Entropy as an Operad Derivation

Tai-Danae Bradley
SandboxAQ · The Master's University

May 26, 2022

a description of Shannon entropy in the language of operads

- preliminaries
- result
- related works ★

a description of Shannon entropy in the language of operads

- preliminaries
- result
- related works ★

The Shannon entropy of a finite probability distribution $p = (p_1, \ldots, p_n)$ is

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

It gives rise to a sequence of real-valued continuous functions on topological simplices,

$$\{H \colon \Delta_n \to \mathbb{R}\}_{n \geq 1},$$

where

$$\Delta_n := \{(p_1, \ldots, p_n) \mid p_i \geq 0 \text{ and } \sum_i p_i = 1\}.$$

a description of
Shannon entropy in
the language of operads

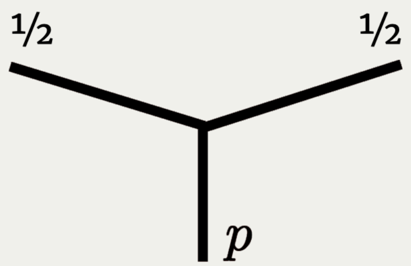
- preliminaries
- result
- related works ★

The chain rule

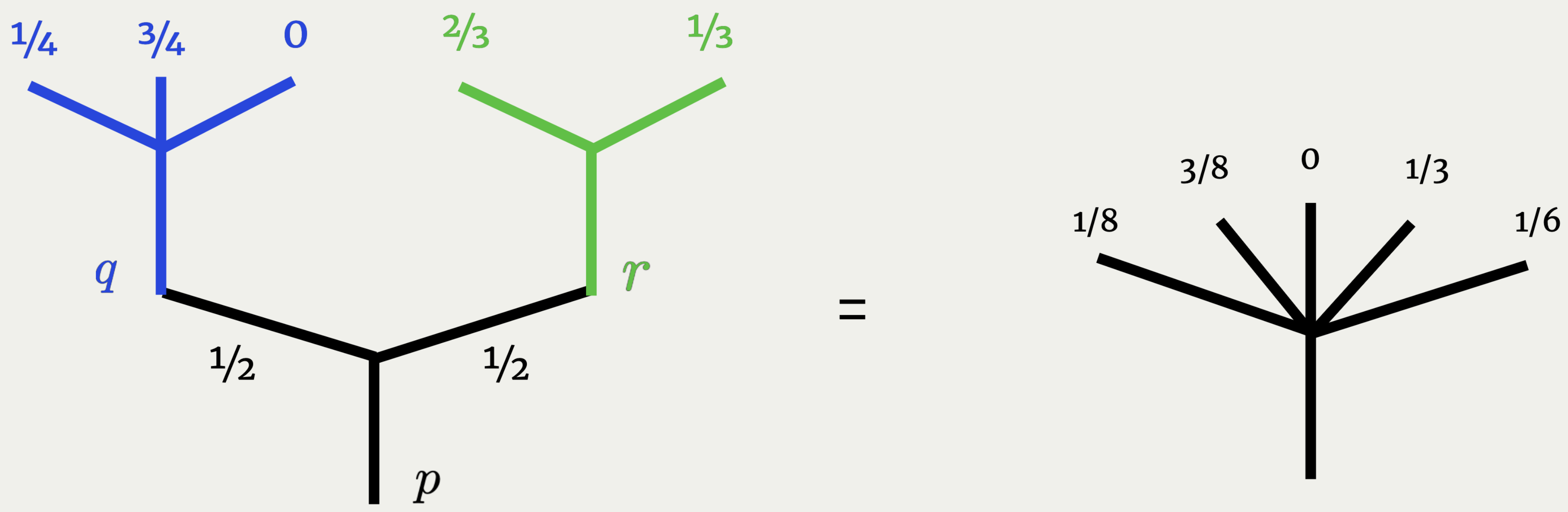
The chain rule tells us how entropy behaves when probability distributions are combined in a certain way. Pictures help illustrate this.

Composing probabilities

It's convenient to illustrate a probability distribution $p = (\frac{1}{2}, \frac{1}{2})$ as a tree:



Since p has two leaves, it may be composed with two probability distributions by multiplication.



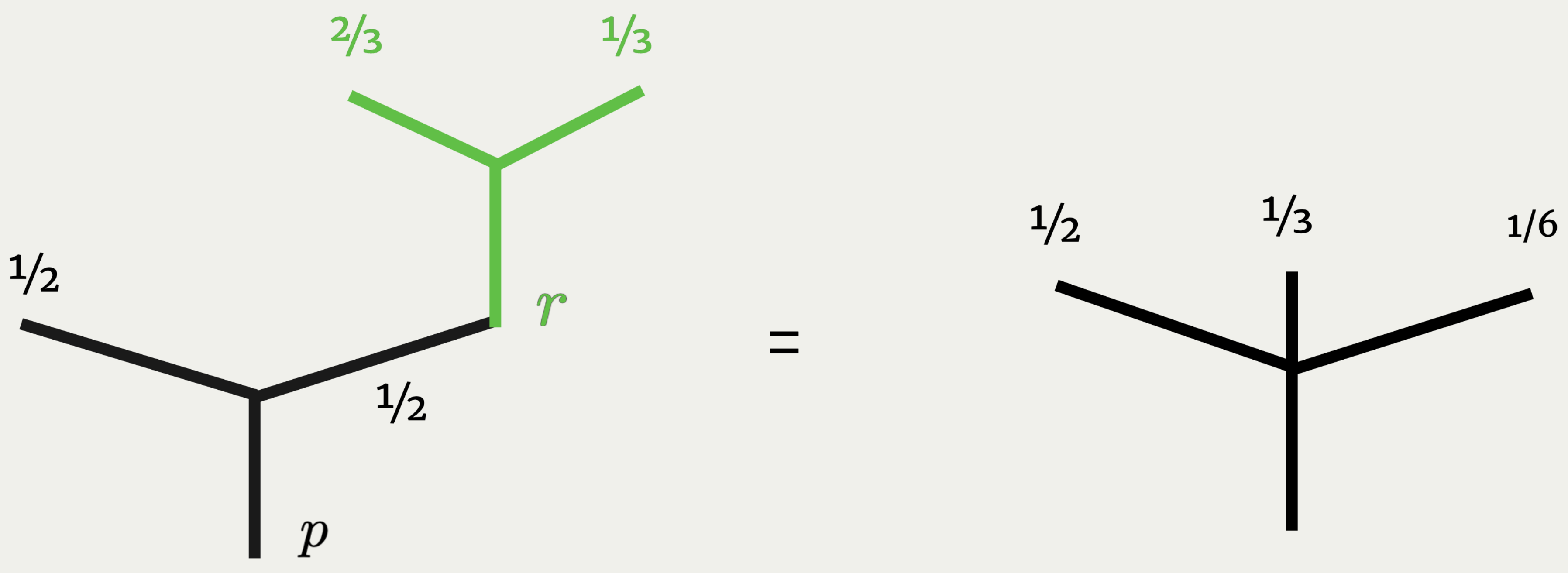
$$p \circ (q, r) \quad = \quad \left(\tfrac{1}{8}, \tfrac{3}{8}, 0, \tfrac{1}{3}, \tfrac{1}{6}\right) \in \Delta_5$$

This composition defines a map on simplicies

$$\circ\colon \Delta_2 \times \Delta_3 \times \Delta_2 \to \Delta_{2+3}.$$

More generally, any probability distribution on n elements may be composed with n probability distributions in a similar way.

We can also compose one leaf at a time.



$$p \circ_2 r = \left(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}\right) \in \Delta_3$$

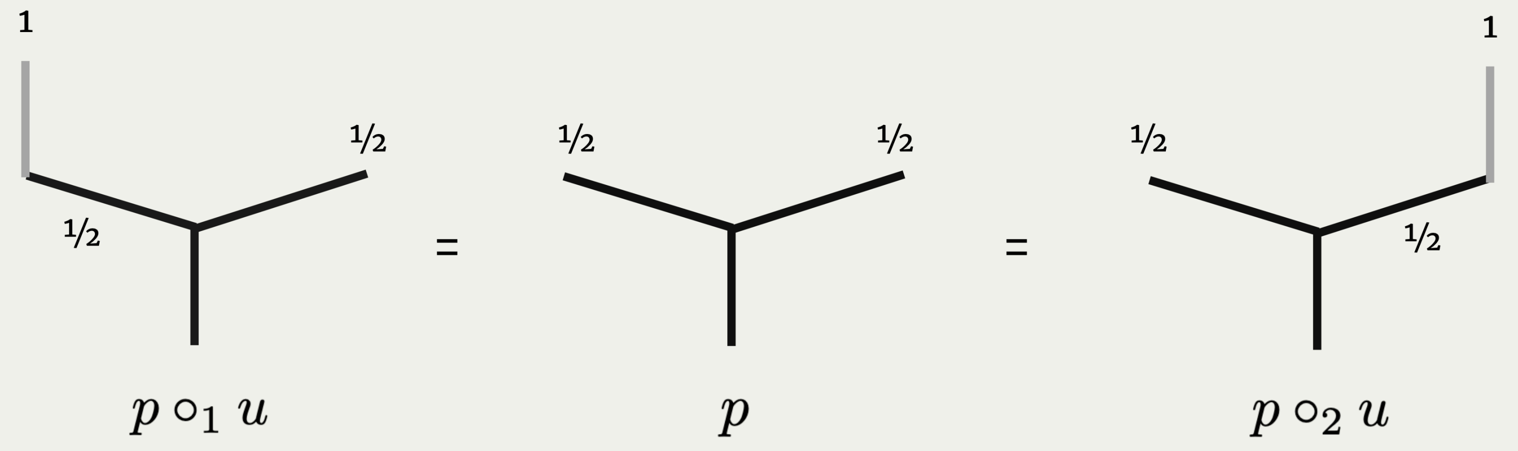
This partial composition defines two maps on simplicies

$$\circ_i \colon \Delta_2 \times \Delta_2 \to \Delta_{(2-1)+2}\,,$$

where $i = 1, 2$.

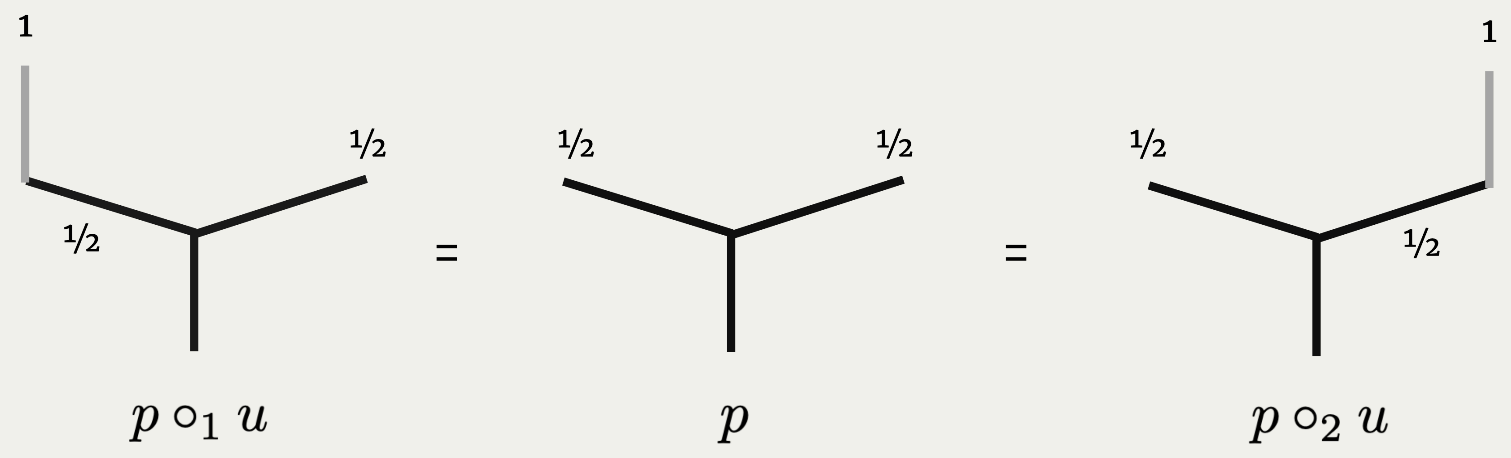
Identity

The distribution on a single element $u = (1) \in \Delta_1$ serves as an identity for this algebraic structure:



Identity

The distribution on a single element $u = (1) \in \Delta_1$ serves as an identity for this algebraic structure:



Punchline

Topological simplices $\Delta_1, \Delta_2, \Delta_3, \ldots$ together with composition maps

$$\circ_i \colon \Delta_n \times \Delta_m \to \Delta_{n+m-1}$$

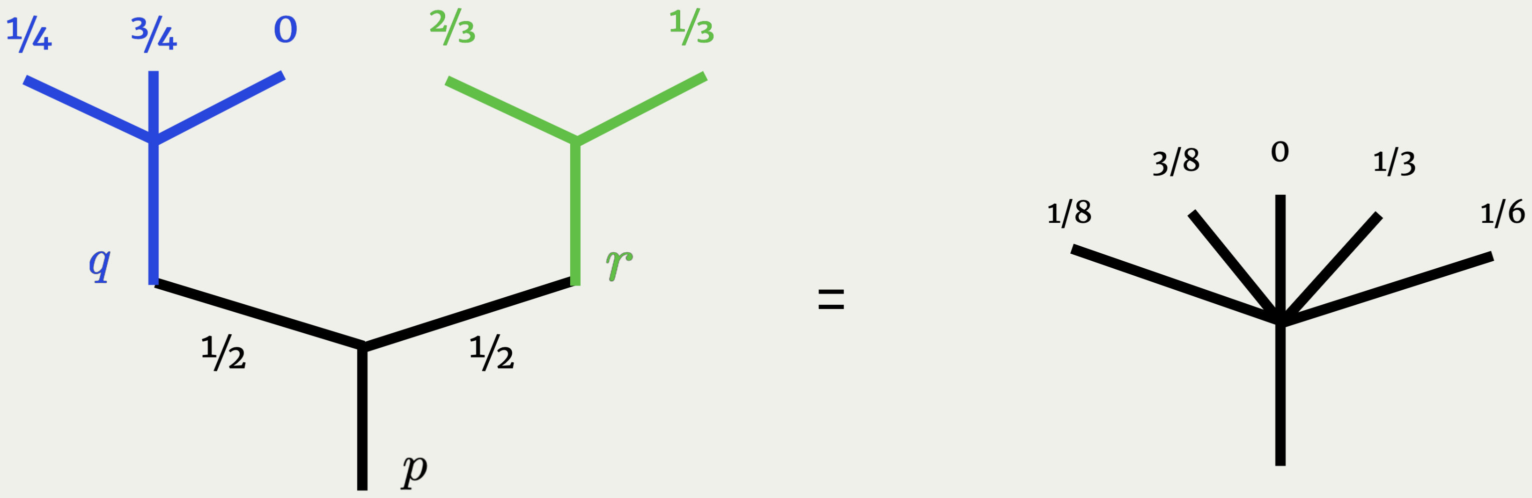
for $1 \leq i \leq n$ (and associativity and unital axioms) form a operad*, an abstract structure with origins in algebraic topology.

* Here, we mean a one object multicategory with no symmetric group action. For more, see chapter 12 of Tom Leinster's book Entropy and Diversity: An Axiomatic Approach.

The chain rule

The collection of continuous functions $\{H\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ satisfies the following rule:

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



$p \circ (q,r)$ = $\left(\frac{1}{8}, \frac{3}{8}, 0, \frac{1}{3}, \frac{1}{6}\right) \in \Delta_5$

The chain rule

The collection of continuous functions $\{H\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ satisfies the following rule:

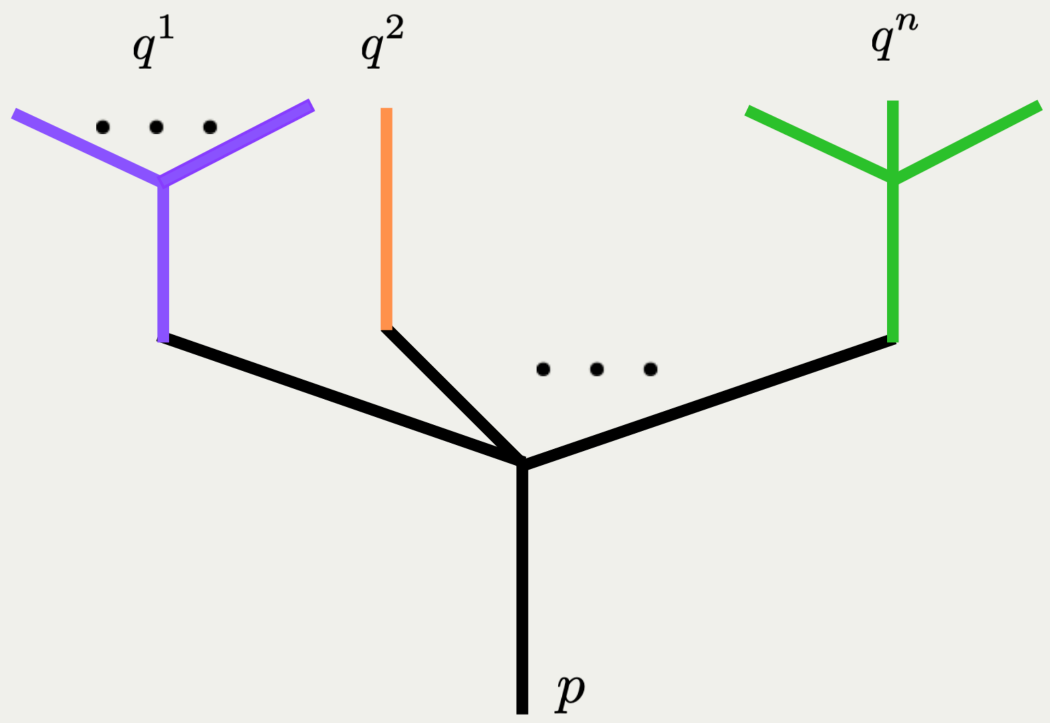
$$H(p \circ (q, r)) = H(p) + \tfrac{1}{2} H(q) + \tfrac{1}{2} H(r)$$

or more generally

$$H(p \circ (q^1, q^2, \ldots, q^n)) = H(p) + \sum_{i=1}^{n} p_i H(q^i)$$

for any probability distributions $p \in \Delta_n$ and $q_i \in \Delta_{k_i}$.

This is "the most important algebraic property of Shannon entropy" (Tom Leinster, Entropy and Diversity: The Axiomatic Approach).



$p \circ (q^1, q^2, \ldots, q^n)$

Theorem (Leinster, '21).

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

1. The functions I
   - are continuous
   - satisfy the chain rule.
2. $I = cH$ for some $c \in \mathbb{R}$.

Theorem (Leinster, '21).

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

1. The functions I
   - are continuous
   - satisfy the chain rule.
2. $I = cH$ for some $c \in \mathbb{R}$.

Some History

- This is Theorem 2.5.1 of Tom's book Entropy and Diversity (2021).
- It's also used in an operadic and category theoretical characterization given by Tom around 2011 (see also Theorem 12.3.1).[1]
- Tom's operadic characterization motivated the 2011 Baez-Fritz-Leinster paper.[2]
- It is based on a 1950s theorem by D. Faddeev, which is similar to Shannon's original characterization.

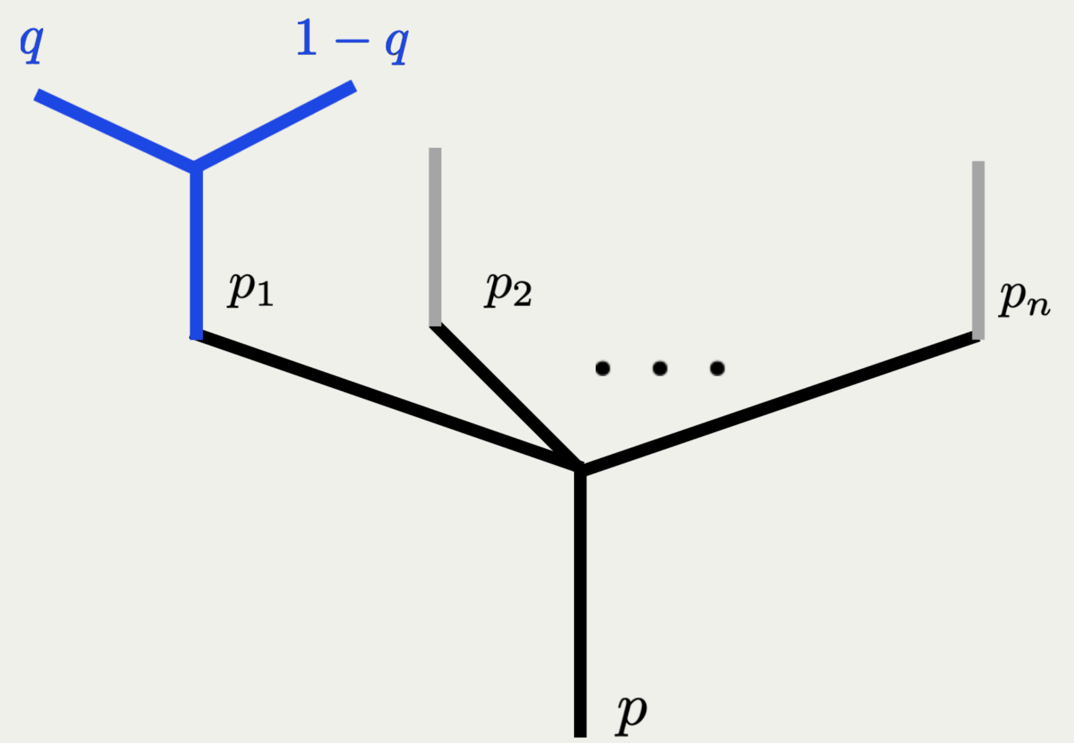
[1] If you're a veteran reader of the n-Category Café, you might remember this result from "An Operadic Introduction to Entropy" (Leinster, The n-Category Café, 2011).

[2] "A characterization of entropy in terms of information loss" Entropy 2011, 13(11), 1945-1957.

Theorem (Faddeev, '57).

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

1. The functions I
   - are continuous
   - satisfy a special case of the chain rule
   - are invariant under bijections
   - satisfy $I(\frac{1}{2}, \frac{1}{2}) = 1$.
2. $I = H$.



$$H(p \circ ((q, 1-q), u, \ldots, u) = H(p) + p_1 H(q, 1-q)$$

Theorem (Faddeev, '57).

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

1. The functions I
   - are continuous
   - satisfy a special case of the chain rule
   - are invariant under bijections
   - satisfy $I(\frac{1}{2}, \frac{1}{2}) = 1$.
2. $I = H$.

Theorem (Shannon, '49).[3]

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

1. The functions I
   - are continuous
   - satisfy the chain rule
   - are monotonic on uniform distributions.
2. $I = cH$ for some $c \in \mathbb{R}$.

[3] I've modified this slightly to draw a connection with Faddeev's and Leinster's theorems. For Shannon's original statement and proof, see Theorem 2 in "The Mathematical Theory of Communication."

"The Mathematical Theory of Communication"

(Shannon, 1949)

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} H(\tfrac{2}{3}, \tfrac{1}{3}).$$

The coefficient $\frac{1}{2}$ is the weighting factor introduced because this second choice only occurs half the time.

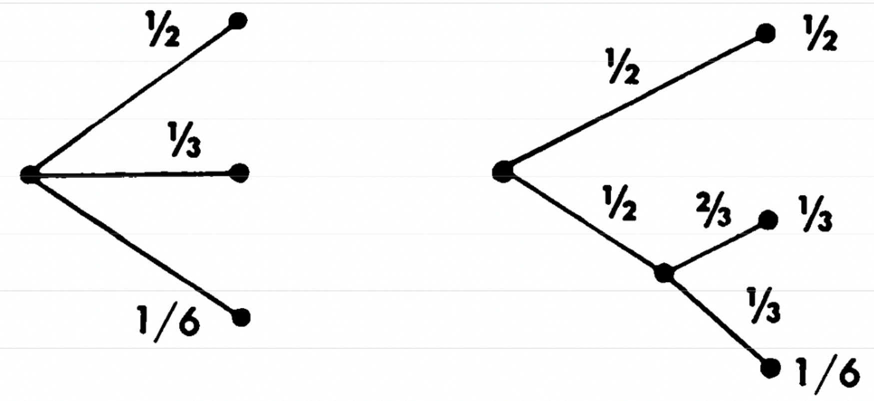


Fig. 6. — Decomposition of a choice from three possibilities.

The chain rule is the product rule

from the perspective of operads.

$$H(p \circ (q^1, \ldots, q^n)) = H(p) + \sum_i p_i H(q^i)$$

The chain rule is the product rule

from the perspective of operads.

In other words,

entropy $\leftrightarrows$ derivations of Δ

Idea:

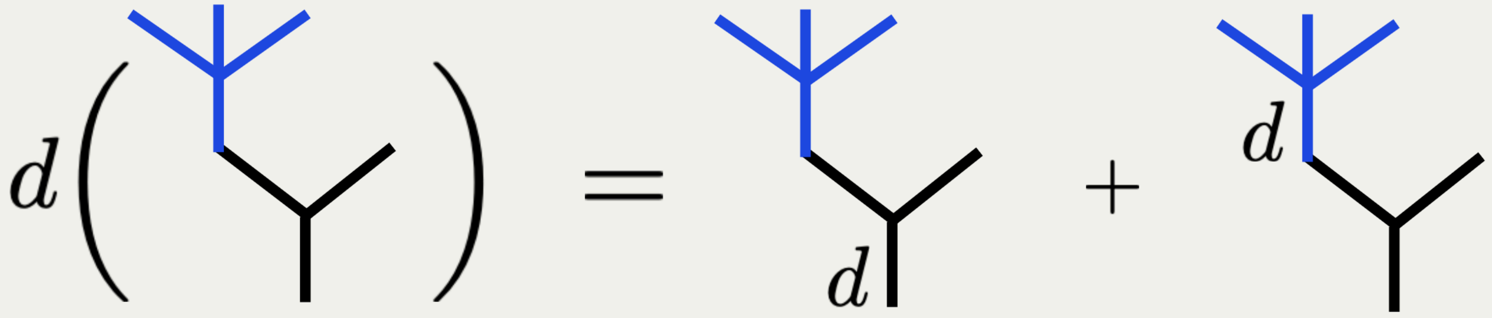
We wish to upgrade entropy $\{H\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ to a sequence of continuous functions

$$\{d\colon \Delta_n \to \blacksquare\}_{n\geq 1}$$

satisfying the product rule:

$$d(p \circ_i q) = dp \circ_i^R q + p \circ_i^L dq$$

for all $p \in \Delta_n$ and $q \in \Delta_m$ and $1 \leq i \leq n$ that recovers the chain rule, in some sense.



Idea:

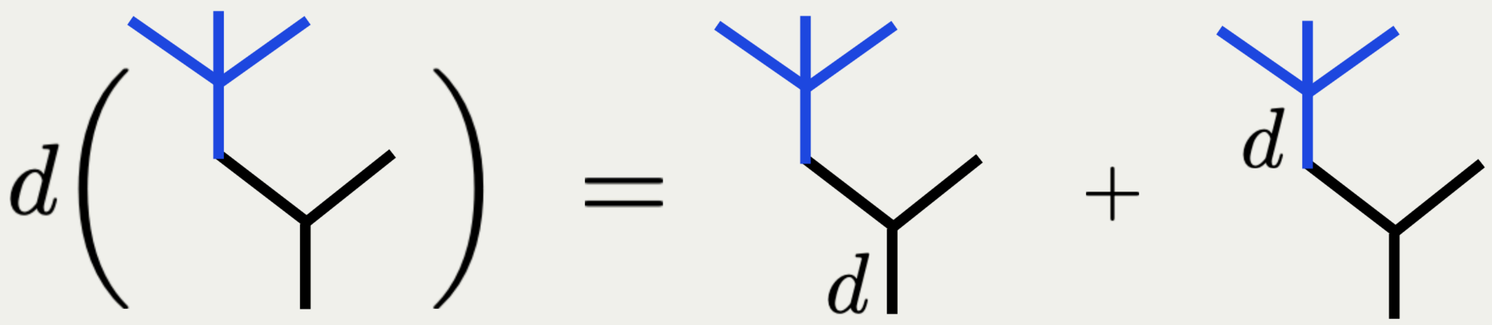
We wish to upgrade entropy $\{H: \Delta_n \to \mathbb{R}\}_{n\geq 1}$ to a sequence of continuous functions

$$\{d: \Delta_n \to \blacksquare\}_{n\geq 1}$$

satisfying the product rule:

$$d(p \circ_i q) = dp \circ_i^R q + p \circ_i^L dq$$

for all $p \in \Delta_n$ and $q \in \Delta_m$ and $1 \leq i \leq n$ that recovers the chain rule, in some sense.



Let $\blacksquare$ be the space of real-valued continuous functions on n-dimensional Euclidean space:

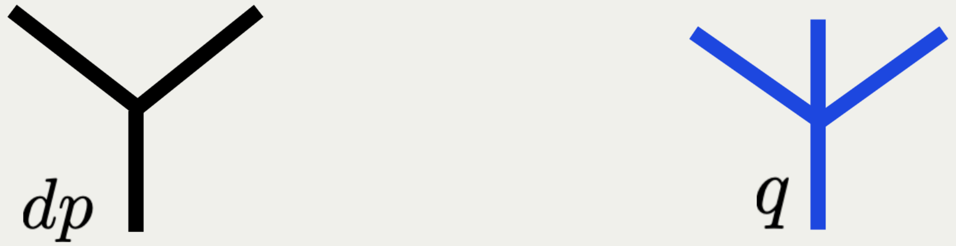
$$p \in \Delta_n \quad \mapsto \quad dp: \mathbb{R}^n \to \mathbb{R}$$

We will call such sequences a derivation of Δ.

Example: right composition $dp \circ_i^R q$

Let $p \in \Delta_2$ be a probability distribution on two elements, and let $dp\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. It can be convex combinations, or constant at entropy, or something else.

Suppose $q \in \Delta_3$. Let's say $q = (\frac{2}{3}, \frac{1}{3}, 0)$.

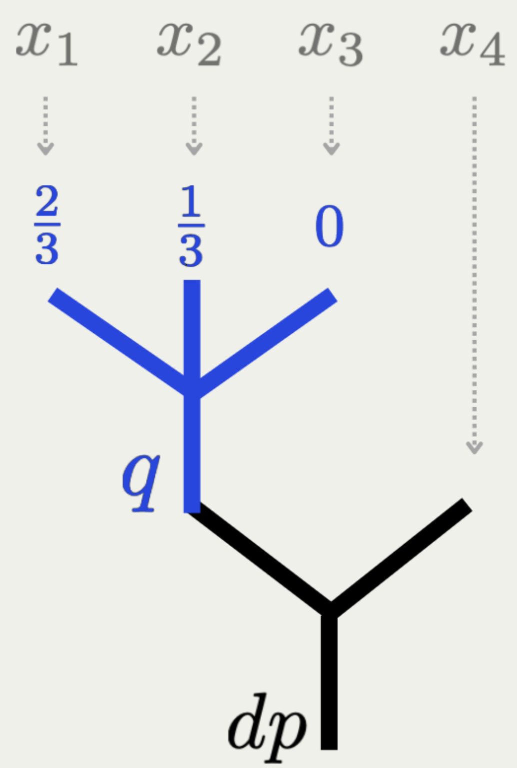


Example: right composition $dp \circ_i^R q$

Let $p \in \Delta_2$ be a probability distribution on two elements, and let $dp\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. It can be convex combinations, or constant at entropy, or something else.

Suppose $q \in \Delta_3$. Let's say $q = (\frac{2}{3}, \frac{1}{3}, 0)$. Define a new function $dp \circ_1^R q\colon \mathbb{R}^4 \to \mathbb{R}$ as follows:

$(dp \circ_1^R q)(\mathbf{x}) =$

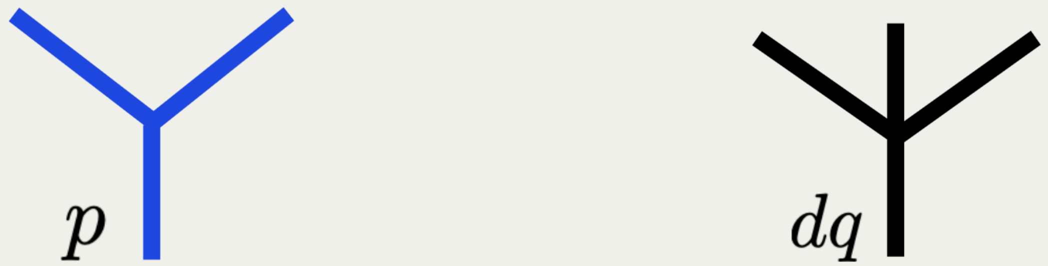


$= dp(\frac{2}{3}x_1 + \frac{1}{3}x_2 + 0x_3, x_4)$

Example: left composition $p \circ_i^L dq$

Given a probability distribution $q \in \Delta_3$, let $dq\colon \mathbb{R}^3 \to \mathbb{R}$ be a continuous function, e.g. convex combinations, or constant at entropy, or something else.

Suppose $p \in \Delta_2$, let's say $p = (\frac{3}{4}, \frac{1}{4})$.

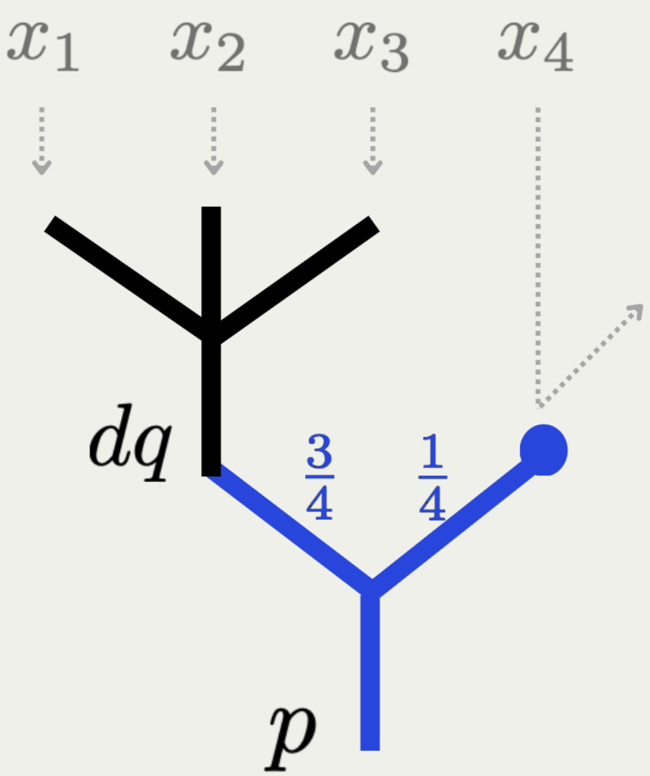


Example: left composition $p \circ_i^L dq$

Given a probability distribution $q \in \Delta_3$, let $dq\colon \mathbb{R}^3 \to \mathbb{R}$ be a continuous function, e.g. convex combinations, or constant at entropy, or something else.

Suppose $p \in \Delta_2$, let's say $p = (\frac{3}{4}, \frac{1}{4})$. Define a new function $p \circ_1^L dq\colon \mathbb{R}^4 \to \mathbb{R}$ as follows:

$(p \circ_1^L dq)(\mathbf{x}) =$



$= \frac{3}{4} dq(x_1, x_2, x_3)$

Putting it all together...

For each $n \geq 1$ let $\blacksquare = \mathbf{End}_n(\mathbb{R})$ denote the space of continuous functions $\mathbb{R}^n \to \mathbb{R}$ equipped with the product topology. We are interested in sequences of continuous functions

$$\{d \colon \Delta_n \to \mathbf{End}_n(\mathbb{R})\}_{n \geq 1}$$

where the codomain is equipped with continuous right/left composition maps:

$$\circ_i^R \colon \mathbf{End}_n(\mathbb{R}) \times \Delta_m \to \mathbf{End}_{n+m-1}(\mathbb{R}) \qquad \circ_i^L \colon \Delta_n \times \mathbf{End}_m(\mathbb{R}) \to \mathbf{End}_{n+m-1}(\mathbb{R})$$

for all $n, m \geq 1$ and $1 \leq i \leq n$, that satisfy the product rule:

$$d(p \circ_i q) = dp \circ_i^R q + p \circ_i^L dq.$$

Shannon entropy defines a derivation.

Shannon entropy defines a sequence of ds satisfying the product rule: $d(p \circ_i q) = dp \circ_i^R q + p \circ_i^L dq$.

Let $n \geq 1$ and for each probability distribution $p \in \Delta_n$ let $dp\colon \mathbb{R}^n \to \mathbb{R}$ be constant at entropy:

$$dp(\mathbf{x}) = H(p) \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Then for any $q \in \Delta_m$ and $\mathbf{x} \in \mathbb{R}^{n+m-1}$ and $1 \leq i \leq n$,

$$d(p \circ_i q)(\mathbf{x}) \ = \ H(p \circ_i q) \overset{\text{chain rule}}{=} H(p) + p_i H(q) \ = \ \underbrace{(dp \circ_i^R q)(\mathbf{x})}_{dp(\text{stuff})} + \underbrace{(p \circ_i^L dq)(\mathbf{x})}_{p_i dq(\text{stuff})}.$$

Conversely...

There's something to say in the other direction, too.

Conversely...

There's something to say in the other direction, too.

Suppose $\{d\colon \Delta_n \to \mathbf{End}_n(\mathbb{R})\}_{n\geq 1}$ is a sequence of continuous functions satisfying the product rule.

Define a new sequence $\{I\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ as follows:

$$I(p) := dp(\mathbf{0}) \quad \forall p \in \Delta_n$$

where $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^n$.

Conversely...

There's something to say in the other direction, too.

Suppose $\{d\colon \Delta_n \to \mathbf{End}_n(\mathbb{R})\}_{n\geq 1}$ is a sequence of continuous functions satisfying the product rule.

Define a new sequence $\{I\colon \Delta_n \to \mathbb{R}\}_{n\geq 1}$ as follows:

$$I(p) := dp(\mathbf{0}) \quad \forall p \in \Delta_n$$

where $\mathbf{0} = (0,\dots,0) \in \mathbb{R}^n$.

The I are continuous, and moreover

$$I(p \circ (q^1,\dots,q^n)) = d(p \circ (q^1,\dots,q^n))(\mathbf{0})$$

$$\overset{\text{can show}}{=} dp(\mathbf{0}) + \sum_i p_i dq^i(\mathbf{0})$$

$$= I(p) + \sum_i p_i I(q^i).$$

By Leinster's theorem, $dp(\mathbf{0}) = I(p) = cH(p)$ for some constant $c \in \mathbb{R}$.

Theorem

entropy $\leftrightarrows$ derivations of Δ

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.

Corollary

the chain rule

Proof: For each $n \geq 1$ let $d\colon \Delta_n \to \mathbf{End}_n(\mathbb{R})$ be constant at entropy $p \mapsto dp \equiv H(p)$. Then d is a derivation, and evaluating at any input $\mathbf{x}$ gives the following:

$$
\begin{aligned}
H(p \circ (q^1, \ldots, q^n)) &= d(p \circ (q^1, \ldots, q^n))(\mathbf{x}) \\
&\overset{\text{can show}}{=} dp(\text{stuff}) + \sum_{i=1}^{n} p_i \, dq^i(\text{stuff}) \\
&= H(p) + \sum_{i=1}^{n} p_i H(q^i).
\end{aligned}
$$

For more details...

See "Entropy as a Topological Operad Derivation" (B., 2021):

arXiv:2107.09581 / Entropy 2021, 23(9), 1195.

This theorem was motivated by a remark of John Baez in "Entropy as a Functor" (nLab, 2011).

- Operads can be defined in any symmetric monoidal category, not just spaces.
- The left/right compositions $\circ_i^L$ and $\circ_i^R$ comprise a bimodule structure, which can be defined on any operad.
- Likewise, derivations valued in a bimodule can be defined on any operad.
- Question: What are the values of arbitrary derivations away from zero, $dp(\mathbf{x}) = ?$

What's more interesting

(to me) is if/how this relates to other results.

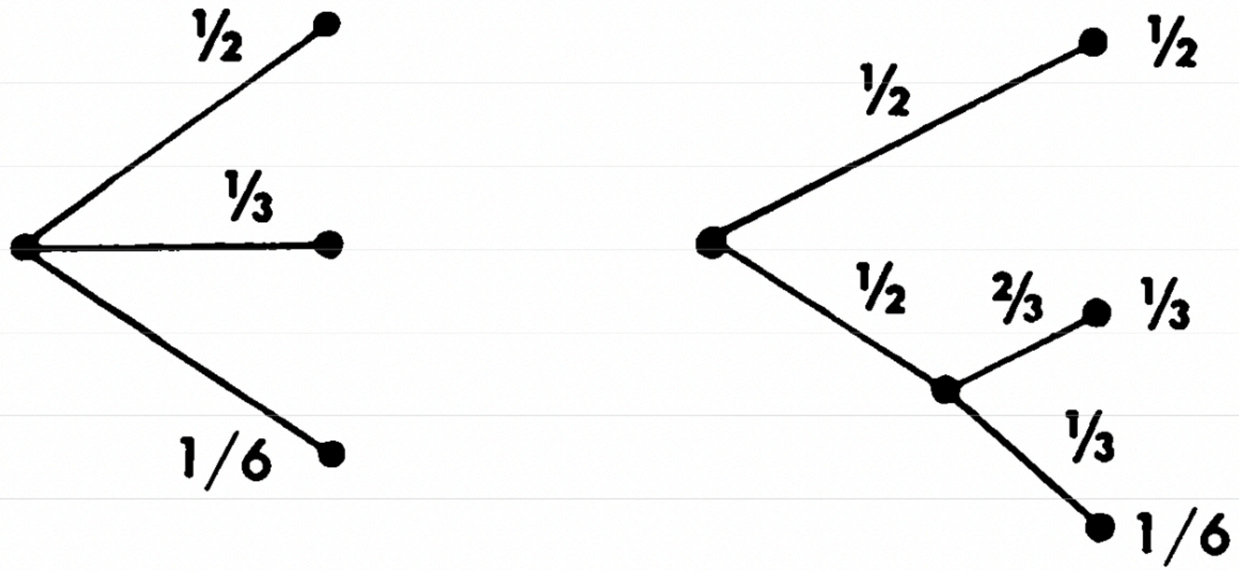


Fig. 6. — Decomposition of a choice from three possibilities.

We define the *conditional entropy* of y, $H_x(y)$ as the average of the entropy of y for each value of x, weighted according to the probability of getting that particular x. That is

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j).$$

This quantity measures how uncertain we are of y on the average when we know x. Substituting the value of $p_i(j)$ we obtain

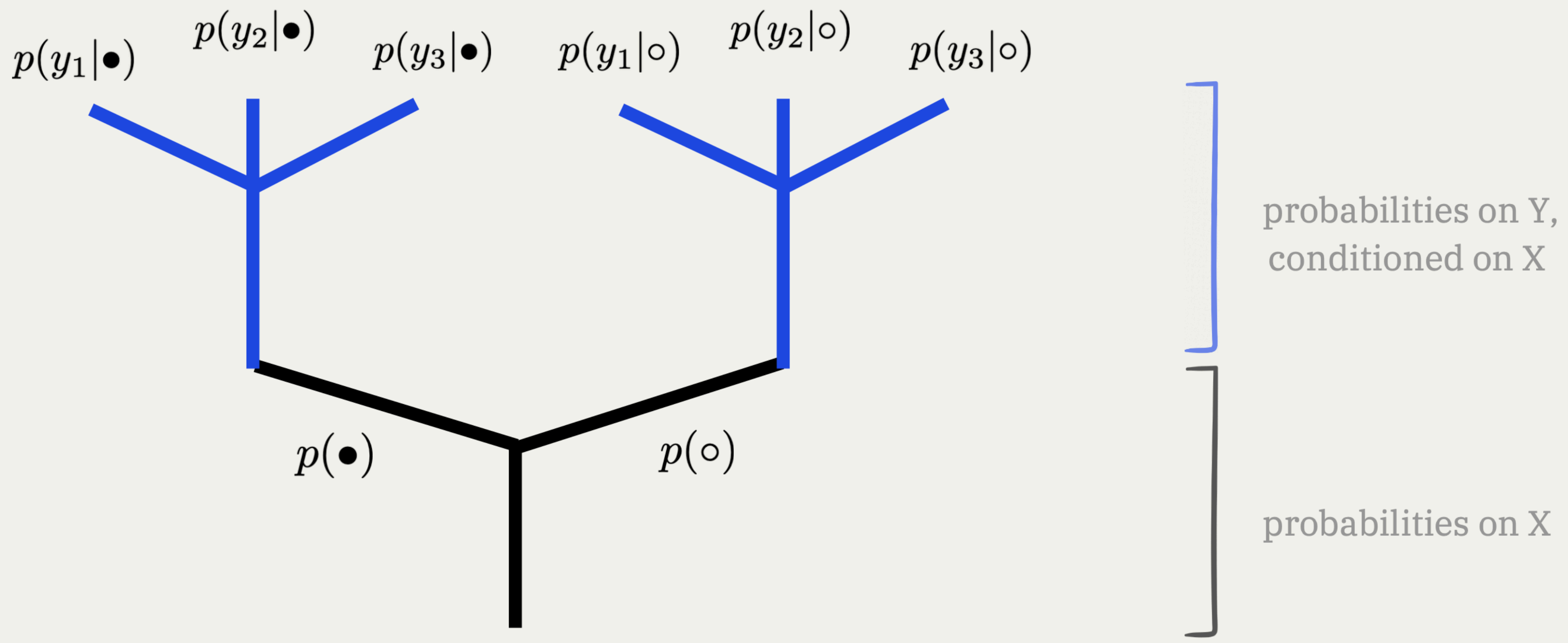
$$\begin{aligned} H_x(y) &= -\sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} p(i,j) \log \sum_j p(i,j) \\ &= H(x,y) - H(x) \end{aligned}$$

or

$$H(x,y) = H(x) + H_x(y).$$

The uncertainty (or entropy) of the joint event x, y is the uncertainty of x plus the uncertainty of y when x is known.

Suppose $x = \{\bullet, \circ\}$ and $y = \{y_1, y_2, y_3\}$, then:



$$H(x, y) = H(x) + xH(y)$$

Conditional entropy satisfies $H(x,y) = H(x) + H(y \mid x)$.

If we let $xH(y) := H(y \mid x)$ then this becomes

$$H(x,y) = H(x) + xH(y)$$

which looks like a derivation. In fact, it's the 1-cocycle condition in group cohomology. (Such maps are called "crossed homomorphisms.")

In 2015, this analogy was fleshed out in full detail, leading to information cohomology where entropy represents the unique 1-cocyle of a certain cochain complex:

- P. Baudot, D. Bennequin, "The Homological Nature of Entropy" (Entropy, 2015)

Just getting started

New ways of thinking about entropy aren't new, but there's also a sense in which this subject is "just getting started" (Baez).

A good place to start is the Categorical Semantics of Entropy symposium at the CUNY ITS (May 2022) - see talks on YouTube.

Operads

Topological simplices:

- Tom Leinster, "Entropy as an Internal Algebra" (chapter 12, Entropy and Diversity)
  - See also, "An Operadic Introduction to Entropy" (2011, n-Category Café)
- Matilde Marcolli and Ryan Thorngren, "Thermodynamic Semirings," (2012)
  - introduce "thermodynamic semirings" as the algebra for a certain operad, related to Δ

Operads

Topological simplices:

- Tom Leinster, "Entropy as an Internal Algebra" (chapter 12, Entropy and Diversity)
    - See also, "An Operadic Introduction to Entropy" (2011, n-Category Café)
- Matilde Marcolli and Ryan Thorngren, "Thermodynamic Semirings," (2012)
    - introduce "thermodynamic semirings" as the algebra for a certain operad, related to Δ

Convex relations:

- John Baez, Owen Lynch, Joe Moeller: "Compositional Thermostatics" (2021)
    - Construct an operad whose algebras are thermostatic systems (e.g. entropy-like functions) in a unifying framework for thermodynamics and classical/ quantum statistical mechanics
    - Connection(s)?

Category Theory

Polynomial functors

- David Spivak, "Polynomial Functors and Shannon Entropy" (2022)

Quantum

- Arthur Parzygnat, "A functorial characterization of von Neumann entropy" (2020)
    - Also see YouTube talk "On characterizing classical and quantum entropy" + references therein

Category Theory

Polynomial functors

- David Spivak, "Polynomial Functors and Shannon Entropy" (2022)

Quantum

- Arthur Parzygnat, "A functorial characterization of von Neumann entropy" (2020)
    - Also see YouTube talk "On characterizing classical and quantum entropy" + references therein

Quantum & homological algebra

- Tom Mainiero, "Homological tools for the quantum mechanic" (2019)
    - entropy appears in the Euler characteristic of a cochain complex associated to a quantum state

Homological algebra & topos theory

- P. Baudot, D. Bennequin, "Homological nature of entropy" (2015)
    - information cohomology

What's the connection?

- operads
- category theory
- polynomial functors
- quantum physics
- homological algebra
- topos theory
- entropy

?