Markov categories and entropy

Paolo Perrone, University of Oxford

Basic idea. Markov categories are a novel framework to describe and treat problems in probability and information theory. One can combine the categorical formalism with the traditional quantitative notions of entropy, mutual information, and data processing inequalities. Several quantitative aspects of information theory can be captured by an enriched version of Markov categories, where the spaces of morphisms are equipped with a divergence or even a metric.

For instance, Markov categories give a notion of determinism for sources and channels, and we can define entropy exactly by measuring how far a source or channel is from being deterministic. This recovers Shannon and Rényi entropies, as well as the Gini-Simpson index used in ecology to quantify diversity, and it can be used to give a conceptual definition of generalized entropy.

Divergences on Markov categories. A *divergence* or *statistical distance* on a set X is a function

$$\begin{aligned} X \times X &\xrightarrow{\;D\;} [0,\infty] \\ (x,y) &\longmapsto D(x \parallel y) \end{aligned}$$

such that $D(x \parallel x) = 0$. In particular, every metric is a divergence.

We can define a category *enriched in divergences* analogously to metrically enriched categories:

Definition 1. *A* divergence on a monoidal category C *amounts to:*

- *For each pair of objects X and Y, a divergence $D_{X,Y}$ on the set of morphisms $X \to Y$, or more briefly just D; such that*
- *The composition of morphisms in the following form*

$$\begin{tikzcd} X \arrow[r, bend left, "f"] \arrow[r, bend right, "f'"'] & Y \arrow[r, bend left, "g"] \arrow[r, bend right, "g'"'] & Z \end{tikzcd}$$

  satisfies the following inequality,

$$D(g \circ f \parallel g' \circ f') \leq D(f \parallel f') + D(g \parallel g'); \tag{1}$$

- *The tensor product of morphisms in the following form*

$$\begin{tikzcd} X \otimes A \arrow[r, bend left, "f\otimes h"] \arrow[r, bend right, "f'\otimes h'"'] & Y \otimes B \end{tikzcd}$$

  satisfies the following inequality,

$$D\big((f \otimes h) \parallel (f' \otimes h')\big) \leq D(f \parallel f') + D(h \parallel h'). \tag{2}$$

We can interpret this definition as the fact that *we can bound the distance between complex configurations in terms of their simpler components*, an enriched version of the principle of compositionality.

Markov categories are particular monoidal categories, and as such, one can enrich them in divergences according to Definition 1 (see [3, Section 2]). Here are two important examples of divergences that we can put on the category $\mathsf{FinStoch}$ of finite alphabets and noisy channels.

Example 1 (The Kullback-Leibler divergence). Let X and Y be finite sets, and let $f, g : X \to Y$ be stochastic matrices. The *relative entropy* or *Kullback-Leibler divergence* between f and g is given by

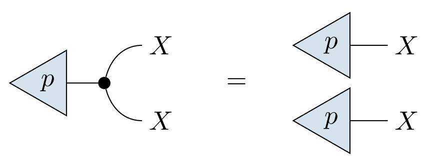
$$D_{KL}(f \parallel g) \coloneqq \max_{x \in X} \sum_{y \in Y} f(y|x) \log \frac{f(y|x)}{g(y|x)},$$

with the convention that $0 \log(0/0) = 0$ and $p \log(p/0) = \infty$ for $p \neq 0$.

Example 2 (The total variation distance). Let X and Y be finite sets, and let $f, g : X \to Y$ be stochastic matrices. The *total variation distance* between f and g is given by

$$D_T(f \parallel g) \coloneqq \max_{x \in X} \frac{1}{2} \sum_{y \in Y} \big| f(y|x) - g(y|x) \big|.$$

Recovering information-theoretic quantities. Recall that a source p on X in a Markov category is called *deterministic* [1, Definition 10.1] if and only if copying its output has the same effect as running it twice independently:



(3)

It is then natural to define as our measure of randomness the discrepancy between the two sides of equation (3).

Definition 2. *Let* C *be a Markov category with divergence* D*. The* entropy *of a source* p *is the quantity*

$$H(p) \coloneqq D\big(\mathrm{copy} \circ p \parallel (p \otimes p)\big), \tag{4}$$

i.e. the divergence between the two sides of (3)*. (Note that the order matters.)*

If we equip $\mathsf{FinStoch}$, with the KL divergence, our notion of entropy recovers exactly Shannon's entropy:

$$H_{KL}(p) = -\sum_{x \in X} p(x) \, \log p(x).$$

If we instead use the total variation distance, our notion of entropy gives the Gini-Simpson index [2], used for example in ecology to quantify diversity:

$$H_T = 1 - \sum_{x \in X} p(x)^2.$$

This approach to information theory recovers other quantities as well, but it also poses a number of open problems. We refer the interested reader to [3].

References

1. Fritz, T.: A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. Adv. Math. **370**, 107239 (2020), arXiv:1908.07021
2. Leinster, T.: Entropy and Diversity. Cambridge University Press (2021)
3. Perrone, P.: Markov categories and entropy (2022), arXiv:2212.11719