entropy

MDPI

Article

Dirichlet Polynomials and Entropy

David I. Spivak and Timothy Hosgood *

Topos Institute, Berkeley, CA 94704, USA; david@topos.institute
* Correspondence: tim@topos.institute

Abstract: A Dirichlet polynomial d in one variable $\mathcal{y}$ is a function of the form $d(\mathcal{y}) = a_n n^{\mathcal{y}} + \cdots + a_2 2^{\mathcal{y}} + a_1 1^{\mathcal{y}} + a_0 0^{\mathcal{y}}$ for some $n, a_0, \ldots, a_n \in \mathbb{N}$. We will show how to think of a Dirichlet polynomial as a set-theoretic bundle, and thus as an empirical distribution. We can then consider the Shannon entropy $H(d)$ of the corresponding probability distribution, and we define its *length* (or, classically, its *perplexity*) by $L(d) = 2^{H(d)}$. On the other hand, we will define a rig homomorphism $h\colon \mathsf{Dir} \to \mathsf{Rect}$ from the rig of Dirichlet polynomials to the so-called *rectangle rig*, whose underlying set is $\mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0}$ and whose additive structure involves the weighted geometric mean; we write $h(d) = (A(d), W(d))$, and call the two components *area* and *width* (respectively). The main result of this paper is the following: the rectangle-area formula $A(d) = L(d)W(d)$ holds for any Dirichlet polynomial d. In other words, the entropy of an empirical distribution can be calculated entirely in terms of the homomorphism h applied to its corresponding Dirichlet polynomial. We also show that similar results hold for the cross entropy.

Keywords: bundle; weighted geometric mean; category theory; Dirichlet polynomial

check for updates

Citation: Spivak, D.I.; Hosgood, T. Dirichlet Polynomials and Entropy. *Entropy* **2021**, *23*, 1085. https://doi.org/10.3390/e23081085

Academic Editor: Karl Svozil

Received: 14 July 2021
Accepted: 19 August 2021
Published: 21 August 2021

1. Introduction

The purpose of this paper is simply to provide another categorical treatment of *entropy* of probability distributions, which turns out to be computed in terms of a rig homomorphism; our treatment also generalises to *cross entropy* (and thus to *Kullback–Leibler divergence*). What is particularly interesting about the treatment outlined here is that we can somewhat "visualise" the notion of entropy in terms of sizes of coding schemes (cf. Section 6). Not only that, but classical entropy is only homomorphic in the product of distributions, whereas the notion that we describe here is homomorphic in both the product and the sum.

A brief outline of this paper is as follows:

Section 2: We recall the definitions of *Dirichlet polynomials* and *set-theoretic bundles*, along with their rig structures, from [1] (one important thing to note is the following: Dirichlet polynomials are well-studied objects in the setting of complex analysis, but we *cannot* apply tools from this area to our setting, because we have only *natural number* coefficients and *non-negative* exponents); we then study the equivalence between these two notions.

Section 3: We explain how *empirical probability distributions* correspond to set-theoretic bundles (and thus to Dirichlet polynomials).

Section 4: We define the *rig homomorphism* $h\colon \mathsf{Dir} \to \mathsf{Rect}$ that we wish to study, whose codomain is a rig encoding the weighted geometric mean; we prove some useful computational results and give some explicit examples.

Section 5: We define the *entropy* $H(d)$ of a Dirichlet polynomial using the classical notion of Shannon entropy; we give some explicit examples; we prove the main result of this paper (Theorem 1), relating entropy to the rig homomorphism defined in the previous section.

Section 6: We try to provide some intuition for the image $h(d)$ of a Dirichlet polynomial under the rig homomorphism, in terms of *coding schemes*.

Section 7: We generalise Theorem 1 to the case of *cross-entropy*, or *Kullback–Leibler divergence*.

Prerequisites

We assume that the reader is familiar with the fundamentals of category theory, such as functors, natural transformations, and (co)products, as covered, for example, in [2] (Chapter 1).

2. Dirichlet Polynomials and Bundles

This section is simply a brief summary of content from [1], repeated here for the convenience of the reader.

Definition 1. *A* Dirichlet polynomial *d in one variable* y *is a function of the form*

$$d(\mathrm{y}) = a_n n^{\mathrm{y}} + \ldots + a_2 2^{\mathrm{y}} + a_1 1^{\mathrm{y}} + a_0 0^{\mathrm{y}}$$

for some $n, a_0, \ldots, a_n \in \mathbb{N}$.

The set of Dirichlet polynomials is clearly closed under addition, and further under multiplication (using the distributive law along with the fact that $m^{\mathrm{y}} \cdot n^{\mathrm{y}} = (m \cdot n)^{\mathrm{y}}$). In fact, it has the structure of a rig*: a "ring without negatives" (or, to be pedantic, a monoid object in commutative monoids). We denote this rig by* Dir*, where the additive unit is* 0*, and the multiplicative unit is* 1^{y}*.*

Note that we can embed $\mathbb{N}$ as a sub-rig of Dir*, by $a \mapsto a \cdot 1^{\mathrm{y}}$; we often use this fact and simply write $a \in$* Dir*.*

Following [1], we can think of Dirichlet polynomials as functors $\mathtt{FSet}^{\mathrm{op}} \to \mathtt{FSet}$, where $\mathtt{FSet}$ is the category of finite sets. Indeed, given a natural number $n \in \mathbb{N}$, the *exponential* n^{y} can be thought of as the Yoneda embedding of the set with n elements, i.e.,

$$n^{\mathrm{y}} = \mathtt{FSet}\,(-, \underline{n})$$

where $\underline{n} = \{1, \ldots, n\}$. (For typographical convenience, we sometimes use the notation n and $\underline{n}$ interchangeably. In particular, we write e.g., $d(0)$ instead of $d(\underline{0})$). Then the addition of exponentials corresponds to the coproduct of the corresponding representable functors (and so multiplication by a natural number a_n corresponds to the a_n-fold coproduct of the representable functor with itself). This means that evaluating a Dirichlet polynomial at some natural number n corresponds to evaluating the corresponding functor on the finite set $\underline{n}$.

Note that 0^{y} is *not* the initial object 0, since

$$0^{\underline{n}} = \begin{cases} 1 & \text{if } n = 0; \\ 0 & \text{if } n \geqslant 1 \end{cases}$$

i.e., $0^{\mathrm{y}} \neq 0$.

Example 1. *The Dirichlet polynomial*

$$d(\mathrm{y}) = 4^{\mathrm{y}} + 4 \cdot 1^{\mathrm{y}}$$

evaluated at 0 *gives*

$$\begin{aligned} d(0) &= \mathtt{FSet}\,(\underline{0}, \underline{4}) \sqcup \left(\sqcup_{i=1}^{4} \mathtt{FSet}\,(\underline{0}, \underline{1}) \right) \\ &\cong \underline{1} \sqcup \underline{4} \\ &\cong \underline{5}, \end{aligned}$$

and, similarly,

$$\begin{aligned} d(1) &= \mathtt{FSet}\,(\underline{1}, \underline{4}) \sqcup \left(\sqcup_{i=1}^{4} \mathtt{FSet}\,(\underline{1}, \underline{1}) \right) \\ &\cong \underline{4} \sqcup \underline{4} \\ &\cong \underline{8}. \end{aligned}$$

Note that, since $1^{\mathrm{y}} = 1$, we can write $d(\mathrm{y}) = 4^{\mathrm{y}} + 4$.

Definition 2. *A* morphism $\varphi \colon d \to e$ *of Dirichlet polynomials is a natural transformation of (contravariant) functors. Denote by* $\mathtt{Dir}$ *the category of Dirichlet polynomials (thought of as functors* $\mathtt{FSet}^{\mathrm{op}} \to \mathtt{FSet}$*), and by* $\mathtt{Dir}\,(d, e)$ *the set of all morphisms* $d \to e$*.*

When we think of Dirichlet polynomials as functors $\mathtt{FSet}^{\mathrm{op}} \to \mathtt{FSet}$, addition is given by the coproduct (disjoint union of sets), and multiplication by the product (Cartesian product of sets). This means that working with Dirichlet polynomials in $\mathtt{Dir}$ really is like working with polynomials, in the sense that addition and multiplication are exactly "as expected".

Example 2. *The only slightly confusing aspect of multiplication in* Dir *is how 0^{y} behaves (since $0^{\mathrm{y}} \neq 0$): if $d(\mathrm{y})$ is a Dirichlet polynomial, then*

$$d(\mathrm{y}) \cdot 0^{\mathrm{y}} = d(0) \cdot 0^{\mathrm{y}},$$

as follows from the aforementioned fact that $0^{\underline{n}}$ is zero for $n \neq 0$, and 1 for $n = 0$.

We can use this general fact for specific computations. For example, let

$$\begin{aligned} d(\mathrm{y}) &:= 3 \cdot 2^{\mathrm{y}} + 1^{\mathrm{y}} \\ e(\mathrm{y}) &:= 4^{\mathrm{y}} + 2^{\mathrm{y}} + 3 \cdot 0^{\mathrm{y}} \end{aligned}$$

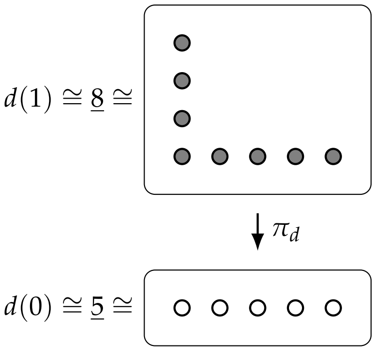
Then

$$\begin{aligned} (d \cdot e)(\mathrm{y}) &= (3 \cdot 8^{\mathrm{y}} + 3 \cdot 4^{\mathrm{y}} + 9 \cdot 0^{\mathrm{y}}) + (4^{\mathrm{y}} + 2^{\mathrm{y}} + 3 \cdot 0^{\mathrm{y}}) \\ &= 3 \cdot 8^{\mathrm{y}} + 4 \cdot 4^{\mathrm{y}} + 2^{\mathrm{y}} + 12 \cdot 0^{\mathrm{y}}. \end{aligned}$$

(and $d + e = 4^{\mathrm{y}} + 4 \cdot 2^{\mathrm{y}} + 1^{\mathrm{y}} + 3 \cdot 0^{\mathrm{y}}$).

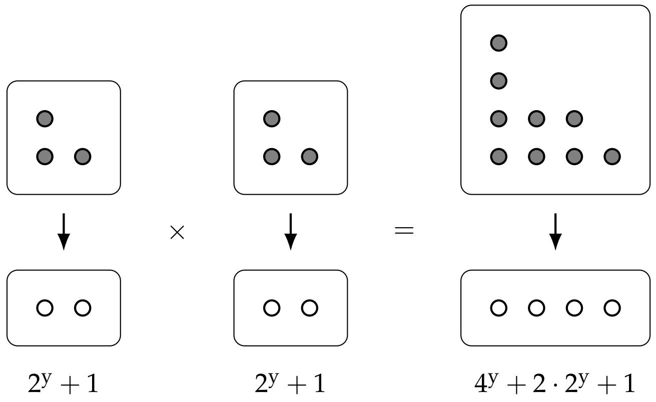
There is a more geometric interpretation of objects of Dir as *set-theoretic bundles*, i.e., (isomorphism classes of) functions $E \to B$, where $E, B \in \mathtt{FSet}$, given as follows: to the Dirichlet polynomial $d \colon \mathtt{FSet}^{\mathrm{op}} \to \mathtt{FSet}$, we associate the function $\pi_d = d(!) \colon d(1) \to d(0)$ induced by the unique function $! \colon \underline{0} \to \underline{1}$. (This vague statement can be upgraded to an equivalence of categories: ([1], [Theorem 4.6]) for more details).

For example, to the polynomial $d(\mathrm{y}) = 4^{\mathrm{y}} + 4 \cdot 1^{\mathrm{y}}$, we associate the bundle


$d(1) \cong \underline{8} \cong$
π_d
$d(0) \cong \underline{5} \cong$

Note that bundles also form a rig, where the sum is given by disjoint union of sets, and the product is given by the Cartesian product of sets, both on the base and the total space. Further, the equivalence between Dirichlet polynomials and bundles respects the rig structures (cf. [1], Theorem 4.6). Because of this, we often switch freely between thinking of Dirichlet polynomials as functors $d \colon \mathtt{FSet}^{\mathrm{op}} \to \mathtt{FSet}$, as bundles $\pi_d \colon d(1) \to d(0)$, and simply as functions of the form $\sum_{j=0}^{n} a_j \cdot j^{\mathrm{y}}$.

Example 3. *We can draw the bundle corresponding to* $(2^{\mathrm{y}}+1)\cdot(2^{\mathrm{y}}+1)$ *as follows:*



Lemma 1. *Let* $d(\mathrm{y}) := \sum_{j=0}^{n} a_j \cdot j^{\mathrm{y}}$ *be a Dirichlet polynomial. Then*

$$|d(0)| = \sum_{j=0}^{n} a_j$$
$$|d(1)| = \sum_{j=0}^{n} a_j j.$$

Proof. This follows from the fact that $n^{\underline{0}} = 1$ and $n^{\underline{1}} = n$, for all $n \in \mathbb{N}$. □

Definition 3. *Let* $d \in \mathsf{Dir}$*. For* $i \in d(0)$*, we define*

$$d[i] := \pi_d^{-1}(i)$$

where $\pi_d \colon d(1) \to d(0)$ *is the bundle corresponding to* d*.*

Using the fact that the sum of bundles is given by the disjoint union of sets, we can use this above definition to write any Dirichlet polynomial d as

$$d(\mathrm{y}) \cong \sum_{i \in d(0)} d[i]^{\mathrm{y}}$$

(where $\sum$ is the coproduct in $\mathtt{FSet}$).

Corollary 1. *Let* $d(\mathrm{y}) := \sum_{i \in d(0)} d[i]^{\mathrm{y}}$ *be a Dirichlet polynomial. Then*

$$|d(1)| = \sum_{i \in d(0)} |d[i]|.$$

Proof. This is, again, simply the fact that $n^{\underline{1}} = n$ for all $n \in \mathbb{N}$. □

Lemma 2. *A morphism* $\varphi \colon d \to e$ *of Dirichlet polynomials is exactly a morphism of the corresponding bundles, i.e., functions* $\varphi_0 \colon d(0) \to e(0)$ *and* $\varphi_1 \colon d(1) \to e(1)$ *such that*

$$\begin{tikzcd} d(1) \arrow[r, "\varphi_1"] \arrow[d, "\pi_d"'] & e(1) \arrow[d, "\pi_e"] \\ d(0) \arrow[r, "\varphi_0"'] & e(0) \end{tikzcd}$$

commutes.

Proof. This statement forms a specific part of ([1], Theorem 4.6), but the proof is simple enough that we give a direct version here. Writing $d(\mathrm{y}) := \sum_{i\in d(0)} d[i]^{\mathrm{y}}$ and $e(\mathrm{y}) := \sum_{i\in e(0)} e[i]^{\mathrm{y}}$, we see that

$$\begin{aligned}\mathrm{Hom}_{[\mathtt{FSet}^{\mathrm{op}},\mathtt{FSet}]}(d,e) &\cong \prod_{i\in d(0)} \mathrm{Hom}_{[\mathtt{FSet}^{\mathrm{op}},\mathtt{FSet}]}\left(d[i]^{\mathrm{y}}, \sum_{j\in e(0)} e[j]^{\mathrm{y}}\right)\\ &\cong \prod_{i\in d(0)} \sum_{j\in e(0)} e[j]^{d[i]}\\ &= \prod_{i\in d(0)} \sum_{j\in e(0)} \mathrm{Hom}_{\mathtt{FSet}}(d[i],e[j])\end{aligned}$$

(the first isomorphism is by the universal property of the coproduct; the second isomorphisms is the Yoneda lemma). However, an element of this set is exactly a bundle morphism: we have, for all $i \in d(0)$, some $j \in e(0)$ along with a function $d[i] \to e[j]$; φ_1 is given by the disjoint union of all these $d[i] \to e[j]$, and φ_0 is given by the the choice of j for each i. □

Definition 4. *Given Dirichlet polynomials $d, e \in \mathtt{Dir}$ such that $d(0) = e(0)$, we denote by $\mathtt{Dir}_{/d(0)}(d,e)$ the set of morphisms $(\varphi_0, \varphi_1)\colon d \to e$ such that $\varphi_0 = \mathrm{id}$.*

Given the correspondence between Dirichlet polynomials and bundles, we might rightly ask why we should prefer to work with the former over the latter. For one possible answer to this, see Section 6.

3. Bundles as Empirical Distributions

The interpretation of Dirichlet polynomials as bundles helps us to understand how they relate to probability theory. Imagine flipping a coin eight times and observing five heads and three tails; we refer to "heads" and "tails" as *outcomes*, and each of the eight flips as *draws*; every draw has an associated outcome.

Consider some bundle $\pi_d\colon d(1) \to d(0)$. We can think of $d(0)$ as the set of outcomes, and $d(1)$ as the set of draws; the fibre $\pi_d^{-1}(x)$ over an outcome $x \in d(0)$ corresponds to all the draws that lead to the outcome x, and so we obtain a probability distribution on $d(0)$ by setting $\mathbb{P}(X = x) = \frac{|\pi_d^{-1}(x)|}{|d(0)|}$. Conversely, any *rational distribution* (i.e., a distribution such that all probabilities are rational numbers) on a finite set arises in this way: take the finite set as the set of outcomes; take the least common multiple of the denominators of all the probabilities as the cardinality of $d(1)$; and then take $\mathbb{P}(X = x) \cdot |d(1)|$ many elements of $d(1)$ to be in the fibre of $x \in d(0)$.

Example 4. *Consider the set $S = \{x_1, x_2, x_3, x_4\}$, endowed with the probability distribution such that*

$$\begin{aligned}\mathbb{P}(x_1) &= \frac{1}{5} & \mathbb{P}(x_2) &= \frac{1}{6}\\ \mathbb{P}(x_3) &= \frac{1}{2} & \mathbb{P}(x_4) &= \frac{2}{15}\end{aligned}$$

Define the sets $d(0) = \underline{4}$ and $d(1) = \underline{30}$, and define the function $\pi_d\colon d(1) \to d(0)$ by

$$\pi(n) = \begin{cases} 1 & \text{if } 0 \leqslant n < 6;\\ 2 & \text{if } 6 \leqslant n < 11;\\ 3 & \text{if } 11 \leqslant n < 26;\\ 4 & \text{if } 26 \leqslant n < 30.\end{cases}$$

Then the empirical probability distribution on the bundle $\pi\colon d(1) \to d(0)$ agrees exactly with the given distribution on S. As a Dirichlet polynomial, this bundle is given (up to relabelling the outcomes) by $d(\mathrm{y}) := 15^{\mathrm{y}} + 6^{\mathrm{y}} + 5^{\mathrm{y}} + 4^{\mathrm{y}}$.

Note that any multiple $m^{\mathrm{y}} \cdot d(\mathrm{y})$ of d (for $m \geqslant 1$) will correspond to the same probability distribution as d itself, but to a different empirical distribution, since it will have m times as many draws.

Under this interpretation of Dirichlet polynomials as empirical distributions, multiplication $d \cdot e$ corresponds to taking the *product distribution*.

Remark 1. *For any $d \in \mathsf{Dir}$, and any $n \in \mathbb{N}$, we can give $|d(n)|$ a combinatorial interpretation: it is the number of ways of choosing n indistinguishable (in the sense that they have the same outcome) draws, i.e., the number of length-n lists of elements of $d[i]$ for some $i \in d(0)$.*

To see this, note that $d(n) = \mathtt{Dir}\,(n^{\mathrm{y}}, d)$ (by Yoneda), and so $d(n)$ is in bijection with the set of bundle morphisms $(\varphi_0, \varphi_1)\colon (n \to 1) \to (d(1) \to d(0))$, which are given exactly by choosing n (possibly repeated) elements of $d(1)$ that all lie in the same fibre (namely the fibre above the point specified by $\varphi_0(1)$).

Remark 2. *Although we deal only with* finite *sets and* rational *probability distributions here, it seems likely that one could follow the methods of [3] and consider colimits of these to obtain analogous results for* arbitrary *probability distributions on* discrete measurable spaces*.*

4. Area and Width

Definition 5. *Define the rig* Rect *as follows. The underlying set is $\mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0}$. The multiplicative structure has unit $(1,1)$, and is given by component-wise multiplication:*

$$(A_1, W_1) \cdot (A_2, W_2) := (A_1 A_2, W_1 W_2).$$

The additive structure has unit $(0,0)$, and is given by real-number addition in the first component, and by weighted geometric mean in the second component:

$$(A_1, W_1) + (A_2, W_2) := \left(A_1 + A_2, \left(W_1^{A_1} W_2^{A_2}\right)^{\frac{1}{A_1 + A_2}} \right).$$

Given an element (A, W) in Rect*, we call A its* area *and W its* width*.*

The fact that Rect is indeed a rig follows from the fact that its multiplication distributes over its addition:

$$\begin{aligned}
(A_1, W_1) \cdot \Big((A_2, W_2) + (A_3, W_3) \Big) &= (A_1, W_1) \cdot \left(A_2 + A_3, \left(W_2^{A_2} W_3^{A_3}\right)^{\frac{1}{A_2 + A_3}} \right) \\
&= \left(A_1 (A_2 + A_3), W_1 \left(W_2^{A_2} W_3^{A_3}\right)^{\frac{1}{A_2 + A_3}} \right) \\
&= \left(A_1 A_2 + A_1 A_3, \left(W_1^{A_2 + A_3} W_2^{A_2} W_3^{A_3}\right)^{\frac{1}{A_2 + A_3}} \right) \\
&= \left(A_1 A_2 + A_1 A_3, \left((W_1 W_2)^{A_2} (W_1 W_3)^{A_3}\right)^{\frac{1}{A_2 + A_3}} \right) \\
&= \left(A_1 A_2 + A_1 A_3, \left((W_1 W_2)^{A_1 A_2} (W_1 W_3)^{A_1 A_3}\right)^{\frac{1}{A_1 A_2 + A_1 A_3}} \right) \\
&= (A_1 A_2, W_1 W_2) + (A_1 A_3, W_1 W_3) \\
&= (A_1, W_1) \cdot (A_2, W_2) + (A_1, W_1) \cdot (A_3, W_3).
\end{aligned}$$

Proposition 1. *There exists a unique rig morphism $h\colon \mathsf{Dir} \to \mathsf{Rect}$ for which*

$$h\colon n^{\mathrm{y}} \mapsto (n, n).$$

Proof. Since every Dirichlet polynomial is just a sum of exponentials, a rig homomorphism is fully determined by its action on exponentials, since it must respect addition. So we just need to show that h does indeed extend to a rig homomorphism, but this follows from the fact that $m^{\mathrm{y}} \cdot n^{\mathrm{y}} = (m \cdot n)^{\mathrm{y}}$. □

Definition 6. *Given a Dirichlet polynomial d, we define its* area $A(d)$ *and its* width $W(d)$ *to be given by the components of $h(d) = (A(d), W(d))$.*

With this definition, along with Proposition 1, we see that

$$A(n^{\mathrm{y}}) = W(n^{\mathrm{y}}) = n.$$

Lemma 3. *Let $d \in \mathsf{Dir}$ and $a \in \mathbb{N}$. Then*

1. $h(a) = (a, 1)$;
2. $A(a \cdot d) = aA(d)$;
3. $W(a \cdot d) = W(d)$.

Proof. Recall that addition (and thus scalar multiplication) in Rect involves the weighted geometric mean in the second component. Then

$$\begin{aligned} h(a) :=& h(a \cdot 1^{\mathrm{y}}) \\ =& a \cdot h(1^{\mathrm{y}}) \\ =& a \cdot (1, 1) \\ =& (a, 1) \end{aligned}$$

which proves (i). For (ii) and (iii), since h is a rig homomorphism (and thus respects addition), it suffices to consider the case where d is an exponential, say $d(\mathrm{y}) = n^{\mathrm{y}}$. However, then

$$\begin{aligned} h(a \cdot d) &= a \cdot h(d) \\ &= a \cdot (n, n) \\ &= \left(an, (\underbrace{n^n n^n \ldots n^n}_{a \text{ times}})^{1/an}\right) \\ &= \left(an, (n^{an})^{1/an}\right) \\ &= (an, n) \end{aligned}$$

i.e., $A(a \cdot d) = aA(d)$ and $W(a \cdot d) = W(d)$, as claimed. □

Corollary 2. *Let $d \in \mathsf{Dir}$. Then*

$$\begin{aligned} A(d) &= |d(1)| \\ W(d)^{A(d)} &= \left|\mathtt{Dir}_{/d(0)}(d, d)\right|. \end{aligned}$$

Proof. Write $d(\mathrm{y}) = a_n \cdot n^{\mathrm{y}} + \ldots + a_1 \cdot 1^{\mathrm{y}} + a_0 \cdot 0^{\mathrm{y}}$. Using Lemma 3, along with Definition 5, we see that

$$\begin{aligned} h(d) &= (a_n n, n) + (a_{n-1}(n-1), n-1) + \ldots + (a_1, 1) \\ &= \left(\sum_{i=0}^{n} a_i i, \left(\prod_{i=0}^{n} i^{a_i i}\right)^{1/\sum_{i=0}^{n} a_i i}\right). \end{aligned}$$

By Lemma 1, the first component (i.e., $A(d)$) is equal to $|d(1)|$; by the same lemma, we can also rewrite the second component (i.e., $W(d)$) as

$$W(d) = \left(\prod_{i=0}^{n} i^{a_i i}\right)^{\frac{1}{A(d)}}$$

so it simply remains to justify why this is equal to $\left|\mathtt{Dir}_{/d(0)}(d,d)\right|^{\frac{1}{A(d)}}$. But a morphism in $\mathtt{Dir}_{/d(0)}(d,d)$ is exactly the data of an endomorphism of each fibre of $\pi_d \colon d(1) \to d(0)$; since there are a_i fibres of size i, endomorphisms of these fibres are in bijection with the a_i-fold product of i^i, which is equal to $i^{a_i i}$, whence the claim. □

Corollary 3. *Let $d \in \mathsf{Dir}$. Then the width $W(d)$ is an algebraic number, i.e., the image of $h \colon \mathsf{Dir} \to \mathsf{Rect}$ lies in the sub-rig whose underlying set is $\mathbb{N} \times \overline{\mathbb{Q}}_{\geqslant 0}$, where $\overline{\mathbb{Q}}$ is the algebraic closure of $\mathbb{Q}$.*

Proof. By Corollary 2, both $W(d)^{A(d)}$ and $A(d)$ are equal to the cardinality of some sets, and thus integer. □

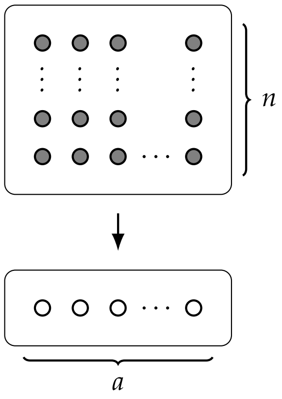
Example 5. *Reassuringly, if we start with a "rectangle", then the area and width are exactly what we might expect. More concretely: consider $d(\mathrm{y}) = a \cdot n^{\mathrm{y}}$ for some $a, n \in \mathbb{N}$; then, by Corollary 2,*

$$A(d) = d(1) = an,$$

and, by direct calculation,

$$W(d) = n.$$

Comparing this to the picture of $a \cdot n^{\mathrm{y}}$, we can explain why we chose the terminology "width" and "area":

$a \cdot n^{\mathrm{y}} =$ 

Indeed, the area is exactly the number of dots in the (upper) rectangle, and the width is its width.

However, this picture now leads us to consider the question of whether or not there is a good meaning we can give to the "length" of this rectangle (which, here, should be equal to a). Indeed, this has been our motivation all along; we will return to this question in Example 9.

Example 6. *The fact that $d(1)$ is "rectangular" in Example 5 makes the terminology look like a numerical coincidence, but we can try to hone our intuition of what this really "means" by considering another example.*

Let's consider $d(\mathrm{y}) = 4^{\mathrm{y}} + 4$, which has area $A(d) = d(1) = 8$. We can calculate its width by using the fact that

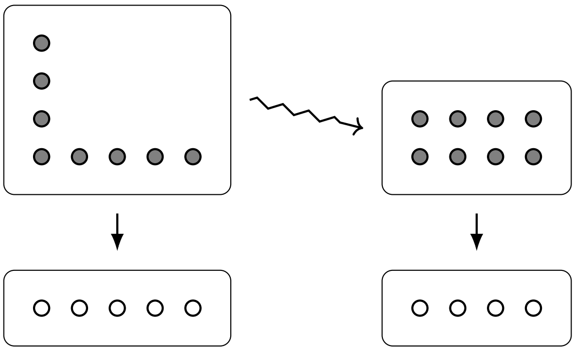
$$\begin{aligned} h(4^{\mathrm{y}}) &= (4,4) \\ h(4) &= h(1^{\mathrm{y}}) + h(1^{\mathrm{y}}) + h(1^{\mathrm{y}}) + h(1^{\mathrm{y}}) \\ &= (4,1) \end{aligned}$$

whence

$$\begin{aligned} h(d) &= (4,4) + (4,1) \\ &= \left(8, (4^4 1^4)^{\frac{1}{8}}\right) \\ &= (8,2) \end{aligned}$$

and so $W(d) = 2$*.*

How, then, does the rectangle with area 8 *and width* 2 *relate to our Dirichlet polynomial* $d(\mathrm{y}) = 4^{\mathrm{y}} + 4$*? That is, what is the process that takes us from* d *to* $4 \cdot 2^{\mathrm{y}}$*? Looking at the pictures of the bundles, we see that the width tells us how our bundle would look if we had the same set (*$d(1)$*) of draws, but with different outcomes, now all* equally likely*:*



Note that, in order to have equally sized fibres, we needed to have 4 *outcomes, not* 5 *(since* $8/2 = 4$*). We make this idea more precise (as well as explain why the rectangle is of size* 4×2 *instead of* 2×4*) in Section 6.*

Example 7. *We have just seen that* $d(\mathrm{y}) = 4^{\mathrm{y}} + 4$ *has* $W(d) = 2$ *and* $A(d) = 8$*, but now let's look at an example where the numbers don't divide so neatly.*

Let $d(\mathrm{y}) = 4^{\mathrm{y}} + 3$*. Then* $A(d) = d(1) = 7$*, and, as in Example 6, we use the fact that*

$$\begin{aligned} h(d) &= (4,4) + (3,1) \\ &= \left(7, (4^4 1^3)^{\frac{1}{7}}\right) \\ &= (7, 2\sqrt[7]{2}) \\ &\approx (7, 2.21) \end{aligned}$$

Of course, now we can't draw a nice rectangle representing the evenly distributed bundle as we did in Example 6 for $4^{\mathrm{y}} + 4$*, since we would have to have an outcome set of size* $7/2.21 \approx 3.17$ *elements, with fibres all of size* 2.21*, but this should come as no surprise, since* 7 *is prime. One might be tempted to solve this problem using groupoid cardinality (cf. [4]), but there are some technical issues here.*

5. Length

Remark 3. *We write* $\log$ *to mean* $\log_2$*.*

Definition 7. *Given a Dirichlet polynomial* $d(\mathrm{y}) := \sum_{i \in d(0)} d[i]^{\mathrm{y}}$*, we define its* entropy $H(d)$ *by*

$$H(d) := -\sum_{i \in d(0)} \frac{|d[i]|}{|d(1)|} \log\left(\frac{|d[i]|}{|d(1)|}\right).$$

We then define its length *(also called the* perplexity*)* $L(d)$ *by*

$$L(d) := 2^{H(d)}.$$

Readers might recognise $H(d)$ as being the *Shannon entropy* of the corresponding probability distribution (cf. [5]). The convention for naming the sides of a rectangle is from [6] (Figure 1).

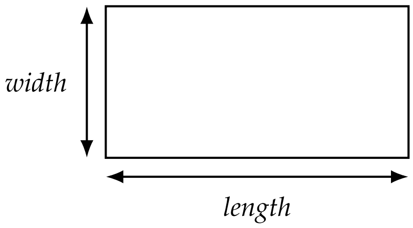
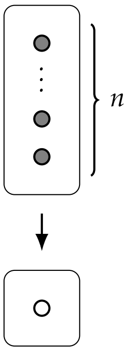


Figure 1. Our convention for naming the sides of a rectangle, from [6].

Example 8. *Consider $d(\mathrm{y}) = n^{\mathrm{y}}$ for some $n \in \mathbb{N}$.*



Then $d(0) = 1$ and $d(1) = n$, and so

$$\begin{aligned} H(d) &= -\sum_{i \in \underline{1}} \frac{n}{n} \log\left(\frac{n}{n}\right) \\ &= -\log 1 \\ &= 0 \end{aligned}$$

whence $L(d) = 2^0 = 1$.

In terms of distributions, this corresponds to the fact that the unique probability distribution on a single outcome has entropy equal to 0 (and so the same is true for any empirical distribution on a single outcome).

Example 9. *Continuing on from Example 5, we can calculate the entropy of a uniform distribution on a many outcomes $d(\mathrm{y}) = a \cdot n^{\mathrm{y}}$ as*

$$\begin{aligned} H(a \cdot n^{\mathrm{y}}) &= -\sum_{i \in \underline{a}} \frac{n}{an} \log\left(\frac{n}{an}\right) \\ &= -\log\left(\frac{1}{a}\right) \\ &= \log a \end{aligned}$$

whence $L(a \cdot n^{\mathrm{y}}) = 2^{\log a} = a$, exactly as desired.

Example 10. *Continuing on from Example 7, recall that* $d(\mathrm{y}) = 4^{\mathrm{y}} + 4$ *has area* $A(d) = 8$ *and width* $W(d) = 2$*. We can further calculate that*

$$\begin{aligned} H(d) &= -\sum_{i \in \underline{5}} \frac{|d[i]|}{8} \log\left(\frac{|d[i]|}{8}\right) \\ &= -\frac{4}{8}\log\left(\frac{4}{8}\right) - 4 \cdot \frac{1}{8}\log\left(\frac{1}{8}\right) \\ &= -\frac{1}{2}\log\left(\frac{1}{16}\right) \\ &= 2 \end{aligned}$$

whence $L(d) = 2^2 = 4$*.*

As for $d(\mathrm{y}) = 4^{\mathrm{y}} + 3$*, recall that its area is* $A(d) = 7$ *and its width is* $2\sqrt[7]{2}$*. Now, its entropy is*

$$\begin{aligned} H(d) &= -\sum_{i \in \underline{4}} \frac{|d[i]|}{7} \log\left(\frac{|d[i]|}{7}\right) \\ &= -\frac{4}{7}\log\left(\frac{4}{7}\right) - 3 \cdot \frac{1}{7}\log\left(\frac{1}{7}\right) \\ &= \frac{\log 7}{\log 2} - \frac{8}{7} \end{aligned}$$

and so its length is

$$L(d) = 2^{\frac{\log 7}{\log 2} - \frac{8}{7}} = \frac{7}{2\sqrt[7]{2}}.$$

Note that, in the above example, even though both $L(d)$ and $W(d)$ have non-integer values, the formula implied by our choice of nomenclature still holds: the area $A(d)$ is equal to the length $L(d)$ times the width $W(d)$. This leads us to our main theorem. It says that the Shannon entropy, which is only homomorphic in products of distributions, can be computed in terms of the width and area, which together are homomorphic in both sums and products of distributions. We will explain this in more detail in Section 6.

Theorem 1. *For all* $d \in \mathsf{Dir}$*, we have the* rectangle-area formula

$$A(d) = L(d)W(d).$$

Proof. In the following, we omit absolute value signs, writing e.g., $d(1)$ instead of $|d(1)$.

First, write

$$d(\mathrm{y}) := \sum_{j=0}^{n} a_j j^{\mathrm{y}} \cong \sum_{i \in d(0)} d[i]^{\mathrm{y}}.$$

Now we can rewrite the length as

$$\begin{aligned}
L(d) &= 2^{H(d)} \\
&= 2^{-\sum_{i\in d(0)} \frac{d[i]}{d(1)} \log\left(\frac{d[i]}{d(1)}\right)} \\
&= \prod_{i\in d(0)} 2^{-\frac{d[i]}{d(1)} \log\left(\frac{d[i]}{d(1)}\right)} \\
&= \prod_{i\in d(0)} \left(2^{-\log\left(\frac{d[i]}{d(1)}\right)}\right)^{\frac{d[i]}{d(1)}} \\
&= \prod_{i\in d(0)} \frac{d(1)}{d[i]}^{\frac{d[i]}{d(1)}} \\
&= \frac{\prod_{i\in d(0)} d(1)^{\frac{d[i]}{d(1)}}}{\prod_{i\in d(0)} d[i]^{\frac{d[i]}{d(1)}}}
\end{aligned}$$

The numerator is then

$$\begin{aligned}
\prod\nolimits_{i\in d(0)} d(1)^{\frac{d[i]}{d(1)}} &= d(1)^{\sum_{i\in d(0)} \frac{d[i]}{d(1)}} \\
&= d(1) \\
&= A(d)
\end{aligned}$$

since $\sum_{i\in d(0)} |d[i]| = |d(1)|$, by Corollary 1, and we can then apply Corollary 2. The denominator is exactly

$$\left(\prod\nolimits_{i\in d(0)} d[i]^{d[i]}\right)^{\frac{1}{d(1)}}$$

and so, by Corollary 2, we only need to justify why $\prod_{i\in d(0)} d[i]^{d[i]}$ is equal to $|\mathtt{Dir}_{/d(0)}(d,d)|$. However, this follows from the definition of an element of the latter set: a choice of map $d[i] \to d[i]$ for all $i \in d(0)$. □

6. Interpreting Area, Length, and Width

We have mentioned many times that Dirichlet polynomials are equivalent to set-theoretic bundles, so the natural question to ask is "*why, then, should we work with the former instead of the latter?*". One answer to this is question is the fact that *entropy does not respect bundle morphisms*: we cannot functorially assign a morphism between entropies to morphisms, since we are working with **arbitrary** morphisms of bundles. (If, however, we restrict to only morphisms given by pushforward, then [7] tells us (via *Faddeev's theorem*) that the only possible functorial definition of entropy is given by the *relative entropy*, i.e., the difference of the entropies of the source and the target). This makes it seem rather bad to work with a *category* (such as that of bundles) instead of simply a *rig* (such as that of Dirichlet polynomials). Of course, this isn't an entirely satisfactory answer, since we *do* care about the notion of morphisms for Dirichlet polynomials (for example, Corollary 2 tells us that the width can be expressed in terms of the number of certain morphisms). In light of Theorem 1, however, we might consider the following possibility: both area and length can be expressed in terms of $d(0)$, $d(1)$, and $d[i]$ (for $i \in d(0)$), and we could *define* the width by $W(d) := A(d)/L(d)$.

A better answer to this question might be the following: the rig homomorphism $h\colon \mathsf{Dir} \to \mathsf{Rect}$ is incredibly simple, since it just maps n^{y} to (n,n); from this computationally simple homomorphism, however, we can recover entropy (as $\log(A(d)/W(d))$), without making any reference to the classical equation that defines it ("negative the sum of probabilities of the log of the probabilities"), but instead relying on the fact that Rect encodes the weighted geometric mean. That is, $H(d)$ is only homomorphic in the product of distributions, whereas the pair $(A(d), W(d))$ is homomorphic in both the product and the sum.

Now, the entropy $H(d) = \log L(d)$ can be understood (via Huffman coding, cf. [8]) as *the average number of bits needed to code a single outcome* (over a long enough message). What is also true, however, is that the width (which is obtained purely "algebraically", i.e., from the rig homomorphism $h\colon \mathsf{Dir} \to \mathsf{Rect}$) gives similar information: by Theorem 1, combined with the previous sentence, $\log W(d)$ is *the average number of bits needed to code the draw, given an outcome* (in the same Huffman coding as before). This answers the question of "*what is special about the bundle defined by the width and length*" with "*it describes the optimal encoding of draws, given outcomes*".

As for the picture in Example 6, we can now understand the hand-wavy explanation a bit better (but still just as hand-wavy-ly): we take our original "half-filled" rectangle $d(1)$ and pour its contents into a new rectangle, of length $L(d)$, and then "slosh the contents around" until they lie flat, and then put a lid on it; the rectangle will be perfectly filled up, and the placement of the lid will be given by $W(d)$.

We also mentioned, in Corollary 3, that the width $W(d)$ of any Dirichlet polynomial d is an algebraic number, but the actual result is slightly more interesting that this: Corollary 2 tells us that $W(d)^{A(d)}$ is equal to the cardinality of the set $\mathtt{Dir}_{/d(0)}(d, d)$. We already know how to understand endomorphisms of d that fix $d(0)$ as endomorphisms of $d(1)$ that fix the outcome; we can understand $W(d)^{A(d)}$ as maps from $A(d)$ to $W(d)$; roughly speaking, such a map $f\colon A(d) \to W(d)$ determines the remaining ambiguity in determining a draw, given its outcome.

7. Cross Entropy

Everything above can be viewed as a specific example of the analogous *cross* notions. That is, given two Dirichlet polynomials, we can define their cross area, cross width, etc. as follows.

Definition 8. *Let $d, e \in \mathsf{Dir}$ be Dirichlet polynomials such that $d(0) = e(0)$. Then we define the* cross entropy *$H(d, e)$ by*

$$H(d,e) = -\sum_{i \in d(0)} \frac{|d[i]|}{|d(1)|} \log\left(\frac{|e[i]|}{|e(1)|}\right)$$

and the cross area, cross width*, and* cross length *by*

$$\begin{aligned} A(d,e) &:= |e(1)| \\ W(d,e) &:= \left|\mathtt{Dir}_{/d(0)}(d,e)\right|^{\frac{1}{|d(1)|}} \\ L(d,e) &:= 2^{H(d,e)} \end{aligned}$$

(respectively).

By definition, $X(d, d) = X(d)$ for $X \in \{A, W, H, L\}$. That is, just as cross entropy is a generalisation of entropy, the notions of cross width, etc., generalise the notions of width, etc.

Remark 4. *Note that we can recover the notion of* relative entropy *(also known as* Kullback–Leibler divergence*) $D_{\mathrm{KL}}(p\|q)$, as studied in [9], from cross entropy:*

$$H(d,e) = H(d) + D_{\mathrm{KL}}(p\|q)$$

(which can also be seen to justify the fact that $H(d, d) = H(d)$).

Remark 5. *Although we have some idea of how to understand these cross notions (e.g., cross area can be understood as the number of "actual" draws, when we think of d as being a potentially inaccurate model for e), the choice of definitions in Definition 8 was chosen simply so that*

1. *we recover the "uncrossed" notions when we take $d = e$, and*
2. *Theorem 2 holds.*

Theorem 2. *For all $d, e \in \mathsf{Dir}$, we have the* cross rectangle-area formula

$$A(d,e) = L(d,e)W(d,e).$$

Proof. This proof follows exactly the same argument as the proof of Theorem 1. □

Author Contributions: Writing—original draft, D.I.S. and T.H. Both authors have read and agreed to the published version of the manuscript.

Funding: The first author acknowledges support from AFOSR grant no. FA9550-20-1-0348.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spivak, D.I.; Myers, D.J. Dirichlet Polynomials form a Topos. *arXiv* **2020**, arXiv:2003.04827.
2. Leinster, T. *Basic Category Theory*; Cambridge University Press: Cambridge, UK, 2014.
3. Fritz, T.; Perrone, P. A probability monad as the colimit of spaces of finite samples. *Theory Appl. Categ.* **2019**, *34*, 170–220.
4. Baez, J.C.; Hoffnung, A.E.; Walker, C.D. Higher-Dimensional Algebra VII: Groupoidification. *Theory Appl. Categ.* **2010**, *24*, 489–553.
5. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
6. KidzSearch Wiki. Rectangle Facts for Kids. Available online: https://wiki.kidzsearch.com/wiki/Rectangle (accessed on 18 August 2021)
7. Baez, J.C.; Fritz, T.; Leinster, T. A Characterization of Entropy in Terms of Information Loss. *Entropy* **2011**, *13*, 1945–1957. [CrossRef]
8. Huffman, D.A. A method for the construction of minimum-redundancy codes. *Proc. IRE* **1952**, *40*, 1098–1101. [CrossRef]
9. Baez, J.C.; Fritz, T. A Bayesian characterization of relative entropy. *Theory Appl. Categ.* **2014**, *29*, 421–456.