Baez-Fritz-Leinster's theorem on information loss

Notes by: Arthur Parzygnat

March 29, 2016

Abstract

In this mostly expository note, we use Fadeev's theorem to prove Baez-Fritz-Leinster's theorem on characterizing entropy in terms of an affine functor [3]. Although a proper treatment involves categorifying the concept of a convex object, we follow the proof given in [3].

In an earlier note [4], we gave a detailed proof of Fadeev's theorem, which is (a slight modification of) the following.

Theorem 1. *Let*

$$\Delta = \coprod_{n=1}^{\infty} \Delta^{n-1} \tag{2}$$

be the coproduct of all $(n-1)$-simplices

$$\Delta^{n-1} := \left\{ (p_1, \ldots, p_n) \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \;\forall\; i = 1, \ldots, n \right\}. \tag{3}$$

Let $J : \Delta \to \mathbb{R}$ be any function satisfying the following properties.

(a) For each n, the diagram

$$\begin{tikzcd} \Delta^{n-1} \arrow[rr, "\sigma"] \arrow[dr, "J"'] & & \Delta^{n-1} \arrow[dl, "J"] \\ & \mathbb{R} & \end{tikzcd} \tag{4}$$

commutes for all $\sigma \in S_n$.

(b) The function $J_1 \equiv J|_{\Delta^1} : \Delta^1 \to \mathbb{R}$ is continuous.

(c) $0 \leq J(\frac{1}{2}, \frac{1}{2}) < \infty$.

(d) For each $n \in \mathbb{Z}^+$,

$$J\big(tp_1, (1-t)p_1, p_2, \ldots, p_n\big) = J(p_1, \ldots, p_n) + p_1 J(t, 1-t) \tag{5}$$

for all $t \in [0, 1]$.

Then $J = cH$, where H is the Shannon entropy function defined by

$$H(p_1, \dots, p_n) := -\sum_{i=1}^{n} p_i \log_2 p_i \tag{6}$$

and $c \in \mathbb{R}_{\geq 0}$ is a constant.

This theorem was proved in [4] using a Lemma from [5] and based on a sketch indicated in [3]. The original proof is due to Fadeev.

In this note, we will use the notions of convex categories, cone categories, convex/affine functors, and proportionalities of such functors. These ideas are currently being developed, so the following is meant to be a sketch for now. In particular, the "definitions" below are not actually definitions. The proofs are due to [3].

Definition 7. A *convex category* is a category $\mathcal{C}$ together with a family of functors $F_\lambda : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ indexed by $\lambda \in [0,1]$ together with natural transformations (often isomorphisms) that "categorify" the following axioms

$$\begin{aligned}
F_0(x,y) &= y && \text{(unit law)} && (8)\\
F_\lambda(x,y) &= F_{1-\lambda}(y,x) && \text{(parametric commutativity)} && (9)\\
F_\lambda\Big(F_\mu(x,y), z\Big) &= F_{\lambda \llcorner \mu}\Big(x, F_{\lambda \lrcorner \mu}(y,z)\Big) && \text{(deformed parametric associativity)} && (10)\\
F_\lambda(x,x) &= x && \text{(idempotency)}, && (11)
\end{aligned}$$

where

$$\lambda \llcorner \mu := \lambda\mu \qquad \& \qquad \lambda \lrcorner \mu := \begin{cases} \frac{\lambda(1-\mu)}{1-\lambda\mu} & \text{if } \lambda\mu \neq 1 \\ \text{arbitrary} & \text{if } \lambda = \mu = 1, \end{cases} \tag{12}$$

which themselves satisfy several "coherence conditions." The functor F_λ is often written as

$$F_\lambda(x,y) = \lambda x \oplus (1-\lambda) y \tag{13}$$

on its inputs.

The fact that this is not a definition is because the terms "categorify" and "coherence conditions" have not been made precise. A precise definition exists, but it is a bit too involved for this note. It is based on a the definition of semiconvex sets of Świrszcz from 1974. There are several examples, two of which we mention now.

Example 14. Let **FinProb** be the category of finite probability spaces and a.e. equivalence classes of measure-preserving functions. For every $\lambda \in [0,1]$, define the convex sum F_λ on objects by

$$\begin{aligned}
\lambda(X,p) \oplus (1-\lambda)(Y,q) &:= \big(X \amalg Y, \lambda p \oplus (1-\lambda) q\big),\\
\text{where } \big(\lambda p \oplus (1-\lambda) q\big)(z) &:= \begin{cases} \lambda p(z) & \text{if } z \in X \\ (1-\lambda) q(z) & \text{if } z \in Y. \end{cases}
\end{aligned} \tag{15}$$

The convex sum of morphisms $(X,p) \xrightarrow{f} (X',p')$ and $(Y,q) \xrightarrow{g} (Y',q')$ is given by

$$\big(\lambda f \oplus (1-\lambda) g\big)(z) := \begin{cases} f(z) & \text{if } z \in X \\ g(z) & \text{if } z \in Y. \end{cases} \tag{16}$$

Remark 17. Note that $F_0\big((X,p),(Y,q)\big) \neq (Y,q)$. In fact, if we did not use a.e. equivalence classes of measure-preserving maps (which [3] do not), then $F_0\big((X,p),(Y,q)\big)$ is not even *isomorphic* to (Y,q). Therefore, our category **FinProb** differs from the one introduced in [3]. Our version is such that $F_0\big((X,p),(Y,q)\big) \cong (Y,q)$.

Example 18. Let $\mathbb{B}\mathbb{R}_{\geq 0}$ be the one-object category whose morphisms are all non-negative numbers with composition addition. Then the convex sum on objects is trivial (because there is only a single object) and the convex sum on morphisms (which are non-negative numbers) is defined by

$$F_\lambda(a,b) := \lambda a + (1-\lambda) b. \tag{19}$$

The last example is also an example of a cone category.

Definition 20. A <u>*cone category*</u> is a symmetric semigroupal category $(\mathcal{C}, \oplus, a, \phi)$, where a is the associator and ϕ is the symmetric braiding, together with a family of functors $k_\lambda : \mathcal{C} \to \mathcal{C}$ for every $\lambda \in \mathbb{R}_{\geq 0}$ and natural transformations (often isomorphisms) "categorifying" the following axioms.

$$k_{\lambda\mu}(x) = k_\lambda\big(k_\mu(x)\big) \qquad \text{(distributivity over multiplication in } \mathbb{R}_{\geq 0}) \tag{21}$$

$$k_\lambda(x+y) = k_\lambda(x) + k_\lambda(y) \qquad \text{(distributivity over addition in } \mathcal{C}) \tag{22}$$

$$k_{\lambda+\mu}(x) = k_\lambda(x) + k_\mu(x) \qquad \text{(distributivity over addition in } \mathbb{R}_{\geq 0}) \tag{23}$$

$$k_1(x) = x \qquad \text{(unit axiom)}, \tag{24}$$

which themselves satisfy several "coherence conditions."

Example 25. The example $\mathbb{B}\mathbb{R}_{\geq 0}$ is actually a cone category since

$$k_\lambda(a) := \lambda a \tag{26}$$

is a non-negative real number for all $\lambda \in \mathbb{R}_{\geq 0}$ and $a \in \mathbb{R}_{\geq 0}$.

However, **FinProb** is *not* a cone category since if p is a probability distribution on X, λp is not a probability distribution on X for all $\lambda \in \mathbb{R}_{\geq 0}$.

Remark 27. A review on what categorification is heuristically is given in [1]. However, what we mean by categorification is actually slightly weaker in the definitions of convex and cone categories since equalities are not always replaced by isomorphisms.

Lemma 28. *Every cone category is a convex category with the convex sums defined by*

$$F_\lambda(x,y) := k_\lambda(x) \oplus k_{1-\lambda}(y). \tag{29}$$

Definition 30. Let $(\mathcal{C}, \{F_\lambda\}_{\lambda\in[0,1]})$ and $(\mathcal{D}, \{G_\lambda\}_{\lambda\in[0,1]})$ be two convex categories. A functor $K : \mathcal{C} \to \mathcal{D}$ is <u>*affine*</u> if

$$K\big(\lambda x \oplus (1-\lambda)y\big) = \lambda K(x) \oplus (1-\lambda)K(y) \tag{31}$$

for all inputs $x, y \in \mathcal{C}$ and $\lambda \in [0,1]$.

Lemma 32. *The functor $S : \mathbf{FinProb} \to \mathbb{BR}_{\geq 0}$ defined by*

$$S\Big((X,p) \xrightarrow{f} (Y,q)\Big) := H(p) - H(q) \equiv -\sum_{x\in X} p(x)\log_2\big(p(x)\big) + \sum_{y\in Y} q(y)\log_2\big(q(y)\big) \tag{33}$$

is affine.

Lemma 34. *Let $(\mathcal{C}, \{F_\lambda\}_{\lambda\in[0,1]})$ be a convex category and $(\mathcal{D}, \oplus, a, \phi, \{k_\lambda\}_{\lambda\in\mathbb{R}_{\geq 0}})$ a cone category. Given an affine functor $K : \mathcal{C} \to \mathcal{D}$ and a constant $c \in \mathbb{R}_{\geq 0}$, the functor $cK : \mathcal{C} \to \mathcal{D}$ defined by*

$$cK(x) := k_c\big(K(x)\big) \tag{35}$$

is an affine functor.

Definition 36. Let $K, L : \mathcal{C} \to \mathcal{D}$ be two affine functors from a convex category $\mathcal{C}$ to a cone category $\mathcal{D}$. K is <u>*proportional*</u> to L if there exists a constant $c \in \mathbb{R}_{\geq 0}$ such that $K = cL$.

Before stating the theorem, we need just one more definition.

Definition 37. A sequence of morphisms

$$(X_n, p_n) \xrightarrow{f_n} (Y_n, q_n) \tag{38}$$

in **FinProb** indexed by $n \in \mathcal{I} \subset \mathbb{N}$ <u>*converges*</u> to a morphism $(X,p) \xrightarrow{f} (Y,q)$ if

(a) there exists an $N \in \mathbb{N}$ for which $X_n = X$, $Y_n = Y$, and $f_n = f$ for all $n \geq N$ and

(b) both

$$\lim_{n\to\infty} p_n = p \qquad \& \qquad \lim_{n\to\infty} q_n = q \tag{39}$$

where this limit is taken in the spaces,

$$\Delta^{|X|-1} \qquad \& \qquad \Delta^{|Y|-1}, \tag{40}$$

respectively.

A functor $K : \mathbf{FinProb} \to \mathbb{BR}_{\geq 0}$ is <u>*continuous*</u> if

$$\lim_{n\to\infty} K(f_n) = K(f) \tag{41}$$

whenever $\{f_n\}$ is a sequence in **FinProb** converging to f.

Remark 42. Although we have definitions of convex categories, functors, and so on, we have not yet been able to provide a useful definition of continuity of convex functors in the abstract setting. This is current work in progress. The definition we've given here is from [3]. Ideally, we would prefer to stay away from universes and also skelleta, which is what is done in [2]. Whatever the appropriate definition is, it should reproduce the one given above for functors from **FinProb** to $\mathbb{BR}_{\geq 0}$.

Theorem 43. *Let $K : \mathbf{FinProb} \to \mathbb{BR}_{\geq 0}$ be a continuous affine functor. Then K is proportional to S, i.e. there exists a constant $c \in \mathbb{R}_{\geq 0}$ such that for any morphism $(X,p) \xrightarrow{f} (Y,q)$ in* **FinProb**,

$$K(f) = c\Big(H(p) - H(q)\Big). \tag{44}$$

Proof. The idea of the proof will be to first use the functor K to define the "K-entropy" for each object (X,p) in **FinProb** and show that K is a difference of these entropies. Then we will prove this assignment satisfies the conditions of Fadeev's theorem and use that to prove the claim.

The set $\{1\}$ with the unit probability distribution $p(1) = 1$ is a finite probability space with a single element. Hence, we denote it by $\{1\}$ instead of $(\{1\}, 1)$. It is a terminal object in **FinProb**, which means that for every object (X,p), there is a unique function $t_{(X,p)} : (X,p) \to \{1\}$, which in this case sends every point of X to the single element 1, that satisfies the universal property that to any other finite probability space (Z,r) with a measure-preserving map $f : (X,p) \to (Y,q)$, there exists a unique measure-preserving map $h : (Y,q) \to \{1\}$ such that $t_{(X,p)} = h \circ f$. This unique map is $h = t_{(Y,q)}$. Using this, we define the <u>*K-entropy*</u> J of (X,p) to be

$$J(X,p) := K(t_{(X,p)}). \tag{45}$$

Thus, to every morphism $(X,p) \xrightarrow{f} (Y,q)$, the diagram

$$\begin{tikzcd} (X,p) \arrow[rr, "f"] \arrow[dr, "t_{(X,p)}"'] & & (Y,q) \arrow[dl, "t_{(Y,q)}"] \\ & \{1\} & \end{tikzcd} \tag{46}$$

commutes. Hence, by this fact and since K is a functor,

$$\begin{tikzcd}[column sep=small] & K(t_{(Y,q)} \circ f) \arrow[dl, equal] \arrow[dr, equal, "(46)"] & \\ K(t_{(Y,q)}) + K(f) \arrow[d, equal, "(45)"'] & & K(t_{(X,p)}) \arrow[d, equal, "(45)"] \\ J(Y,q) + K(f) & & J(X,p) \end{tikzcd} \tag{47}$$

which gives

$$K(f) = J(X,p) - J(Y,q). \tag{48}$$

This shows that knowledge of J on all objects in **FinProb** is sufficient to determine K. Hence, we will use J to prove that the assumptions in Fadeev's theorem hold.

(a) Let $(X,p) \xrightarrow{f} (Y,q)$ be an isomorphism with inverse denoted by f^{-1}. Then $K(f)=0$ because

$$K(f)+K(f^{-1}) = K(f\circ f^{-1}) = K(\mathrm{id}_{(Y,q)}) \overset{(48)}{=} 0 \tag{49}$$

but by assumption both $K(f)\geq 0$ and $K(f^{-1})\geq 0$ and from this it follows that

$$K(f)=0 \tag{50}$$

for all isomorphisms f. In particular, for any (X,p) in **FinProb**, a bijection $\sigma : X \to X$ of sets induces a measure-preserving isomorphism $f_\sigma : (X,p) \to (X,\sigma^*p)$, where

$$(\sigma^*p)(x) := p\big(\sigma(x)\big). \tag{51}$$

Thus,

$$J(X,p) = J(X,\sigma^*p). \tag{52}$$

Furthermore, because every object (X,p) of **FinProb**, with $|X|=n$, is isomorphic to $(\{1,\dots,n\},p)$ (for some non-unique choice of isomorphism), I is determined by its values on all objects of **FinProb** of the form $(\{1,\dots,n\},p)$ for some $n\in\mathbb{Z}^+$. Therefore, in particular, J defines a function $\Delta\to\mathbb{R}_{\geq 0}$.

(b) By assumption, K is continuous. In particular, let

$$\Big\{(\{1,2\},p_n) \xrightarrow{f_n} \{1\}\Big\}_n \tag{53}$$

be a sequence of morphisms in **FinProb** converging to $(\{1,2\},p) \xrightarrow{f} \{1\}$. Then

$$J(\{1,2\},p) = K(f) = \lim_{n\to\infty} K(f_n) = \lim_{n\to\infty} J(\{1,2\},p_n) \tag{54}$$

and so $J|_{\Delta^1} : \Delta^1 \to \mathbb{R}_{\geq 0}$ is continuous.

(c) By assumption, $0\leq K(f)<\infty$ for all morphisms f in **FinProb**. In particular,

$$0 \leq K\big(t_{(\{1,2\},(\frac{1}{2},\frac{1}{2}))}\big) = J\left(\{1,2\},\left(\frac{1}{2},\frac{1}{2}\right)\right) < \infty. \tag{55}$$

(d) Let $n \in \mathbb{Z}^+$ and let $(p_1, \ldots, p_n) := \big(p(1), \ldots, p(n)\big)$ be a probability measure on the set $\{1, \ldots, n\}$. First suppose that $p_1 \neq 1$. Then, since K is affine

$$\begin{aligned} J\big(\{1,\ldots,n\},(p_1,\ldots,p_n)\big) &\overset{(50)}{=} J\bigg(p_1\{1\} \oplus (1-p_1)\left(\{2,\ldots,n\},\left(\frac{p_2}{1-p_1},\ldots,\frac{p_n}{1-p_1}\right)\right)\bigg) \\ &\overset{(31)}{=} p_1 J\big(\{1\}\big) + (1-p_1)J\left(\{2,\ldots,n\},\left(\frac{p_2}{1-p_1},\ldots,\frac{p_n}{1-p_1}\right)\right) \\ &= (1-p_1)J\left(\{2,\ldots,n\},\left(\frac{p_2}{1-p_1},\ldots,\frac{p_n}{1-p_1}\right)\right), \end{aligned} \tag{56}$$

where the last equality follows from the fact that

$$J\big(\{1\}\big) = K\left(\{1\} \xrightarrow{\mathrm{id}_{\{1\}}} \{1\}\right) = J\big(\{1\}\big) - J\big(\{1\}\big) = 0. \tag{57}$$

Now let $t \in [0,1]$. Using this and affinity again, we obtain

$$\begin{aligned} &J\bigg(p_1\Big(t\{1\} \oplus (1-t)\{1\}\Big) \oplus (1-p_1)\left(\{2,\ldots,n\},\left(\frac{p_2}{1-p_1},\ldots,\frac{p_n}{1-p_1}\right)\right)\bigg) \\ &\quad\overset{(31)}{=} p_1 J\Big(\{1,2\},(t,1-t)\Big) + (1-p_1)J\left(\{2,\ldots,n\},\left(\frac{p_2}{1-p_1},\ldots,\frac{p_n}{1-p_1}\right)\right) \\ &\quad\overset{(56)}{=} p_1 J\Big(\{1,2\},(t,1-t)\Big) + J\Big(\{1,\ldots,n\},(p_1,\ldots,p_n)\Big). \end{aligned} \tag{58}$$

Since all the conditions in Fadeev's theorem holds, there exists a constant $c \in \mathbb{R}_{\geq 0}$ such that $J = cH$. Hence, by (48),

$$K = cS. \tag{59}$$

$\square$

References

[1] John C. Baez and James Dolan, *Categorification*, Higher Category Theory, 1998, pp. 1–36. http://arxiv.org/abs/math/9802029.

[2] John C. Baez and Tobias Fritz, *A Bayesian Characterization of Relative Entropy*, Theory and Applications of Categories **29** (2014), no. 16, 421–456. http://arxiv.org/abs/1402.3067.

[3] John C. Baez, Tobias Fritz, and Tom Leinster, *A Characterization of Entropy in Terms of Information Loss*, Entropy **13** (2011), 1945–1957. http://arxiv.org/abs/1106.1791.

[4] Arthur Parzygnat, *Fadeev's theorem on entropy.* Available at https://arthurparzygnat.com/expository-notes/.

[5] Alfréd Rényi, *On Measures of Entropy and Information*, Proc. Fourth Berkeley Symp. on Math. Statist. and Prob. **1** (1961), 547–561.