Categorical Information Geometry

Paolo Perrone(✉)

University of Oxford, Oxford, UK
paoloperrone@cs.ox.ac.uk
http://www.paoloperrone.org

Abstract. Information geometry is the study of interactions between random variables by means of metric, divergences, and their geometry. Categorical probability has a similar aim, but uses algebraic structures, primarily monoidal categories, for that purpose. As recent work shows, we can unify the two approaches by means of enriched category theory into a single formalism, and recover important information-theoretic quantities and results, such as entropy and data processing inequalities.

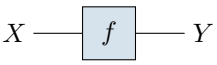
Keywords: Category Theory · Markov Categories · Graphical Models · Divergences · Information Geometry

1 Metrics and Divergences on Monoidal Categories

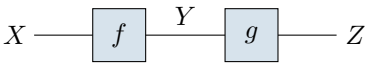
A *monoidal category* [15, Sect. VII.1] is an algebraic structure used to describe processes that can be composed both sequentially and in parallel. Before introducing their metric enrichment, we sketch their fundamental aspects and their graphical representation. See the reference above for the full definition and for the details.

1.1 Monoidal Categories and Their Graphical Calculus

First of all, a category C consists of *objects*, which we can view as spaces of possible states, or alphabets, and which we denote by capital letters such as X, Y, A, B. We also have *morphisms* or *arrows* between them. A morphism $f : A \to B$ can be seen as a process or a channel with input from A and output in B. Graphically, we represent objects as wires and morphisms as boxes, to be read from left to right.



Morphisms can be composed sequentially, with their composition represented as follows,

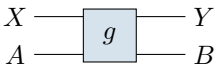


Supported by Sam Staton's ERC grant "BLaSt – A Better Language for Statistics".

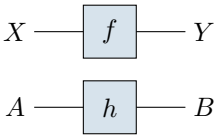
F. Nielsen and F. Barbaresco (Eds.): GSI 2023, LNCS 14071, pp. 268–277, 2023.
https://doi.org/10.1007/978-3-031-38271-0_27

and having a *category* means that the composition is associative and unital. A relevant example is the category $\mathsf{FinStoch}$ of finite sets, which we view as finite alphabets, and stochastic matrices between them, which we view as noisy channels. A stochastic matrix $f : X \to Y$ is a matrix of entries $f(y|x) \in [0, 1]$, which we can view as transition probabilities, such that $\sum_y f(y|x) = 1$ for each $x \in X$.

A *monoidal structure* on C is what allows us to have morphisms with several inputs and outputs, represented as follows.

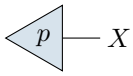


This is accomplished by forming, for each two object X and A, a new object, which we call $X \otimes A$. This assignment is moreover *functorial*, in the sense of a two-variable functor $\otimes : \mathsf{C} \times \mathsf{C} \to \mathsf{C}$, meaning that we also multiply *morphisms*. Given $f : X \to Y$ and $g : A \to B$ we get a morphism $f \otimes g : X \otimes A \to Y \otimes B$, which we represent as follows,



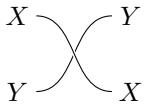
and which we can interpret as executing f and g independently and in parallel.

We also have morphisms with *no* inputs or outputs. For example, a *state* or *source* is a morphism with no inputs.



This is accomplished by means of a distinguished object I, called the *unit*, with the property that $X \otimes I \cong I \otimes X \cong X$, so that it behaves similarly to a neutral element for the tensor product. In $\mathsf{FinStoch}$, I is the one-element set: stochastic matrices $I \to X$ are simply probability measures on X.

A monoidal category is then a category C equipped with a distinguished object I, called the *unit*, and a product functor $\otimes : \mathsf{C} \times \mathsf{C} \to \mathsf{C}$ which is associative and unital up to particular isomorphisms. This makes the structure analogous to a monoid, hence the name. A monoidal category is *symmetric* whenever there is a particular involutive isomorphism



for each pair of objects X, Y, analogously to commutative monoids. For the details, see once again [15, Sect. VII.1].

Let's now equip these structures with metrics and divergences.

1.2 Metrics and Divergences, and a Fundamental Principle

Definition 1. *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 $D(x \parallel x) = 0$.

We call the pair (X, D) a divergence space.

We call the divergence D strict *if $D(x \parallel y) = 0$ implies $x = y$.*

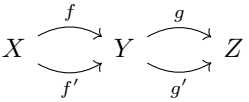
Every metric is a strict divergence which is moreover finite, symmetric, and satisfies a triangle inequality. The Kullback-Leibler divergence (see the next section) is an example of a non-metric divergence.

Definition 2. *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

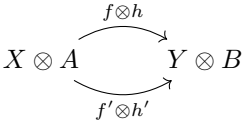
- *The composition of morphisms in the following form*



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*

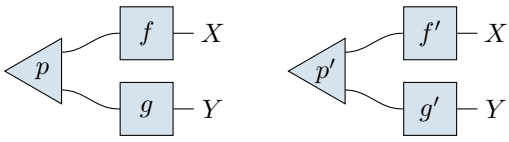


satisfies the following inequality,

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

We can interpret this definition in terms of the following *fundamental principle of categorical information geometry*: *We can bound the distance between complex configurations in terms of their simpler components.*

For example, the distance or divergence between the two systems depicted below



is bounded by $D(p,p') + D(f,f') + D(g,g')$. More generally, for any string diagrams of any configuration, the distance or divergence between the resulting constructions will always be bounded by the divergence between the basic building blocks.

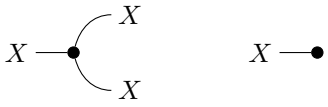
An important consequence of this principle, which can be obtained by setting, for example, $p = p'$ and $f = f'$ but not $g = g'$ in the example above, is that *adding the same block to both sides, in any sequential or parallel direction, cannot increase the distance or divergence*. This is a wide generalization of Shannon's data processing inequalities, which say that the divergence between two sources cannot increase by processing them in the same way. (See [18, Sect. 2.1])

In the next two sections we are going to see two main examples of monoidal categories with divergences: Markov categories, in particular FinStoch, and categories of couplings.

2 Markov Categories and Divergences

Markov categories are particular monoidal categories with a structure that makes them very well suited for modeling probabilistic processes.[1] They were defined in their current form in [5], building up on previous work (see Sect. 2.1).

A Markov category is a symmetric monoidal category where each object X is equipped with two particular maps called "copy" and "discard", and represented as follows.



Note that the copy map has output $X \otimes X$ and the discard map has output I, i.e. "no output". These maps have to satisfy some properties (commutative comonoid axioms) which ensure that the interpretation as "copy" and "discard" maps is indeed consistent. See [5] as well as [18] for more details on this.

Example 1 (The category FinStoch*).* We can construct a category of finite alphabets and noisy channels, called FinStoch, as follows.

- Its objects are finite sets, which we denote by X, Y, Z, etc.
- A morphism $X \to Y$ is a *stochastic matrix*, i.e. a matrix of nonnegative entries with columns indexed by the elements of X, and rows indexed by the elements of Y,

$$\begin{aligned} X \times Y &\xrightarrow{\ f\ } [0,1] \\ (x,y) &\longmapsto f(y|x) \end{aligned}$$

[1] Despite the name, Markov categories are not only suited to model Markov processes, but arbitrary stochastic processes. Indeed, arbitrary joint distributions can be formed, and the *Markov property* states that the stochastic dependencies between the variables are faithfully represented by a particular graph. If the graph is (equivalent to) a single chain, we have a Markov process. In general, the graph is more complex. In this respect, Markov categories are similar to, but more general than, Markov random fields. See [6] for more details on this.

such that each column sums to one,

$$\sum_{y \in Y} f(y|x) = 1 \qquad \text{for every } x \in X.$$

We can interpret $f(y|x)$ as a conditional or transition probability from state $x \in X$ to state $y \in Y$, or we can interpret f as a family of probability measures f_x over Y indexed by the elements of X.

FinStoch is a Markov category with the copy maps $X \to X \otimes X$ given by mapping x to (x, x) deterministically for each $x \in X$, and discard maps given by the unique stochastic matrix $X \to 1$, i.e. a row matrix of entries 1.

2.1 A Brief History of the Idea

The first known study of some aspects of probability theory via categorical methods is due to Lawvere [12], where he defined the category FinStoch outlined above, as well as its generalization to arbitrary measurable spaces. Some of those ideas reappeared in the work of Giry [9] in terms of monads. Lawvere was also the first to see the potential of enriched category theory in metric geometry [13], although it seems that he never used these ideas to study probability theory.

The same categories of probabilistic mappings were defined independently by Chentsov [1], and used to set the stage for the (differential) geometry of probability distributions [2]. Interestingly, Chentsov's work involves *categories* of probabilistic mappings as well as their *geometry*, but he never merged the two approaches into a geometric enrichment of the category of kernels (most likely because at that time, enriched category theory was still in its infancy).[2] The influence of Chentsov on the present work is therefore two-fold, and the main challenge of this work is integrating his two approaches, geometric and categorical, into one unified formalism.

Markov categories, and the more general GS or CD categories, first appeared in [8] in the context of graph rewriting. Similar structures reappeared independently in the work of Golubtsov [10], and were applied for the first time to probability, statistics and information theory. The idea of using "copy" and "discard" maps to study probability came independently to several other authors, most likely initially unaware of each other's work, such as Fong [4], Cho and Jacobs [3], and Fritz [5]. (Here we follow the conventions and terminology of [5].)

Finally, the idea to use both category theory and geometry to study the properties of entropy was inspired by the work of Gromov [11]. This work has a similar philosophy, but follows a different approach.

For more information on the history of these ideas, we refer the reader to [5, Introduction], to [7, Remark 2.2], and to [18, Introduction].

[2] The *geometry of the category of Markov kernels* studied by Chentsov in [2, Sects. 4 and 6] is not *metric geometry*, it is a study of invariants in the sense of Klein's Erlangen Program. More related to the present work are, rather, the *invariant information characteristics* of Sect. 8 of [2]. Much of classical information geometry, and hence indirectly this work, is built upon those notions.

2.2 Divergences on Markov Categories

As Markov categories are monoidal categories, one can enrich them in divergences according to Definition 2 (see also [18, Sect. 2]).

Here are two important examples of divergences that we can put on FinStoch.

Example 2 (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

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

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

Example 3 (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

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

See [18] for why these examples satisfy the conditions of Definition 2. This in particular implies that all these quantities satisfy a very general version of the data processing inequality, see the reference above for more information.

Remark 1. It is well known that the KL divergence and the total variation distance, as well as Rényi's α-divergences, are special cases of *f-divergences* [16]. It is still an open question whether all f-divergences give an enrichment on FinStoch. However, Tsallis' q-divergences do not [18, Sect. 2.3.4].

3 Categories of Couplings and Divergences

Besides Markov categories, another example of divergence-enriched categories relevant for the purposes of information theory are *categories of couplings*. The category FinCoup has

- As objects, finite probability spaces, i.e. pairs (X, p) where X is a finite set and p is a probability distribution on it;
- As morphisms $(X, p) \to (Y, q)$, *couplings* of p and q, i.e. probability measures s on $X \otimes Y$ which have p and q as their respective marginals;
- The identity $(X, p) \to (X, p)$ is given by the pushforward of p along the diagonal map $X \to X \otimes X$;
- The composition of couplings is given by the *conditional product*: for $s : (X, p) \to (Y, q)$ and $y : (Y, q) \to (Z, r)$

$$(t \circ s)(x, z) := \sum_y \frac{s(x, y)\, t(y, z)}{q(y)},$$

where the sum is taken over the $y \in Y$ such that $q(y) > 0$.

More information about this category, and its generalization to the continuous case, can be found in [17].

The two choices of divergence outlined in the previous section also work for the category FinCoup.

Example 4 (Kullback-Leibler divergence). Let (X, p) and (Y, q) be finite probability spaces, and let s and t be couplings of p and q. The Kullback-Leibler divergence

$$D_{KL}(s \parallel t) \coloneqq \sum_{x,y} s(x,y) \log \frac{s(x,y)}{t(x,y)}$$

can be extended to a divergence on the whole of FinCoup, i.e. the conditions of Definition 2 are satisfied.

Example 5 (Total variation distance). Let (X, p) and (Y, q) be finite probability spaces, and let s and t be couplings of p and q. The total variation distance

$$D_T(s,t) \coloneqq \frac{1}{2} \sum_{x,y} \big| s(x,y) - t(x,y) \big|$$

can be extended to a divergence on the whole of FinCoup, i.e. the conditions of Definition 2 are satisfied.

The category FinCoup is moreover an *enriched dagger category*. A coupling $(X, p) \to (Y, q)$ can also be seen as a coupling $(Y, q) \to (X, p)$, and this choice does not have any effect on the metrics or divergences. This property is analogous to, but independent from, the symmetry of the distance in a metric space.

Categories of couplings and Markov categories are tightly related, for more information see [5, Definition 13.7 and Proposition 13.8]. Further links between the two structures will be established in future work.

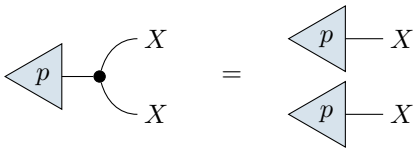
4 Recovering Information-Theoretic Quantities

One of the most interesting features of categorical information geometry is that *basic information-theoretic quantities can be recovered from categorical prime principles*. These include Shannon's entropy and mutual information for discrete sources. Here are some examples, more details can be found in [18].

4.1 Measures of Randomness

Markov categories come equipped with a notion of *deterministic morphisms*, [5, Definition 10.1]. Let's review here the version for sources.

Definition 3. *A source p on X in a Markov category is called* deterministic *if and only if copying its output has the same effect as running it twice independently:*



(3)

Let's try to interpret this notion. First of all, if p is a source which outputs deterministically a single element x of X, then both sides output the ordered pair (x, x), and hence they are equal. Instead, if p is random, the left-hand side will have perfectly correlated output, while the right-hand side will display identically distributed, but independent outputs.

In $\mathsf{FinStoch}$ Eq. (3) reduces to

$$p(x) = p(x)^2$$

for all $x \in X$, so that deterministic sources are precisely those probability distributions p whose entries are only zero and one, i.e. the "Dirac deltas". It is then natural to define as our measure of randomness the discrepancy between the two sides of Eq. (3).

Definition 4. *Let* C *be a Markov category with divergence* D*. The* entropy *of a source* p *is the quantity*

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

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

Example 6. In $\mathsf{FinStoch}$, equipped with the KL divergence, our notion of entropy recovers exactly Shannon's entropy:

$$\begin{aligned} H_{KL}(p) &= D_{KL}\big(\operatorname{copy} \circ p \parallel (p \otimes p)\big) \\ &= \sum_{x,x' \in X} p(x)\, \delta_{x,x'}\, \log \frac{p(x)\, \delta_{x,x'}}{p(x)\, p(x')} \\ &= -\sum_{x \in X} p(x)\, \log p(x). \end{aligned}$$

Example 7. $\mathsf{FinStoch}$, equipped with the total variation distance, our notion of entropy gives the Gini-Simpson index [14], used for example in ecology to quantify diversity:

$$\begin{aligned} H_T &= \frac{1}{2} \sum_{x,x' \in X} \big| p(x)\, \delta_{x,x'} - p(x)\, p(x') \big| \\ &= \frac{1}{2} \sum_{x \in X} p(x) \left(1 - p(x) + \sum_{x' \neq x} p(x') \right) \\ &= 1 - \sum_{x \in X} p(x)^2. \end{aligned}$$

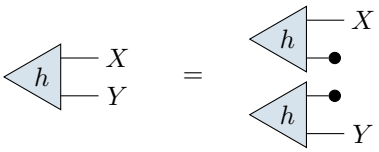
Rényi's α-entropies can also be obtained in this way (see [18, Sect. 4.2.2]), while it is still unclear whether Tsallis' q-entropies can be obtained in this way for $q \neq 2$ (see [18, Question 4.4]).

The fundamental principle of Sect. 1 implies a data processing inequality for entropy generalizing the traditional one. See [18, Sect. 4] for more details.

4.2 Measures of Stochastic Interaction

Just as for determinism, Markov categories are equipped with a notion of stochastic and conditional independence [5, Definition 12.12 and Lemma 12.11]. For sources it reads as follows.

Definition 5. *A joint source h on $X \otimes Y$ in a Markov category displays* independence *between X and Y if and only if*



(5)

For discrete probability measures, this is exactly the condition

$$p(x, y) = p(x)\, p(y),$$

i.e. that p is the product of its marginals. It is a natural procedure in information theory to quantify the stochastic dependence of the variables X and Y by taking the divergence between both sides of the equation.

Definition 6. *Let C be a Markov category with a divergence D. The* mutual information *displayed by a joint source h on $X \otimes Y$ is the divergence between the two sides of Eq. (5),*

$$\mathrm{I}_D(h) \coloneqq D\big(h \parallel (h_X \otimes h_Y)\big).$$

Note that the order of the arguments of D matters.

In $\mathsf{FinStoch}$, with the KL divergence, one recovers exactly Shannon's mutual information. This is well known fact in information theory, and through our formalism, it acquires categorical significance. Using other notion of divergences one can obtain other analogues of mutual information, such as a total variation-based one. Moreover, once again the fundamental principle of Sect. 1 implies a data processing inequality for mutual information generalizing the traditional one. See [18, Sect. 3] for more on this.

Acknowledgements. The author would like to thank Tobias Fritz, Tomáš Gonda and Sam Staton for the helpful discussions and feedback, the anonymous reviewers for their constructive comments, and Swaraj Dash for the help with translating from Russian.

References

1. Chentsov, N.N.: The categories of mathematical statistics. Dokl. Akad. Nauk SSSR **164**, 511–514 (1965)
2. Chentsov, N.N.: Statistical decision rules and optimal inference. Nauka (1972)

3. Cho, K., Jacobs, B.: Disintegration and Bayesian inversion via string diagrams. Math. Struct. Comput. Sci. **29**, 938–971 (2019). https://doi.org/10.1017/S0960129518000488
4. Fong, B.: Causal theories: a categorical perspective on Bayesian networks. Master's thesis, University of Oxford (2012). arXiv:1301.6201
5. Fritz, T.: A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. Adv. Math. **370**, 107239 (2020). arXiv:1908.07021
6. Fritz, T., Klingler, A.: The d-separation criterion in categorical probability (2022). arXiv:2207.05740
7. Fritz, T., Liang, W.: Free GS-monoidal category and free Markov categories (2022). arXiv:2204.02284
8. Gadducci, F.: On the algebraic approach to concurrent term rewriting. Ph.D. thesis, University of Pisa (1996)
9. Giry, M.: A categorical approach to probability theory. In: Banaschewski, B. (ed.) Categorical Aspects of Topology and Analysis. LNM, vol. 915, pp. 68–85. Springer, Heidelberg (1982). https://doi.org/10.1007/BFb0092872
10. Golubtsov, P.V.: Axiomatic description of categories of information transformers. Problemy Peredachi Informatsii **35**(3), 80–98 (1999)
11. Gromov, M.: In search for a structure, Part 1: on entropy (2013). https://www.ihes.fr/ gromov/wp-content/uploads/2018/08/structre-serch-entropy-july5-2012.pdf
12. Lawvere, F.W.: The category of probabilistic mappings (1962). Unpublished notes
13. Lawvere, W.: Metric spaces, generalized logic and closed categories. Rendiconti del seminario matematico e fisico di Milano **43** (1973). http://www.tac.mta.ca/tac/reprints/articles/1/tr1abs.html
14. Leinster, T.: Entropy and Diversity. Cambridge University Press (2021)
15. Mac Lane, S.: Categories for the Working Mathematician. Graduate Texts in Mathematics, vol. 5, 2nd edn. Springer, New York (1998). https://doi.org/10.1007/978-1-4757-4721-8
16. Morozova, E., Chentsov, N.N.: Natural geometry on families of probability laws. Itogi Nauki i Tekhniki. Sovremennye Problemy Matematiki. Fundamental'nye Napravleniya **83**, 133–265 (1991)
17. Perrone, P.: Lifting couplings in Wasserstein spaces (2021). arXiv:2110.06591
18. Perrone, P.: Markov categories and entropy (2022). arXiv:2212.11719