black box is a temporary place-holder for a new kind of output that we explain now. Recall that entropy assigns numbers to probability distributions. Our derivation, on the other hand, will assign *functions* to probability distributions. Explicitly, a derivation d of the operad of topological simplices assigns a *continuous function* $d(p)\colon \mathbb{R}^n \to \mathbb{R}$ to each probability distribution p in Δ_n. We have seen an example of such an assignment already in a "representation of an operad" mentioned in our introduction to operads. Indeed, we noted previously that, because operads are abstract, it is better to work with concrete representations of them in practice. Our new version of a derivation takes this to heart. So to summarize, the first step in making the connection between entropy and derivations is to represent each probability distribution p in Δ_n by a continuous function $d(p)\colon \mathbb{R}^n \to \mathbb{R}$. And this gives a clue to the black box above. It is precisely the set of all continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$. This set can further be made into a topological space, albeit one that is harder to visualize than simplices.[8] Even so, let $\hom(\mathbb{R}^n, \mathbb{R})$ denote this topological space of functions. Then we can summarize this discussion with the following informal definition.

Definition (Informal)**.** *A* ***derivation of the operad of topological simplices*** *is a collection of continuous functions* $\{d\colon \Delta_n \to \hom(\mathbb{R}^n, \mathbb{R})\}$ *that satisfies an appropriate version of the Leibniz rule, "*$d(p \circ q) = d(p) \circ q + q \circ d(q)$*" for any probability distributions* p *and* q*.*

Of course, care must be taken to explain the desired Leibniz rule in the scare quotes above. The expression $p \circ q$ does not make much sense, for instance. So far we have only used the symbol $\circ$ to denote the composition of *multiple* probability distributions with an arbitrary p in Δ_n, as in the tree-grafting picture in Figure 9. To have an appropriate version of the Leibniz rule, however, we need only compose a *single* probability distribution with p. But this problem is no problem at all. A single probability distribution q may indeed be composed with p, and we have already done so in the example shown in Figure 6. That is, we can simply graft the root of some q onto any *one* of the leaves of some p. The definition asks that an appropriate version of the Leibniz rule holds for each of those ways.

To make sense of the Leibniz rule in the context of the operad of topological simplices, we also need to make sense of the expressions $d(p) \circ q$ and $q \circ d(q)$ that appear in the formula. Notably, both expressions involve the combination of two things that are not the same; p is a probability distribution, whereas $d(q)$ is a function. What does it mean to combine the two? We have been faced with this question once before. What kind of mathematical structure

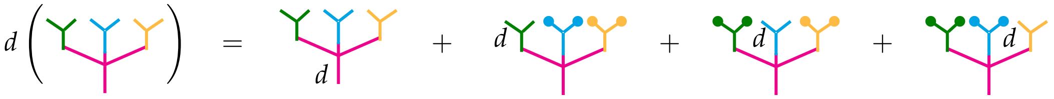
[8] Formally speaking, we can equip the set of continuous functions $\mathbb{R}^n \to \mathbb{R}$ with what is known as the "product topology," which turns it into a topological space that is easy to work with in this setting.

enables one to "multiply" elements from different sets in a meaningful way? As discussed above, the answer lies in a *bimodule structure*. Pinning down such details is another contribution of [Bra21]. The paper gives a formal definition of a "bimodule over an operad," and it shows that the collection of topological spaces $\hom(\mathbb{R}^n, \mathbb{R})$ for each $n = 1, 2, 3, \ldots$ admits such a structure. Curiously enough, *this* part of the mathematics explains the reason that entropy appears to be missing a pink dot when compared with the traditional Leibniz rule in (9). See [Bra21, Example 3] for more details. So, a part of the mystery has now been solved.

With these definitions in hand, the rest of the mathematics falls into place as well. First, it can be shown that every derivation of the operad of topological simplices satisfies a version of the chain rule. This upgraded rule looks roughly like the following:

$$d(p \circ (q^1, q^2, \ldots, q^n)) \overset{\text{sort of}}{=} d(p) + p_1 d(q^1) + p_2 d(q^2) + \cdots + p_n d(q^n),$$

which is analogous to the original chain rule in Equation (5). The "sort of" hovering over the equals sign means interested readers are encouraged to take a look at the *true* equation, which is given in [Bra21, Proposition 1]. Either way, in our graphical notation this new version of the chain rule essentially says that the function $d(p \circ (q^1, q^2, \ldots, q^n))$ is obtained by applying d to each of the trees representing the individual probability distributions. Below is a picture of this rule in the case when $n = 3$. The "dots" on the leaves can be ignored—they are part of the bimodule structure, whose explanation we omit.



Finally, with the proper definition of "derivation" in place, the main theorem of [Bra21]—and the climax of our discussion—follows immediately. One can show there is a way to go back and forth between Shannon entropy and derivations of the operad of topological simplices. More specifically, we can always use Shannon entropy to define a derivation and, conversely, every derivation knows about Shannon entropy. Here is the formal statement of the theorem:

Theorem 2 (Bradley, 2021)**.** *Shannon entropy defines a derivation of the operad of topological simplices, and for every derivation of this operad, there exists a point at which it is given by a constant multiple of Shannon entropy.*

► *In more detail.* The statement of the theorem does not tell the reader exactly *how* the correspondence works, so let us provide a few more details for those who are curious. One direction is quite easy. To show that Shannon entropy defines a derivation, we need to use entropy to construct a collection of continuous functions $\{d \colon \Delta_n \to \hom(\mathbb{R}^n, \mathbb{R})\}$ and verify that it satisfies the appropriate version of the Leibniz rule hinted at above. Here is how to construct such a collection: for each natural number n and for each probability distribution p in Δ_n, define the function $d(p)\colon \mathbb{R}^n \to \mathbb{R}$ to be constant at entropy; that is, define $d(p)(x) = H(p)$ for all points x in $\mathbb{R}^n$. The proof that this defines a derivation is straightforward, requiring nothing more than arithmetic. The other direction of the correspondence is slightly more involved. It says that if $\{d\colon \Delta_n \to \hom(\mathbb{R}^n, \mathbb{R})\}$ is any derivation, then for each natural number n there exists a point x in $\mathbb{R}^n$ so that $d(p)(x) = cH(p)$ for some real number c. The proof of this part of the theorem uses the result of Leinster-Faddeev mentioned previously in Theorem 1, which is the reason that both theorems involve a *constant multiple* of entropy. What's more, one can easily show that the special point x is actually zero, that is, $d(p)(0) = cH(p)$.

So, the mystery surrounding entropy and derivations is now solved. *Or is it?* We have reached the end of this article, but now there are new mysteries to explore. For instance, at the time of writing, it is not known whether there is a meaningful interpretation of derivations evaluated at other points besides zero, or whether the operad has other derivations besides the one that is constant at entropy. Further, it turns out that the original chain rule for entropy in Equation (5)—the very equation that caught the attention of Baez and Leinster [Bae11, Lei21]—is merely a corollary to the chain rule for derivations and the main theorem above. So, perhaps the chain rule is just a shadow of something more. Our "final layer" of entropy is almost certainly not the final layer at all.

Venturing Into New Mathematics

In closing, let us revisit a question asked towards the beginning of this article: "What does it mean to discover new mathematics?" We have now seen an in-depth example. We began with a survey of the landscape of higher mathematics and a basic introduction to Shannon entropy. We then began to peel back the layers of entropy one by one, culminating in the intriguing chain rule. Curiosity compelled us to ask where such a rule fit into the mathematical landscape, and this led us to pursue an interesting connection between entropy and derivations. Generally speaking, one way to approach such mysteries is to search the mathematical literature to see if the mystery has already been solved. If no such solution exists, then the mathematician is prompted to forge ahead. That was indeed the case for us. And that discovery—namely, a new way to think about entropy from a pure mathematical perspective—is the content of [Bra21].

But why is this new perspective worth sharing? Recall that information theory traditionally has little to do with either abstract algebra or topology, as each subject seemingly resides in separate, distant sectors of the scientific and mathematical landscapes. But now we have seen in great detail that entropy, algebra, and topology are intricately intertwined with one another. Recent events also indicate that the work in [Bra21] is merely one of several related connections between entropy and higher mathematics found in the past several years

[BB15, EVG15, Mai19, Bae11]. Perhaps these timely discoveries are a clue that there is more interesting, more fundamental mathematics waiting to be discovered. And because entropy is at the heart of it all, it is particularly intriguing to wonder about the new insights that such mathematics could lend to the study of physics and the natural world. In the words of German physicist Max Plank quoted earlier: "...the external world represents something independent of us, something absolute which we confront, and the search for the laws valid for this absolute appeared to me the most beautiful scientific task in life." The beauty of such a task is especially enriched when the explorer has sure confidence that the answers may indeed be found.

Colossians 1:16–17

Bibliography

[Ada04] Colin C. Adams. *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots*. American Mathematical Society, Providence, RI, 2004.

[Bae11] John C. Baez. Entropy as a functor. *nLab*, 2011. Blog post. Available online: https://www.ncatlab.org/johnbaez/show/Entropy+as+a+functor (accessed 30 December 2021).

[BB15] Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17(5):3253–3318, 2015. https://doi.org/10.3390/e17053253.

[Bek03] Jacob D. Bekenstein. Information in the holographic universe. *Scientific American*, 289(2):61, 2003.

[Bra17a] Tai-Danae Bradley. What is an operad? part 1. *Math3ma*, 2017. Blog post. Available online: https://www.math3ma.com/blog/what-is-an-operad-part-1 (accessed 30 December 2021).

[Bra17b] Tai-Danae Bradley. What is an operad? part 2. *Math3ma*, 2017. Blog post. Available online: https://www.math3ma.com/blog/what-is-an-operad-part-2 (accessed 30 December 2021).

[Bra18] Tai-Danae Bradley. An invitation to category theory. *Chalkdust Magazine*, 08:22–25, 2018. Available online: http://chalkdustmagazine.com/features/an-invitation-to-category-theory/.

[Bra21] Tai-Danae Bradley. Entropy as a topological operad derivation. *Entropy*, 23(9), 2021. https://doi.org/10.3390/e23091195.

[Cha21] Matthew Chalmers. Witten reflects. *CERN Courier*, 2021. Available online: https://cerncourier.com/a/witten-reflects/ (accessed 22 December 2021).

[CVJ21] Gunnar Carlsson and Mikael Vejdemo-Johansson. *Topological Data Analysis with Applications*. Cambridge University Press, Cabridge, UK, 2021.

[Dys09] Freeman Dyson. Birds and frogs. *Notices of the AMS*, 56(2):212–223, 2009.

[EVG15] Philippe Elbaz-Vincent and Herbet Gangl. Finite polylogarithms, their multiple analogues and the Shannon entropy. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information. GSI 2015.*, volume 9389 of *Lecture Notes in Computer Science*, pages 277–285. Springer, Cham., 2015. https://doi.org/10.1007/978-3-319-25040-3_31.

[Fad56] Dmitry K. Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk*, 11:227–231, 1956. (In Russian).

[Gan26] Solomon Gandz. The origin of the term "algebra". *The American Mathematical Monthly*, 33(9):437–440, 1926.

[GBGL08] Timothy Gowers, June Barrow-Green, and Imre Leader, editors. *The Princeton Companion to Mathematics*. Princeton University Press, Princeton, NJ, 2008.

[Ghr14] Robert Ghrist. *Elementary Applied Topology*. Createspace, 1st edition, 2014.

[HE21] Kelsey Houston-Edwards. How mathematicians use homology to make sense of topology. *Quanta Magazine*, 2021. Available online: https://www.quantamagazine.org/how-mathematicians-use-homology-to-make-sense-of-topology-20210511/ (accessed 22 December 2021).

[Lei] Tom Leinster. The categorical origins of entropy. Talk at the Geometrical and Topological Structures of Information Conference at CIRM, 2017. Video online: https://www.youtube.com/watch?v=JgNy2ZUqdZI (accessed 22 December 2021).

[Lei21] Tom Leinster. *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press, Cambridge, UK, 2021.

[Mai19] Tom Mainiero. Homological tools for the quantum mechanic. 2019. arXiv preprint: arXiv:1901.02011.

[Maz07] Barry Mazur. When is one thing equal to some other thing? 2007. Available online: https://people.math.harvard.edu/~mazur/preprints/when_is_one.pdf (Accessed 30 December 2021).

[MSS02] Martin Markl, Steven Shnider, and Jim Stasheff. *Operads in Algebra, Topology and Physics*. Mathematical surveys and monographs. American Mathematical Society, Providence, RI, 2002.

[Nic01] James Nickel. *Mathematics: Is God Silent?* Ross House Books, Vallecito, CA, 2001.

[Pla48] Max Planck. Wissenschaftliche selbstbiographie. *Physikalische Abhandlungen*, 3:374, 1948.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[Smo01] Lee Smolin. *Three Roads to Quantum Gravity*. Basic Books, New York, NY, 2001.

[Sta04] Jim Stasheff. What is... an operad? *Notices Amer. Math. Soc.*, 51:630–631, 2004.

[Wee20] Jeffrey R. Weeks. *The Shape of Space*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, 2020.

[Wit19] Sarah J. Witherspoon. *Hochschild Cohomology for Algebras*. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2019.