

ENTROPY AS A TOPOLOGICAL OPERAD DERIVATION

TAI-DANAE BRADLEY

ABSTRACT. We share a small connection between information theory, algebra, and topology—namely, a correspondence between Shannon entropy and derivations of the operad of topological simplices. We begin with a brief review of operads and their representations with topological simplices and the real line as the main example. We then give a general definition for a derivation of an operad in any category with values in an abelian bimodule over the operad. The main result is that Shannon entropy defines a derivation of the operad of topological simplices, and that for every derivation of this operad there exists a point at which it is given by a constant multiple of Shannon entropy. We show this is compatible with, and relies heavily on, a well-known characterization of entropy given by Faddeev in 1956 and a recent variation given by Leinster.

1. INTRODUCTION

In this article, we describe a simple connection between information theory, algebra, and topology. To motivate the idea, consider the function $d: [0, 1] \rightarrow \mathbb{R}$ defined by

$$d(x) = \begin{cases} -x \log x & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$$

This map satisfies an equation reminiscent of the Leibniz rule from Calculus, $d(xy) = d(x)y + xd(y)$ for all $x, y \in [0, 1]$. In other words, d is a nonlinear derivation [Lei21], (Lemma 2.2.6). This derivation may also bring to mind the Shannon entropy of a probability distribution. Indeed, a probability distribution on a finite set $\{1, \dots, n\}$ for $n \geq 1$ is a tuple of nonnegative real numbers $p = (p_1, \dots, p_n)$ satisfying $\sum_{i=1}^n p_i = 1$, and the **Shannon entropy** of p is defined to be

$$H(p) = - \sum_{i=1}^n p_i \log p_i = \sum_{i=1}^n d(p_i).$$

Although d is not linear, this may prompt one to wonder about settings in which Shannon entropy itself is a derivation. We describe one such setting below by showing a correspondence between Shannon entropy and derivations of the operad of topological simplices.

1.1. Motivation. As evidenced by recent work, the intersection of information theory and algebraic topology is fertile ground. In 2015 tools of information cohomology were introduced in [BB15] by Baudot and Bennequin who construct a certain cochain complex for which entropy represents the unique cocycle in degree

SANDBOX@ALPHABET, MOUNTAIN VIEW, CA 94043
E-mail address: tai.danae@math3ma.com.

1. In the same year, Elbaz-Vincent and Gangl approached entropy from an algebraic perspective and showed that what are known as information functions of degree 1 behave “a lot like certain derivations” [EVG15]. A few years prior in 2011, Baez, Fritz, and Leinster gave a category theoretical characterization of entropy in [BFL11], which was recently extended to the quantum setting by Parzygnat in [Par20]. In preparation of that 2011 result, Baez remarked in the informal article [Bae11] that entropy appears to behave similarly to a derivation in a certain operadic context, an observation we verify and make explicit below. Cohomological ideas are also explored in Mainiero’s recent work, where entropy is found to appear in the Euler characteristic of a particular cochain complex associated to a quantum state [Mai19]. Upon taking inventory, one thus has the sense that entropy behaves somewhat similar to “ d of something,” for some (co)boundary-like operator d . The present article is in this same vein. Notably, once a few simple definitions are in place, the mathematics is quite straightforward. Even so, we feel it is worth sharing if for no other reason than to provide a glimpse at yet another algebraic and topological facet of entropy.

1.2. Background. To start, our work is based on a particular characterization of Shannon entropy that is compatible with an operadic viewpoint. Let Δ^n denote the standard topological n -simplex for $n \geq 0$,

$$\Delta^n := \{(p_0, p_1, \dots, p_n) \in \mathbb{R}^{n+1} \mid 0 \leq p_i \leq 1 \text{ and } \sum_{i=0}^n p_i = 1\},$$

where Δ^0 denotes the unique probability distribution on the one-point set. More generally, any probability distribution $p = (p_0, \dots, p_n)$ on an $n + 1$ -element set is a point in Δ^n . Given $n + 1$ probability distributions $q^i = (q_0^i, \dots, q_{k_i}^i) \in \Delta^{k_i}$ where $i = 0, 1, \dots, n$, they may be composed with p simultaneously to obtain a point in $\Delta^{k_0+k_1+\dots+k_n+n}$ denoted by

$$p \circ (q^0, q^1, \dots, q^n) := (p_0 q_0^0, \dots, p_0 q_{k_0}^0, p_1 q_1^1, \dots, p_1 q_{k_1}^1, \dots, p_n q_n^n, \dots, p_n q_{k_n}^n).$$

As shown in [Lei21] and reviewed below, this composition of probabilities finds a natural home in the language of operads. Furthermore, it plays a key role in a well-known 1956 characterization of Shannon entropy due to D. K. Faddeev [Fad56]. A proof of a slight variation of Faddeev’s result was recently given by Leinster [Lei21], (Theorem 2.5.1). That is the version we quote here.

Theorem 1 (Faddeev-Leinster). *Let $\{F: \Delta^n \rightarrow \mathbb{R}\}_{n \geq 0}$ be a sequence of functions. The following are equivalent:*

(1) *the functions F are continuous and satisfy*

$$(1) \quad F(p \circ (q^0, \dots, q^n)) = F(p) + \sum_{i=0}^n p_i F(q^i)$$

where $n \geq 0$ and $p \in \Delta^n$ and $q^i \in \Delta^{k_i}$ with $k_0, k_1, \dots, k_n \geq 0$;

(2) *$F = cH$ for some $c \in \mathbb{R}$.*

To make the connection with derivations, let us introduce some notation. Given a probability distribution $p \in \Delta^n$ let $\bar{p}: \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ denote the function that maps a point $x = (x_0, \dots, x_n)$ to the standard inner product $\langle p, x \rangle = \sum_{i=0}^n p_i x_i$. Then, when $F = H$, Equation (1) may be rewritten as

$$(2) \quad H(p \circ (q^0, \dots, q^n)) = H(p) + \bar{p}(H(q^0), \dots, H(q^n)).$$

This equation is one hint that entropy might be a derivation, although a “ q ” is notably absent from the first term on the right-hand side. As a further teaser, Baez explored an algebraic interpretation of Equation (2) in the informal article [Bae11], where the reader is reminded that Shannon entropy is a derivative of the partition function of a probability distribution with respect to Boltzmann’s constant, considered as a formal parameter. In that article, Equation (2) follows in a few short lines from this computation. One is thus motivated to look for a general framework of operad derivations for which Equation (2) is an example. This is what we describe below.

Section 3 reviews the definition of operads and representations of them. We will recall that the collection of topological simplices admits the structure of an operad as in [Lei21] and that \mathbb{R} gives rise to a representation of it. In Section 4, we define an abelian bimodule M over any operad \mathcal{O} and the notion of a derivation of \mathcal{O} with values in M . With these definitions in place, Equation (2) will find a generalization in Proposition 1, and the main result will quickly follow.

Theorem. *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.*

2. ACKNOWLEDGEMENTS

I thank Darij Grinberg, Joey Hirsh, Tom Leinster, Jim Stasheff, and John Terilla for helpful discussions as well as the anonymous referees for their insightful feedback.

3. BACKGROUND: OPERADS AND THEIR REPRESENTATIONS

In an introduction to operads, it is helpful to first think about algebras. An algebra A is a vector space V equipped with a bilinear map $\mu: V \times V \rightarrow V$ thought of as multiplication. Depending on whether μ satisfies a particular relation, the algebra will usually be described by an appropriate qualifier. For instance, if $\mu(v, w) = \mu(w, v)$ for all $v, w \in V$, then A is called a *commutative algebra*; if $\mu(\mu(u, v), w) = \mu(u, \mu(v, w))$ for all $u, v, w \in V$, then A is called an *associative algebra*, and so on. Behind each of these algebras is a particular operad that encodes the behavior of the multiplication map μ . To motivate the formal definition, it is helpful to visualize μ as a planar binary rooted tree and more generally to imagine an arbitrary n -ary operation as a planar rooted tree with n leaves. There is a natural way to compose such operations. For instance, when f is a 3-ary operation and g is a 4-ary operation, they may be composed to obtain a 6-ary operation by using the output of g as one of the inputs of f as illustrated in Figure 1. There g has been grafted into the *second* leaf of the tree associated to f , and so we denote that choice with the subscript “ \circ_2 ” in the figure. There are two other composites $f \circ_1 g$ and $f \circ_3 g$, which are not shown but are obtained similarly.

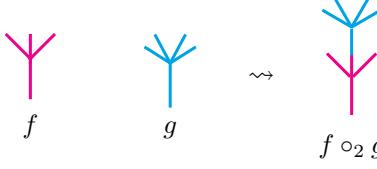


FIGURE 1. One of the three ways to compose a 4-ary operation g with a 3-ary operation f .

In general, there are n ways to compose an m -ary operation with an n -ary operation, and the resulting operation will always have arity $m + n - 1$. This composition should further satisfy some sensible associativity and unital axioms, and the collection of all such operations with their compositions is called an **operad**. The concept has origins in category theory [Lam69] and has been used extensively in algebraic topology and homotopy theory [May72, BV73, LV12, Val12, Sta04] with applications in physics as well [Mar96, MSS02]. Operads may be defined in any symmetric monoidal category, and for ease of exposition below, we will assume all categories \mathbf{C} are concrete (that is, all objects have underlying sets) so that we may refer to *elements* in a given object of \mathbf{C} . Indeed, the main example to have in mind is the category of topological spaces.

Definition 1. Let \mathbf{C} be a symmetric monoidal category with monoidal product \otimes . An **operad** in \mathbf{C} consists of a sequence of objects $\{\mathcal{O}(1), \mathcal{O}(2), \dots\}$ together with morphisms

$$\circ_i: \mathcal{O}(n) \otimes \mathcal{O}(m) \rightarrow \mathcal{O}(n + m - 1)$$

in \mathbf{C} for all $n, m \geq 1$ and $1 \leq i \leq n$ and an operation $1 \in \mathcal{O}(1)$ satisfying the following:

- (i) [associativity] For all $p \in \mathcal{O}(n)$ and $q \in \mathcal{O}(m)$ and $r \in \mathcal{O}(k)$,

$$(p \circ_j q) \circ_i r = \begin{cases} (p \circ_i r) \circ_{j+k-1} q & \text{if } 1 \leq i \leq j-1 \\ p \circ_j (q \circ_{i-j+1} r) & \text{if } j \leq i \leq j+m-1 \\ (p \circ_{i-m+1} r) \circ_j q & \text{if } i \geq j+m \end{cases}$$

- (ii) [identity] The operation $1 \in \mathcal{O}(1)$ acts as an identity in the sense that

$$1 \circ_1 p = p \circ_i 1 = p$$

for all $p \in \mathcal{O}(n)$ and $1 \leq i \leq n$.

The definition is conceptually simple despite its cumbersome appearance. For instance, Figure 2 illustrates the associativity requirements listed in item (i).

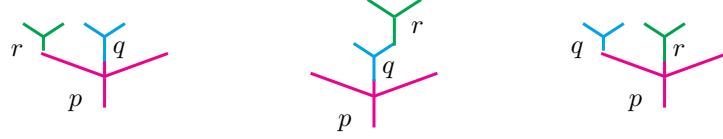


FIGURE 2. Associativity in an operad. Left) First composing q with p and then r is the same as first composing r with p and then q . The order in which this is performed does not matter. (Right) The same is true if r appears to the right, rather than the left, of q . (Middle) Likewise, r may first be composed with q and their composite may then be composed with p , or q may be first composed with p followed by r . Again, the order does not matter.

As mentioned above, one often thinks of the elements $\mathcal{O}(n)$ as abstract n -to-1 operations, and the morphisms \circ_i specify a way to compose them. It is common to begin indexing the sequence of objects at $n = 0$ to account for 0-ary operations, but as we will soon see, our main example of an operad in Example 2 will have no 0-ary operations, and so our definition starts with $\mathcal{O}(1)$. We do not consider an action of the symmetric group and so \mathcal{O} is sometimes called a *non-symmetric operad*, but we will simply call it an operad. In the special case when C is the category of vector spaces with linear maps and \otimes is the tensor product, \mathcal{O} is often called a *linear operad*. When it is the category Top of topological spaces with continuous maps and \otimes is the Cartesian product, \mathcal{O} is often called a *topological operad*.

Example 1. Given a set X , the **endomorphism operad** is $\text{End}_X = \{\text{End}_X(1), \text{End}_X(2), \dots\}$ where $\text{End}_X(n) := C(X^n, X)$ denotes the set of all functions from the n -fold Cartesian product X^n to X . The unit operation in $\text{End}_X(1)$ is the identity function $\text{id}_X : X \rightarrow X$. If $f \in C(X^n, X)$ and $g \in C(X^m, X)$ are a pair of functions, then for each $i = 1, \dots, n$ the composition $f \circ_i g$ is obtained by using the output of g as the i th input of f . Explicitly, given $(x_1, \dots, x_{n+m-1}) \in X^{n+m-1}$,

$$(f \circ_i g)(x_1, \dots, x_{n+m-1}) := f(x_1, \dots, x_{i-1}, g(x_i, \dots, x_{i+m-1}), x_{i+m}, \dots, x_{n+m-1}).$$

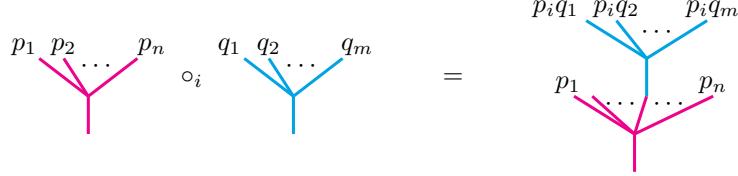
The simultaneous composition of several functions may also be considered. That is, given n functions $g_i \in C(X^{k_i}, X)$ where $i = 1, \dots, n$ they may be composed with f simultaneously to obtain a new function $f \circ(g_1, \dots, g_n) \in C(X^{k_1+\dots+k_n}, X)$, which is again defined by using the outputs of the g_i as the inputs of f . Explicitly, given $(x_1, \dots, x_{k_1+\dots+k_n}) \in X^{k_1+\dots+k_n}$, we have

$$(f \circ(g_1, \dots, g_n))(x_1, \dots, x_{k_1+\dots+k_n}) = f(g_1(x_1, \dots, x_{k_1}), \dots, g_n(x_{k_1+\dots+k_{n-1}+1}, \dots, x_{k_1+\dots+k_n}))$$

Example 2. The simplices $\Delta^0, \Delta^1, \Delta^2, \dots$ give rise to a topological operad called the **operad of topological simplices** $\Delta = \{\Delta_1, \Delta_2, \dots\}$ where $\Delta_n := \Delta^{n-1}$. The unit operation in Δ_1 is the unique probability distribution on a one-point set. If $p = (p_1, \dots, p_n) \in \Delta_n$ and $q = (q_1, \dots, q_m) \in \Delta_m$ are probability distributions, then the composition $p \circ_i q$ is obtained by multiplying each of the m coordinates of q by p_i and then replacing the i^{th} coordinate of p with the resulting m -tuple. Explicitly,

$$p \circ_i q := (p_1, \dots, p_i q_1, \dots, p_i q_m, \dots, p_n) \in \Delta_{n+m-1}.$$

Equivalently, the distribution p may be visualized as a planar tree with n leaves labeled by the probabilities p_1, \dots, p_n and similarly for q . Then the composition $p \circ_i q$ is obtained by “painting” each of the leaves of q with the probability p_i and grafting the resulting tree into the i^{th} leaf of p as below. Notice the sum of the probabilities on the leaves on the composite tree is 1.



As an example, if $p = (\frac{1}{6}, \dots, \frac{1}{6})$ represents the probability distribution of rolling a six-sided die and $q = (\frac{1}{2}, \frac{1}{2})$ is that of a fair coin toss, then $p \circ_3 q = (\frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ is a point in Δ_7 , whose picture is shown on the left of Figure 3.

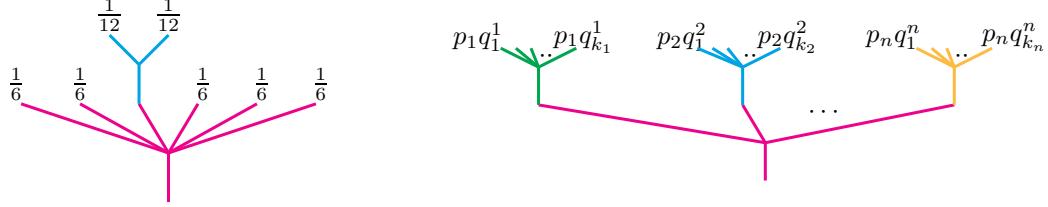


FIGURE 3. (Left) A picture of the composition $p \circ_3 q$ when p is the probability distribution associated to a six-sided die and q is that of a fair coin toss. (Right) The simultaneous composition of n probability distributions $q^i \in \Delta_{k_i}$ with a given $p \in \Delta_n$.

Further recall that if we have n different distributions $q^i = (q_1^i, \dots, q_{k_i}^i) \in \Delta_{k_i}$ where $i = 1, \dots, n$, then we may compose them with p simultaneously to obtain the following point in $\Delta_{k_1 + \dots + k_n}$,

$$p \circ (q^1, \dots, q^n) = (p_1 q_1^1, \dots, p_1 q_{k_1}^1, p_2 q_1^2, \dots, p_2 q_{k_2}^2, \dots, p_n q_1^n, \dots, p_n q_{k_n}^n).$$

This simultaneous composition is illustrated by the tree on the right in Figure 3.

Just as groups come to life when considering representations of them, so operads come to life when each abstract n -ary operation is mapped to a concrete n -ary operation on a particular object. This assignment is traditionally called an *algebra* of the operad, but we prefer the more descriptive name *representation*.

Definition 2. Let \mathcal{O} be an operad in the category of sets. A **representation of \mathcal{O}** , or an **\mathcal{O} -representation**, is set X together with functions

$$\varphi_n: \mathcal{O}(n) \rightarrow \text{End}_X(n) \quad \text{for } n \geq 1$$

that respect the operad unit and compositions. That is, $\varphi_n(1) = 1$ and

$$\varphi_{n+m-1}(p \circ_i q) = \varphi_n(p) \circ_i \varphi_m(q)$$

for all $p \in \mathcal{O}(n), q \in \mathcal{O}(m)$ and $1 \leq i \leq n$.

Importantly, one may also wish to define a representation of an operad in any symmetric monoidal category \mathbf{C} whenever “ $\text{End}_X(n)$ ” is in fact an object in \mathbf{C} . It must consist of an object X together with a family of morphisms $\mathcal{O}(n) \rightarrow \text{End}_X(n)$ in \mathbf{C} that are compatible with the operad unit and compositions. This holds, for instance, when the monoidal category \mathbf{C} is also closed—that is, when it is equipped with an internal hom functor that is compatible with the monoidal product. Monoidal closure, however, will not be required in our work, which primarily concerns the category \mathbf{Top} of topological spaces. Indeed, the main example to have in mind is when $\mathcal{O} = \Delta$ is the operad of simplices and $X = \mathbb{R}$ is the real line in \mathbf{Top} . In this case, we define $\text{End}_{\mathbb{R}}(n) := \mathbf{Top}(\mathbb{R}^n, \mathbb{R})$ to be the space of continuous functions $\mathbb{R}^n \rightarrow \mathbb{R}$ equipped with the product topology. Now, consider the continuous maps $\varphi_n: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)$ given by $p \mapsto \varphi_n(p)$ where $\varphi_n(p)(x) := \langle p, x \rangle = \sum_{i=1}^n p_i x_i$ whenever $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Then, it is simple to check that $\varphi_{n+m-1}(p \circ_i q) = \varphi_n(p) \circ_i \varphi_m(q)$ for all p, q , and i and that $\varphi_n(1) = 1$ for all n , and so \mathbb{R} is a representation of Δ .

4. DERIVATIONS OF THE OPERAD OF SIMPLICES

With these basic definitions in hand, the present goal is to define a mapping d out of the topological operad Δ that satisfies an appropriate version of the Leibniz rule,

$$(3) \quad d(p \circ_i q) = dp \circ_i q + p \circ_i dq \quad (\text{desideratum})$$

for all $p \in \Delta_n$ and $q \in \Delta_m$ and for all $1 \leq i \leq n$. This desired equation suggests the codomain of d should be a (bi)module over Δ that is, moreover, an abelian monoid. This motivates the following two definitions, the first of which is a slight generalization of that given by Markl in [Mar96].

Definition 3. Let $\mathcal{O} = \{\mathcal{O}(1), \mathcal{O}(2), \dots\}$ be an operad in a symmetric monoidal category \mathbf{C} . A **bimodule over \mathcal{O}** , or simply an **\mathcal{O} -bimodule**, is a collection of objects $M = \{M(1), M(2), \dots\}$ in \mathbf{C} together with morphisms

$$\begin{aligned} \circ_i^L &= \mathcal{O}(n) \otimes M(m) \rightarrow M(n+m-1) && (\text{left composition}) \\ \circ_i^R &= M(n) \otimes \mathcal{O}(m) \rightarrow M(n+m-1) && (\text{right composition}) \end{aligned}$$

in \mathbf{C} for each $1 \leq i \leq n$ such that whenever

$$p \otimes q \otimes r \in \begin{cases} M(n) \otimes \mathcal{O}(m) \otimes \mathcal{O}(k), \text{ or} \\ \mathcal{O}(n) \otimes M(m) \otimes \mathcal{O}(k), \text{ or} \\ \mathcal{O}(n) \otimes \mathcal{O}(m) \otimes M(k) \end{cases}$$

the following holds:

$$(4) \quad (p \circ_j q) \circ_i r = \begin{cases} (p \circ_i r) \circ_{j+k-1} q & \text{if } 1 \leq i \leq j-1 \\ p \circ_j (q \circ_{i-j+1} r) & \text{if } j \leq i \leq j+m-1 \\ (p \circ_{i-m+1} r) \circ_j q & \text{if } i \geq j+m. \end{cases}$$

The associativity requirements displayed in Equation (4)—and hence the intuition behind them—are completely analogous to those defining operads as illustrated in Figure 2. The only difference here is that one of the three operations may

come from the bimodule rather than the operad. Here is the main example to have in mind.

Example 3. As every algebra is a bimodule over itself, so every representation of \mathcal{O} is an \mathcal{O} -bimodule in a straightforward way. Indeed, in the case of the topological operad of simplices, the maps comprising the Δ -representation structure on \mathbb{R} induce a Δ -bimodule structure on $\text{End}_{\mathbb{R}}$. However, we will make use of a slight variant of this bimodule structure. Right composition will be defined in the expected way, though left composition will not. Explicitly, we define the left and right composition maps

$$\begin{aligned}\circ_i^L: \Delta_n \times \text{Top}(\mathbb{R}^m, \mathbb{R}) &\longrightarrow \text{Top}(\mathbb{R}^{n+m-1}, \mathbb{R}) \\ \circ_i^R: \text{Top}(\mathbb{R}^n, \mathbb{R}) \times \Delta_m &\longrightarrow \text{Top}(\mathbb{R}^{n+m-1}, \mathbb{R})\end{aligned}$$

as follows. Given a probability distribution $p \in \Delta_n$ and a continuous function $f: \mathbb{R}^m \rightarrow \mathbb{R}$, define left composition by $p \circ_i^L f := \bar{p} \circ (0, \dots, 0, f, 0, \dots, 0)$, where the composition on the right-hand side is defined as in the simultaneous composition in the endomorphism operad of \mathbb{R} illustrated in Example 1, and where each 0 denotes the zero function $\mathbb{R} \rightarrow \mathbb{R}$. Here, recall that $\bar{p}: \mathbb{R}^n \rightarrow \mathbb{R}$ maps a point x to the standard inner product $\langle p, x \rangle$ as introduced in Section 1. Unwinding this, left composition thus evaluates explicitly as $(p \circ_i^L f)(x_1, \dots, x_{n+m-1}) = p_i f(x_i, \dots, x_{i+m-1})$. In words, the value of the left composite $p \circ_i^L f: \mathbb{R}^{n+m-1} \rightarrow \mathbb{R}$ at a point x is computed by evaluating f at the m -tuple of x beginning at the i^{th} coordinate and scaling that output by p_i . All other coordinates of x are ignored. The picture to have in mind is that below, where the bold dots are imagined to be “plugs” that prevent the surplus coordinates from playing a role. In this picture, $n = 3$ and $m = 2$.

$$(p \circ_2^L f)(x_1, x_2, x_3, x_4) = \begin{array}{c} x_2 \\ \diagdown \quad \diagup \\ x_1 \bullet \quad f \quad x_3 \\ \diagup \quad \diagdown \\ x_4 \bullet \\ p \end{array} = p_2 f(x_2, x_3)$$

Given a probability distribution $q \in \Delta_m$ and a continuous function $g: \mathbb{R}^n \rightarrow \mathbb{R}$, define right composition by

$$\begin{aligned}(g \circ_i^R q)(x_1, \dots, x_{n+m-1}) &:= g(x_1, \dots, x_{i-1}, \sum_{k=1}^m q_k x_{i+k-1}, x_{i+m}, \dots, x_{n+m-1}).\end{aligned}$$

This may be understood visually as well. The value of the right composite $g \circ_i^R q: \mathbb{R}^{n+m-1} \rightarrow \mathbb{R}$ at a point x is computed by taking the inner product of q with the m -tuple of x beginning at the i^{th} coordinate and using that number as the i^{th} input of g with all other coordinates of x falling into place as in the picture below. There are no “plugs” in this instance since all coordinates of x play a role.

$$(g \circ_2^R q)(x_1, x_2, x_3, x_4) = \begin{array}{c} x_2 \\ \diagdown \quad \diagup \\ x_1 \quad q \quad x_3 \\ \diagup \quad \diagdown \\ x_4 \\ g \end{array} = g(x_1, q_1 x_2 + q_2 x_3, x_4)$$

These examples suggest the inner product notation is a convenient choice. Given $N \geq 1$ and $k \leq N$ and a point $x \in \mathbb{R}^N$, let $\mathbf{x}_{i,k} \in \mathbb{R}^k$ denote the k -subtuple of x beginning at the i^{th} coordinate:

$$\mathbf{x}_{i,k} := (x_i, \dots, x_{i+k-1}).$$

Then given any point $x \in \mathbb{R}^{n+m-1}$, the left and right composition maps may be written more succinctly as

$$\begin{aligned} (p \circ_i^L f)(x) &= p_i f(\mathbf{x}_{i,m}) \\ (g \circ_i^R q)(x) &= g(x_1, \dots, x_{i-1}, \langle q, \mathbf{x}_{i,m} \rangle, x_{i+m}, \dots, x_{n+m-1}). \end{aligned}$$

We will use this notation below and will always write \mathbf{x}_i in lieu of $\mathbf{x}_{i,m}$ since the context will make it clear that \mathbf{x}_i must be an m -tuple. The boldface font is used to distinguish a tuple \mathbf{x}_i from a real number x_i . Finally, note that the maps \circ_i^L and \circ_i^R are continuous since f and g are continuous, and moreover that the associativity requirements in Equation (4) are analogous to those illustrated in Figure 2, so it is straightforward to verify they are satisfied. In particular, the zero functions appearing in the definition of \circ_i^L simplify the situation greatly. For instance, several of the associativity requirements follow from the simple fact that multiplying an input x_i by a probability and then mapping the result to zero is the same as first mapping the input to zero and then multiplying that zero by a probability. So $\text{End}_{\mathbb{R}}$ is indeed a Δ -bimodule.

Next, recall that the desired Leibniz rule in Equation (3) suggests the bimodule should be equipped with a notion of addition. This motivates the following definition.

Definition 4. Let \mathcal{O} be an operad in a symmetric monoidal category C . An \mathcal{O} -bimodule M is an **abelian \mathcal{O} -bimodule** if each $M(n)$ is an abelian monoid in C ; that is, if for each $n = 1, 2, \dots$ the following hold:

- (i) [associativity, commutativity] there is a morphism $\mu_n: M(n) \times M(n) \rightarrow M(n)$ in C such that $\mu_n(\mu_n(a, b), c) = \mu_n(a, \mu_n(b, c))$ and $\mu_n(a, b) = \mu_n(b, a)$ for all $a, b, c \in M(n)$,
- (ii) [identity] there is an element $1 \in M(n)$ such that $\mu_n(1, a) = a = \mu_n(a, 1)$ for all $a \in M(n)$.

As the primary example, consider $\text{End}_{\mathbb{R}}$ viewed as a Δ -bimodule as described in Example 3. For each n , define $\mu_n: \text{End}_{\mathbb{R}}(n) \times \text{End}_{\mathbb{R}}(n) \rightarrow \text{End}_{\mathbb{R}}(n)$ by pointwise addition, meaning that for each $f, g \in \text{End}_{\mathbb{R}}(n)$ we have $\mu_n(f, g) = f + g$ where $(f + g)(x) := f(x) + g(x)$ for all $x \in \mathbb{R}^n$. The identity element in $\text{End}_{\mathbb{R}}(n)$ is the constant map at zero. Moreover each μ_n is continuous and inherits associativity and commutativity from \mathbb{R} . In this way, $\text{End}_{\mathbb{R}}$ is an abelian Δ -bimodule.

Remark 1. Notice that the Δ -bimodule composition maps \circ_i^L and \circ_i^R distribute over sums in the abelian Δ -bimodule $\text{End}_{\mathbb{R}}$. In other words, for all continuous functions $f, g \in \text{End}_{\mathbb{R}}(n)$ and for all probability distributions $q \in \Delta_m$,

$$(f + g) \circ_i^R q = f \circ_i^R q + g \circ_i^R q, \quad 1 \leq i \leq n$$

and similarly for left composition \circ_i^L . This follows directly from pointwise addition.

With this setup in mind, our desideratum in Equation (3) is now realized in the following definition.

Definition 5. Let \mathcal{O} be an operad in a category C and let M be an abelian \mathcal{O} -bimodule. A **derivation of \mathcal{O} valued in M** is sequence of morphisms $\{d_n: \mathcal{O}(n) \rightarrow M(n)\}$ in C satisfying

$$(5) \quad d_{n+m-1}(p \circ_i q) = d_n p \circ_i^R q + p \circ_i^L d_m q$$

for all $p \in \mathcal{O}(n), q \in \mathcal{O}(m)$ and for all $1 \leq i \leq n$.

In the special case when \mathcal{O} is a linear operad, this definition coincides with that given by Markl in [Mar96]. In what follows, we omit the subscripts and simply write d instead of d_n . Now, suppose $\mathcal{O} = \Delta$ is the operad of topological simplices and $\text{End}_{\mathbb{R}}$ is equipped with the structure of an abelian Δ -bimodule given above. Here is the picture to have in mind for Equation (5):

$$d \left(\begin{array}{c} \text{blue line} \\ \text{pink line} \end{array} \right) = \begin{array}{c} \text{blue line} \\ \text{pink line} \end{array} + \bullet \begin{array}{c} d \\ \text{blue line} \\ \text{pink line} \end{array} \bullet$$

On the right-hand side we have used the “plug” notation introduced in Example 3, which can also be understood explicitly by evaluating d at a point $x \in \mathbb{R}^{n+m-1}$,

$$\begin{aligned} d(p \circ_i q)(x) &= (dp \circ_i^R q)(x) + (p \circ_i^L dq)(x) \\ &= dp(x_1, \dots, \langle q, \mathbf{x}_i \rangle, \dots, x_{n+m-1}) + p_i dq(\mathbf{x}_i). \end{aligned}$$

Of particular interest is the behavior of a derivation $\{d: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)\}$ when it is applied to a simultaneous composition of probability distributions. A derivation applied to the composite $(p \circ_j q) \circ_i r$ for probability distributions $p \in \Delta_n, q \in \Delta_m$, and $r \in \Delta_k$ can be understood in a convenient picture when q and r are composed onto different leaves of p ; that is, when $1 \leq i \leq j-1$ or $i \geq j+m$. This follows straightforwardly from a repeated application of d . Indeed, by definition we have $d((p \circ_j q) \circ_i r) = d(p \circ_j q) \circ_i^R r + (p \circ_j q) \circ_i^L dr$ and by applying the Leibniz rule again to the first summand, this is equal to $(dp \circ_j^R q + p \circ_j^L dq) \circ_i^R r + (p \circ_j q) \circ_i^L dr$, which we can expand to obtain $(dp \circ_j^R q) \circ_i^R r + (p \circ_j^L dq) \circ_i^R r + (p \circ_j q) \circ_i^L dr$ since composition distributes over sums as noted in Remark 1. We will identify this function with the picture below in lieu of the cumbersome notation.

$$d \left(\begin{array}{c} \text{green line} \\ \text{pink line} \end{array} \right) = \begin{array}{c} \text{green line} \\ \text{pink line} \end{array} + d \begin{array}{c} \text{green line} \\ \text{pink line} \end{array} \bullet \begin{array}{c} \bullet \\ \text{blue line} \end{array} + \bullet \begin{array}{c} \text{green line} \\ \text{pink line} \end{array} d \bullet$$

Importantly, the obvious generalization of the formula holds for any simultaneous composition $p \circ (q^1, \dots, q^n)$ for any $p \in \Delta_n$ and $q^i \in \Delta_{k_i}$ where $i = 1, \dots, n$. This again follows directly from repeated applications of Equation (5), as illustrated below.

$$d \left(\begin{array}{c} \text{green} \\ \text{pink} \\ \text{blue} \end{array} \right) = \begin{array}{c} \text{green} \\ \text{pink} \\ \text{blue} \end{array} + d \begin{array}{c} \text{green} \\ \text{pink} \\ \text{blue} \end{array} + \begin{array}{c} \text{green} \\ \text{pink} \\ \text{blue} \end{array} + \begin{array}{c} \text{green} \\ \text{pink} \\ \text{blue} \end{array}$$

This is summarized in the following proposition.

Proposition 1. *Let $p \in \Delta_n$ and $q^i \in \Delta_{k_i}$ for $n, k_1, \dots, k_n \geq 1$ and let $\{d: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)\}$ be a derivation of the operad of topological simplices. Then for any point $x \in \mathbb{R}^{k_1+\dots+k_n}$,*

$$d(p \circ (q^1, \dots, q^n))(x) = dp(\langle q^1, x_1 \rangle, \dots, \langle q^n, x_n \rangle) + \sum_{i=1}^n p_i dq^i(x_i).$$

Finally, the main result follows.

Theorem 2. *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.*

Proof. For each $n \geq 1$ define $d: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)$ by $p \mapsto dp$ where $dp(x) = H(p)$ is constant for all $x \in \mathbb{R}^n$. Then, d is continuous since H is continuous. Moreover, if $p = (p_1, \dots, p_n) \in \Delta_n$ and $q = (q_1, \dots, q_m) \in \Delta_m$ are probability distributions, then for any $x \in \mathbb{R}^{m+n-1}$ and $1 \leq i \leq n$, we have

$$\begin{aligned} d(p \circ_i q)(x) &= H(p \circ_i q) = - \left(\sum_{k=1}^{i-1} p_k \log p_k + p_i \sum_{k=1}^m q_k \log(p_i q_k) + \sum_{k=i+1}^n p_k \log p_k \right) \\ &= - \left(\sum_{k=1}^{i-1} p_k \log p_k + p_i \log p_i \sum_{k=1}^m q_k + p_i \sum_{k=1}^m q_k \log q_k + \sum_{k=i+1}^n p_k \log p_k \right) \\ &= - \left(\sum_{k=1}^n p_k \log p_k + p_i \sum_{k=1}^m q_k \log q_k \right) \\ &= H(p) + p_i H(q) \\ &= (dp \circ_i^R q + p \circ_i^L dq)(x), \end{aligned}$$

where the last line follows since $(dp \circ_i^R q)(x)$ is computed by evaluating the function dp at some point, and this function is assumed to be constant at $H(p)$.

Conversely, suppose $\{d: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)\}$ is a derivation. For each $n \geq 1$ define a function $F: \Delta_n \rightarrow \mathbb{R}$ by $F(p) = dp(0)$ where $0 = (0, \dots, 0) \in \mathbb{R}^n$. Then F is continuous since d is continuous, and Proposition 1 further implies that

$$\begin{aligned}
F(p \circ (q^1, \dots, q^n)) &= d(p \circ (q^1, \dots, q^n))(0) \\
&= dp(\langle q^1, \mathbf{0}_1 \rangle, \dots, \langle q^n, \mathbf{0}_n \rangle) + \sum_{i=1}^n p_i dq^i(\mathbf{0}_i) \\
&= dp(0) + \sum_{i=1}^n p_i dq^i(0) \\
&= F(p) + \sum_{i=1}^n p_i F(q^i).
\end{aligned}$$

From the Faddeev–Leinster result in Theorem 1, it follows that $dp(0) = F(p) = cH(p)$ for some $c \in \mathbb{R}$. \square

Notice that the important Equation (2) mentioned in the introduction is obtained as a corollary. Indeed, if for each $n \geq 1$ the map $d: \Delta_n \rightarrow \text{End}_{\mathbb{R}}(n)$ is defined to be constant at entropy $p \mapsto dp \equiv H(p)$, then d is a derivation by Theorem 2 and so Proposition 1 yields the following by evaluating $d(p \circ (q^1, \dots, q^n))$ at any point.

Corollary. *Let $p \in \Delta_n$ and $q^i \in \Delta_{k_i}$ with $1 \leq i \leq n$. Then*

$$H(p \circ (q^1, \dots, q^n)) = H(p) + \sum_{i=1}^n p_i H(q^i).$$

As a closing remark, Faddeev’s characterization of entropy in Theorem 1 can be reexpressed using the language of category theory and operads as in [Lei21], (Theorem 12.3.1). We have omitted this language here but invite the reader to explore the full category theoretical story in Chapter 12 of Leinster’s book.

REFERENCES

- [Bae11] John C. Baez. Entropy as a functor, 2011. Blog post. Available online: <https://www.ncatlab.org/johnbaez/show/Entropy+as+a+functor>.
- [BB15] Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17(5):3253–3318, 2015.
- [BFL11] John C. Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. *Entropy*, 13:1945–1957, 2011. doi:10.3390/e13111945.
- [BV73] J. M. Boardman and R. Vogt. Homotopy invariant algebraic structures on topological spaces. volume 347 of *Lecture Notes in Mathematics*. Springer, 1973.
- [EVG15] Philippe Elbaz-Vincent and Herbert 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.
- [Fad56] D. K. Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk*, 11:227–231, 1956. (In Russian).
- [Lam69] Joachim Lambek. Deductive systems and categories ii. standard constructions and closed categories. In P. Hilton, editor, *Category Theory, Homology Theory and their Applications, I (Battelle Institute Conference, Seattle, 1968)*, volume 68 of *Lecture Notes in Mathematics*. Springer, 1969.
- [Lei21] Tom Leinster. *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press, 2021.
- [LV12] Jean-Louis Loday and Bruno Vallette. *Algebraic Operads*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012.
- [Mai19] Tom Mainiero. Homological tools for the quantum mechanic. 2019. arXiv:1901.02011.

- [Mar96] Martin Markl. Models for operads. *Communications in Algebra*, 24(4):1471–1500, 1996. arXiv: arxiv.org/abs/hep-th/9411208.
- [May72] J.P. May. The geometry of iterated loop spaces. volume 271 of *Lecture Notes in Mathematics*. Springer, 1972.
- [MSS02] Martin Markl, Steven Shnider, and Jim Stasheff. *Operads in Algebra, Topology and Physics*. Mathematical surveys and monographs. American Mathematical Society, 2002.
- [Par20] Arthur J. Parzygnat. A functorial characterization of von neumann entropy. 2020. arXiv:2009.07125.
- [Sta04] Jim Stasheff. What is... an operad? *Notices Amer. Math. Soc.*, 51:630–631, 2004.
- [Val12] Bruno Vallette. Algebra + homotopy = operad. 2012. arXiv:1202.3245.