An Operadic Tale of Entropy

Emmanouil Sfinarolakis

"***Entropy: An ethereal beast whose roar echoes through the realms of mathematics, physics, biology and computer science...***"

AGQ CDT
Department of Mathematics and Computer Science
Heriot-Watt University

Overview

- introduce Shannon Entropy
- entropy as *indicator*
- entropy as *quantifier*
- entropy as *derivation*

"Surprise": Layman's Approach

Goal: quantify *"surprise"* of event x

Step 1: educated guess $S(x) \propto \frac{1}{P(x)}$

Step 2: thought experiment

(!!!) $P(H) = 1 \implies S(H) = 1 \neq 0$ FIX: $S(x) \propto \log_b \left(\frac{1}{P(x)}\right) = 0$ ✓

(!!!) $P(T) = 0 \implies S(T) \longrightarrow \infty \neq 0$ FIX: $S(x) \propto P(x) \log_b \left(\frac{1}{P(x)}\right) := 0$ ✓

Surprise Function: $\boxed{S(x) = -P(x) \log_b P(x)}$

Shannon Entropy

Definition (Shannon Entropy)

Let

- $X = (\mathcal{X}, P)$
- $\mathcal{X} = (x_1, x_2, \ldots, x_n) =$ outcomes
- $P = (P_1, P_2, \ldots, P_n) =$ probability distribution

$$P \in [0,1]^n \quad \& \quad \sum_{i=1}^{n} P_i = 1.$$

Then, the *Shannon entropy* associated to X is:

$$H(X) \coloneqq \sum_{x \in \mathcal{X}} S(x) = -\sum_{i=1}^{n} P_i \log_b (P_i)$$

Variability & Complexity of Structures

START: $\mathbb{A}$ = structure, theory, collection of objects, model, system, etc

GOAL: describe how complex $\mathbb{A}$ is

IDENTIFY: outcomes + probability $\longrightarrow X_{\mathbb{A}} = (\mathcal{X}, P)$

NEED: *Indicator Function* $H : X_{\mathbb{A}} \longrightarrow [0, \infty]$

QUESTION: *What would a good indicator of complexity be?*

Indicator of Variability: Requirements

- **R1: Continuity** $H = H(P)$

– small change in P $\implies$ small change in H $H(P+\varepsilon) = H(P) + \epsilon$

Example (Coin)

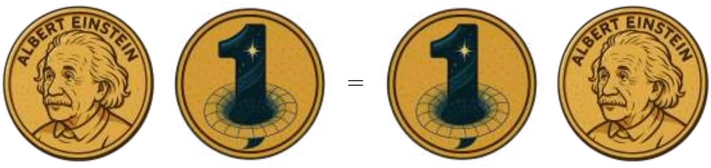
- Fair Coin: $P(E,H,T) = \left(0, \frac{1}{2}, \frac{1}{2}\right)$
- Fair Coin + Edge: $P(E,H,T) = \left(\frac{1}{6000}, \frac{1}{2}\cdot\frac{5999}{6000}, \frac{1}{2}\cdot\frac{5999}{6000}\right)$

$$\approx (0.0001667, 0.4999167, 0.4999167)$$

Indicator of Variability: Requirements

- **R2: Symmetry**

– re-order outcomes $x_i \implies H$ unchanged $\qquad H(P) = H(\sigma(P))$



Indicator of Variability: Requirements

- **R3: Maximality**

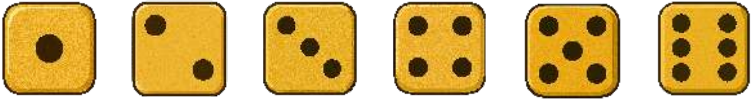
– uniform distribution $U(n) \implies H$ maximal $\qquad H(P) \le H(U(n))$

$>$

Indicator of Variability: Requirements

- **R4: Monotonicity**

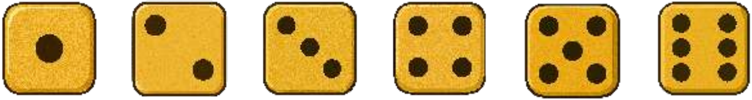
– $U(n)$: # outcomes $\uparrow \implies H \uparrow$ $\qquad H(U(n)) < H(U(n+1))$



$>$

- **R5: Expansibility**

– outcome of $P = 0 \implies H$ unchanged $\qquad H(P_1, \ldots, P_n) = H(P_1, \ldots, P_n, 0)$

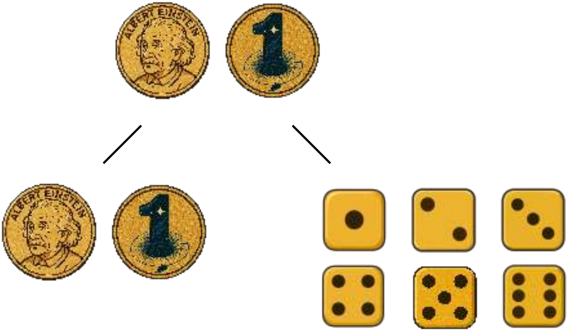


$+$

Indicator of Variability: Requirements

- **R6: Disintegration**

$$- \text{ combine } \begin{cases} \text{macroscopic distribution } P \\ \text{family } Q \text{ of microscopic distrib. } Q^i \end{cases} \qquad H(PQ) = H(P) + \sum_i P_i H(Q^i)$$



Unique Candidate!

Theorem (Khinchin, 1953, 5 axioms)

Shannon entropy is the ONLY function that satisfies the continuity, symmetry, maximality, expansibility *and disintegration conditions.*

expansibility $\longleftrightarrow$ monotonicity

Theorem (Faddeev, 1956, $3+1$ axioms)

Shannon entropy is the ONLY function that satisfies binary (2D) *continuity,* binary positivity at least at one point, *symmetry and* binary *disintegration.*

Theorem (Leinster, 2021, $3+1$ axioms)

Shannon entropy is the ONLY function that satisfies the continuity, symmetry and disintegration conditions, IF the domain is finite, strictly-positive distributions.

Information Content

START: $\mathbb{A}$ = structure, theory, collection of objects, model, system, etc

IDENTIFY: outcomes + probability $\longrightarrow X_{\mathbb{A}} = (\mathcal{X}, P)$

GOAL: quantify amount of information acquired due to observation of event i with probability P_i

NEED: *Information Function* $I : X_{\mathbb{A}} \longrightarrow [0, \infty]$

QUESTION: *What are the fundamental properties of information?*

Information Content: Requirements

- **R1: Continuity**

– should depend ONLY on P $\quad I = I(P)$

– small change in $P \implies$ small change in I $\quad I(P+\varepsilon) = I(P) + \epsilon$

- **R2: Non-Negativity**

– outcome cannot reduce knowledge; new information ≥ 0 $\quad I(P) \geq 0$

- **R3: Boundedness**

– uniform distribution $U(n) \implies I$ maximal $\quad I(P) \leq I(U(n))$

- **R4: Antitonicity**

– Inverse Probability Law

– rare events carry more information $\quad P_1 < P_2 \implies I(P_1) > I(P_2)$

- **R5: Additivity**

– independent events $\implies$ additive information $\quad I(P_1 \cdot P_2) = I(P_1) + I(P_2)$

Unique Candidate!

Theorem (Shannon)

Shannon information $I(P) = -\log_b(P)$ *is the ONLY function that satisfies the continuity, antitonicity and additivity conditions.*

Entropy as a Quantifier

$$\text{variability} \uparrow \implies \# \text{ questions} \uparrow$$
$$\text{entropy} = \text{measure of variability}$$

Theorem (Shannon)

Shannon entropy measures the optimal average number of binary questions needed to identify an element from a distribution.

- Non-Binary Questions?

$$H(P) = -\sum_i P_i \log_b(P_i)$$

- $b = \#$ answers per question

Is Entropy a Derivation?

Reminder: derivation = linear map + Leibniz's rule

Def: Let $d : [0,1] \longrightarrow \mathbb{R}$ (surprise function) such that

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

Note: $d(xy) = -xy\log(xy) = -xy[\log(x) + \log(y)] = -xy\log(x) - xy\log(y)$

$\Longrightarrow$ $d(xy) = d(x)y + xd(y)$

$\Longrightarrow$ $d \models$ Leibniz's rule

BUT: d not linear! $\Longrightarrow$ not *classical* derivation

AND: $H(P) = \sum_i d(P_i) \xrightarrow{d \neq \text{linear}} H(P) \neq d\left(\sum_i P_i\right) \Longrightarrow \neg$ linearity

Treasure Hunt

In what sense

in which framework

under what conditions

X

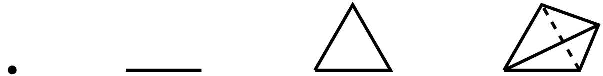
IS entropy a derivation?

Topological Simplices

Idea: identify $P = (P_1, \ldots, P_n)$ as a *point* in $\mathbb{R}^n$

Definition (Topological n-Simplex)

$$\Delta^n := \{\ (P_1, \ldots, P_n) \in \mathbb{R}^n \mid 0 \le P_i \le 1\ \forall i,\ \sum_{i=1}^{n} P_i = 1\ \}$$

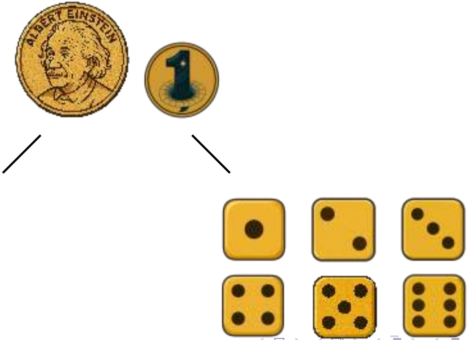


Δ^1 Δ^2 Δ^3 Δ^4

Composition of Probability Distributions

Example (Simultaneous Composition)

- $P = \left(\frac{2}{3}, \frac{1}{3}\right) \in \Delta^2$ (unfair coin)
- family $Q = (Q_1, Q_2) : \begin{cases} Q_1 = (1) \in \Delta^1 & \text{(do nothing)} \\ Q_2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right) \in \Delta^6 & \text{(fair die)} \end{cases}$
- $PQ = P \circ (Q_1, Q_2) = \left(\frac{2}{3} \cdot 1, \frac{1}{3} \cdot \frac{1}{6}, \ldots, \frac{1}{3} \cdot \frac{1}{6}\right) = \left(\frac{2}{3}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}\right) \in \Delta^7$



Definition (Pointwise Evaluation)

- $H = (H_1, H_2 \dots)$: family of functions $H_i : \Delta^i \longrightarrow [0, \infty]$
- $Q = (Q_1, Q_2, \dots, Q_n)$: family of probability distributions $Q_i \in \Delta^{k_i}$

The *pointwise evaluation* of H on Q is

$$\begin{aligned} H(Q) &= H(Q_1, Q_2, \dots, Q_n) \\ &= \big(H_{k_1}(Q_1), H_{k_2}(Q_2), \dots, H_{k_n}(Q_n)\big) \end{aligned}$$

Example

- $Q = (Q_1, Q_2)$; $\quad Q_1 = (1) \in \Delta^1 \quad \& \quad Q_2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right) \in \Delta^6$
- $H(Q) = (H_1(Q_1), H_6(Q_2))$

ALMOST Derivation

DEFINE:

$$\begin{cases} x = (x_1, \ldots, x_n) \in \mathbb{R}^n \\ P = (P_1, \ldots, P_n) \in \Delta^n \end{cases} \implies \boxed{\langle P, x \rangle = \sum_{i=1}^{n} P_i x_i} \implies \begin{cases} \overline{P} : \mathbb{R}^n \longrightarrow \mathbb{R} \\ \qquad x \longmapsto \langle P, x \rangle \end{cases}$$

NOTE:

$$H_m(PQ) = H_n(P) + \sum_{i=1}^{n} P_i H_{k_i}(Q_i)$$

$$\implies H_m(PQ) = H_n(P) + \overline{P}(H_{k_1}(Q_1), \ldots, H_{k_n}(Q_n))$$

$$\implies \boxed{H_m(PQ) = H_n(P) + \overline{P}(H(Q))}$$

OBSERVE:
- Q missing from 1st term on RHS $\implies$ *almost* derivation
- H *different inputs* $\implies$ operad of functions
- P *different dimensions* (of $\mathbb{R}^n$) $\implies$ operad of spaces

Operads I

- n-ary operations $\iff$ planar rooted trees with n leaves
- composition $\circ_i$ $\iff$ grafting tree on i-th leaf
- $\mathcal{O}(n)$ = collection of ALL n-ary operations

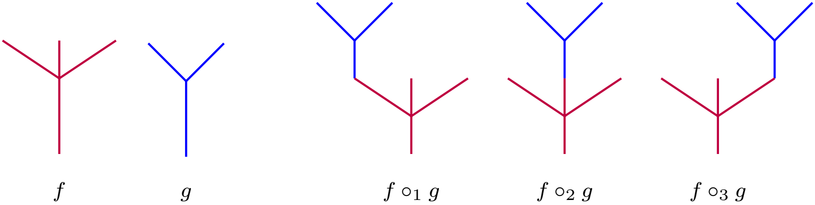
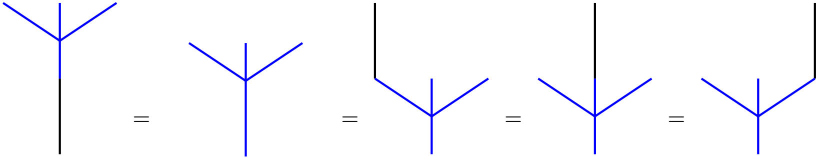


Figure: (3-ary) $f \circ g$ (2-ary) in 3 WAYS

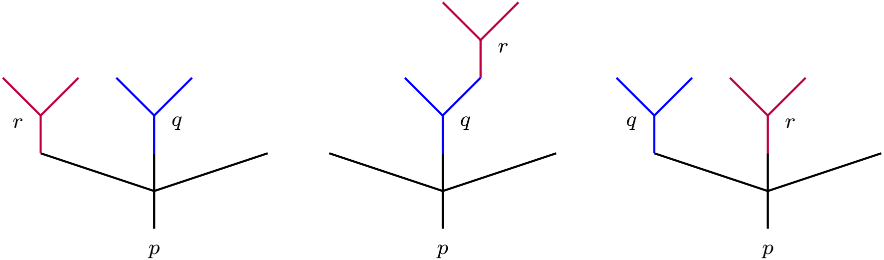
- identity of composition $\Longrightarrow$ $1 \circ p = p = p \circ 1$



$1 \circ_1 p$ $\quad p \quad$ $p \circ_1 1$ $\quad p \circ_2 1$ $\quad p \circ_3 1$

Operads III

- associativity of composition $p \circ q \circ r$
- relative position of $q, r \implies$ 3 cases $\implies$ 3 associativity axioms



$(p \circ q) \circ r = (p \circ r) \circ q$ $\quad$ $(p \circ q) \circ r = p \circ (q \circ r)$ $\quad$ $(p \circ q) \circ r = (p \circ r) \circ q$

Definition (Non-Symmetric Operad)

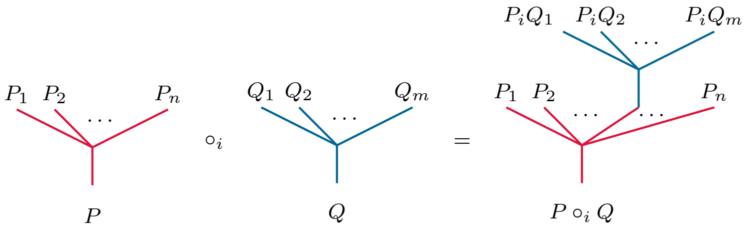
Let $(C, \otimes)$ a symmetric monoidal category. An *operad* $\mathcal{O}$ in C consists of:

- $\mathcal{O} = (\mathcal{O}(1), \mathcal{O}(2), \ldots)$ sequence of objects in C
- $\circ_i : \mathcal{O}(n) \otimes \mathcal{O}(m) \longrightarrow \mathcal{O}(n+m-1)$ morphisms in C
- $1 \in \mathcal{O}(1)$ (*identity element*)

satisfying the *identity* & *associativity* $(\times 3)$ axioms.

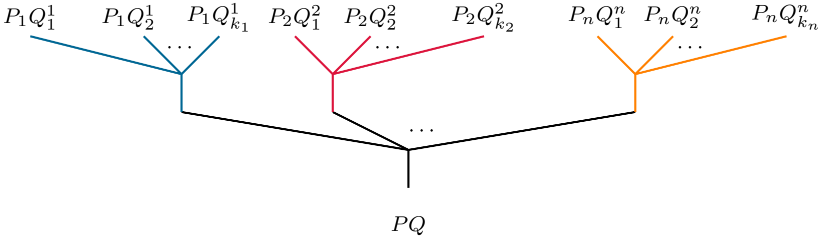
Definition (Δ)

- $(\mathrm{Top}, \times)$: category of topological spaces with continuous maps
- $\Delta = (\Delta^1, \Delta^2, \dots)$
- $P \in \Delta^n \iff$ planar rooted tree with n leaves
- composition $\circ_i \iff$ grafting tree on i-th leaf
- $1 \in \Delta^1 \iff$ unique P on a 1-point set $\qquad 1 = (1)$



Simultaneous Composition as Tree

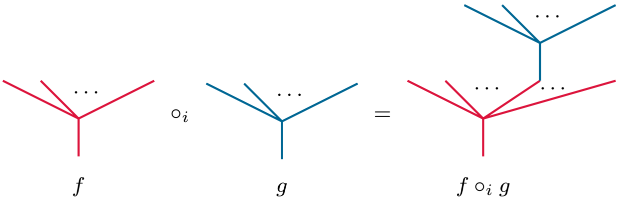
- $P = (P_1, \dots, P_n)$
- $Q = (Q^1, Q^2, \dots, Q^n)$



Endomorphism Operad $\mathrm{End}_{\mathbb{R}}$

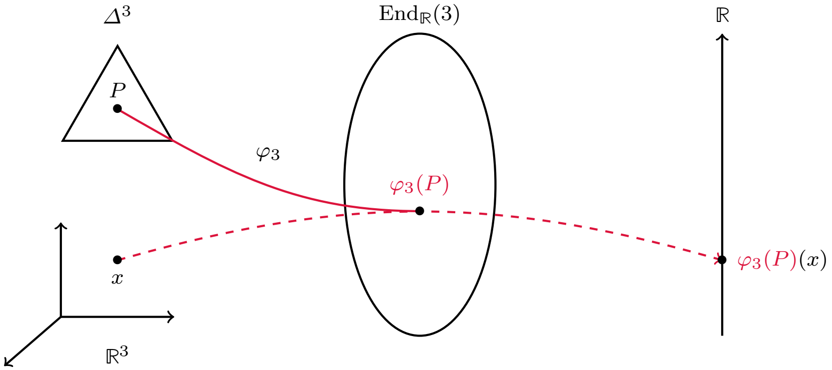
Definition ($\mathrm{End}_{\mathbb{R}}$)

- $(\mathrm{Top}, \times)$: category of topological spaces with continuous maps
- $\mathrm{End}_{\mathbb{R}}(n)$: space of continuous functions $\mathbb{R}^n \longrightarrow \mathbb{R}$ (product topology)
- $\mathrm{End}_{\mathbb{R}} = (\mathrm{End}_{\mathbb{R}}(1), \mathrm{End}_{\mathbb{R}}(2), \ldots)$
- $f \in \mathrm{End}_{\mathbb{R}}(n) \Longleftrightarrow$ planar rooted tree with n leaves
- composition $\circ_i \Longleftrightarrow$ grafting tree on i-th leaf
- $1 \in \mathrm{End}_{\mathbb{R}}(1) \Longleftrightarrow$ identity function $\mathrm{id}_{\mathbb{R}} : \mathbb{R} \longrightarrow \mathbb{R}$



$\mathrm{End}_{\mathbb{R}}$: A Bridge

- $\begin{cases} \varphi_n : \Delta^n \longrightarrow \mathrm{End}_{\mathbb{R}}(n) \\ \qquad P \longmapsto \langle P, \bullet \rangle \end{cases}$ where $\varphi_n(P)(x) := \langle P, x \rangle = \sum_{i=1}^{n} P_i x_i$



Derivation out of Δ

GOAL: define map d out of Δ obeying Leibniz's rule

$$d(P \circ_i Q) = d(P) \circ_i Q + P \circ_i d(Q) \qquad (\textit{desideratum})$$

KEY: observe RHS

(i) addition $(+)$

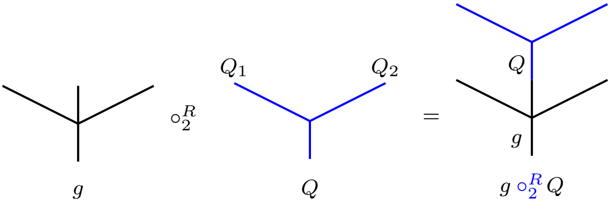
(ii) order does not matter for $+$

(iii) 1^{st} term: right multiplication $(\circ_i) \implies \circ_i^R$

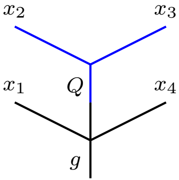
(iv) 2^{nd} term: left multiplication $(\circ_i) \implies \circ_i^L$

NEED: abelian bimodule $(\mathrm{End}_{\mathbb{R}}, +, \circ_i^R, \circ_i^L)$ over Δ

Right Composition in $\mathrm{End}_{\mathbb{R}}$



$(g \circ_2^R Q)(x_1, x_2, x_3, x_4) \;=$

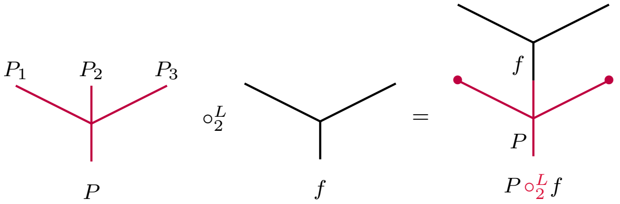


$= \; g(x_1, Q_1 x_2 + Q_2 x_3, x_4)$

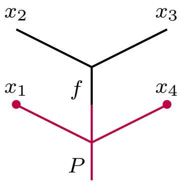
Figure: (above) Right composition = weighted−collapse evaluation

(below) Right composition acting on $x = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$

Left Composition in $\mathrm{End}_{\mathbb{R}}$



$(P \circ_2^L f)(x_1, x_2, x_3, x_4) \quad =$



$= \quad P_2\, f(x_2, x_3)$

Figure: (above) Left composition = scaled local evaluation

(below) Left composition acting on $x = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$

Derivation of Δ in $\mathrm{End}_{\mathbb{R}}$

Definition

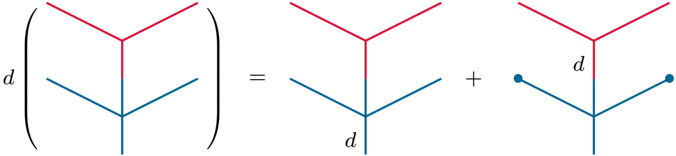
A derivation of Δ in the bimodule $\mathrm{End}_{\mathbb{R}}$ over Δ is:

- $d = (d_1, d_2, \ldots)$ — sequence of morphisms
- $d_n : \Delta^n \longrightarrow \mathrm{End}_{\mathbb{R}}(n)$ — morphisms
- $\forall\ P \in \Delta^n,\ Q \in \Delta^m$ — $(\forall\ 1 \leq i \leq n)$

$$d_{n+m-1}(P \circ_i Q) = d_n(P) \circ_i^R Q + P \circ_i^L d_m(Q)$$

Derivation of Composition

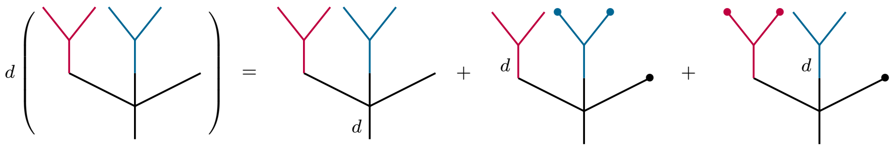
- $P \in \Delta^3$
- $Q \in \Delta^2$
- omit subscripts $\implies$ d instead of d_n



$$d(P \circ_2 Q) = d(P) \circ_2^R Q + P \circ_2^L d(Q)$$

Derivation of Simultaneous Composition

- $P \in \Delta^3$, $Q_1 \in \Delta^2$, $Q_2 \in \Delta^2$
- $Q = (Q_1, Q_2)$
- simultaneous composition $PQ \implies$ repeated composition $(P \circ_1 Q_1) \circ_2 Q_2$
- iterate d



$$d(PQ) = d\left((P \circ_1 Q_1) \circ_2 Q_2\right)$$

$$= d(P \circ_1 Q_1) \circ_2^R Q_2 + P \circ_1^L Q_1 \circ_2^L d(Q_2)$$

$$= d(P) \circ_1^R Q_1 \circ_2^R Q_2 + P \circ_1^L d(Q_1) \circ_2^R Q_2 + P \circ_1^L Q_1 \circ_2^L d(Q_2)$$

Entropy as Derivation

Theorem (Classification)

(i) *Shannon entropy H defines a derivation*

$$\begin{cases} d : \Delta \longrightarrow \mathrm{End}_{\mathbb{R}} \\ \quad P \longmapsto dP, \qquad dP(x) \equiv H(P) \end{cases}$$

of the operad Δ of topological simplices.

(ii) (Converse:) *Every derivation on the operad Δ*

$$d : \Delta \longmapsto \mathrm{End}_{\mathbb{R}}$$

is a scalar multiple of Shannon entropy when evaluated at the origin:

$$dP(0) = \alpha H(P)$$

- space of derivations: 1D vector space; basis = Shannon entropy
- $x = (0, \ldots, 0) \in \mathbb{R}^n, \ \forall n \implies$ dimension-independent anchor

Conclusion

- Entropy is a *functional* on probability distributions
- Entropy is an *indicator* of complexity/variability
- Entropy is a *measure* of symmetry breaking (deviation from uniformity)
- Entropy is a *quantifier* of the average amount of information produced by a stochastic source
- Entropy is a *functional* over geometric objects
- Entropy is a *derivation* of the operad of topological simplices