Compositional Abstraction Error and a Category of Causal Models (Supplementary material)

Eigil F. Rischel[1] **Sebastian Weichwald**[2]

[1]Department of Computer & Information Sciences, University of Strathclyde, United Kingdom
[2]Department of Mathematical Sciences, University of Copenhagen, Denmark

Appendix

Table of Contents

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

A KL-DIVERGENCE IS ARBITRARILY FAR FROM SATISFYING THE TRIANGLE INEQUALITY

We recall the definition of Kullback-Leibler divergence.

Definition A.1 (Kullback-Leibler divergence). *For discrete probability distributions q, p on the same probability space $\mathbf{X}$, the KL-divergence from q to p is defined as $D(p|q) = \sum_{x \in \mathbf{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$.*

Proposition A.2 (KL-divergence fails the triangle inequality). *For any $\epsilon > 0$, there exist probability measures p, q, r on $\mathbb{N}$ so that*

$$D(q|p) < \epsilon \text{ and } D(r|q) < \epsilon \text{ while } D(r|p) = \infty.$$

Proof. Let $p(x) = 1/2^x$, $q(x) = Z/x^3$, and $r(x) = Z'/x^2$, where Z, Z' are normalization constants so that

$$\sum_{x=1}^{\infty} q(x) = \sum_{x=1}^{\infty} r(x) = 1.$$

Observe that

$$\begin{aligned} D(q|p) &= \sum_{x \in \mathbb{N}} q(x) \ln\left(\frac{q(x)}{p(x)}\right) = \sum_{x \in \mathbb{N}} \ln\left(\frac{2^x Z}{x^3}\right) \frac{Z}{x^3} \\ \text{and} \quad D(r|q) &= \sum_{x \in \mathbb{N}} r(x) \ln\left(\frac{r(x)}{q(x)}\right) = \sum_{x \in \mathbb{N}} \ln\left(\frac{x^3 Z'}{x^2 Z}\right) \frac{Z'}{x^2}, \end{aligned}$$

both series of which are convergent. Thence, $D(q|p)$ and $D(r|q)$ are both bounded (the specific bound is not important here).

On the other hand,

$$D(r|p) = \sum_{x \in \mathbb{N}} r(x) \ln\left(\frac{r(x)}{p(x)}\right) = \sum_{x \in \mathbb{N}} \ln\left(\frac{2^x Z'}{x^2}\right) \frac{Z'}{x^2},$$

which is *divergent*. Thence, $D(r|p) = \infty$.

Now for each $\alpha \in [0, 1]$, let $p_\alpha(x) = Y_\alpha \left(\frac{\alpha}{p(x)} + \frac{1-\alpha}{q(x)}\right)^{-1}$, and $r_\alpha(x) = \alpha r(x) + (1 - \alpha) q(x)$, where Y_α is a suitable normalization constant so that these are probability distributions. The series involved in computing the normalisation constant for p_α is convergent by the harmonic-arithmetic inequality, which also implies that $Y_\alpha \geq 1$. Y_α is a continuous function of α.

Now the claim is that $D(q|p_\alpha), D(r_\alpha|q) \to 0$ as $\alpha \to 0$, while $D(r_\alpha|p_\alpha) = \infty$ for all $\alpha > 0$.

First consider

$$D(q|p_\alpha) = \sum_{x \in \mathbb{N}} \frac{Z}{x^3} \ln\left(Z \frac{\alpha 2^x + (1-\alpha) x^3/Z}{Y_\alpha x^3}\right)$$

Since $\ln$ is increasing, we can bound this sum by

$$\sum_{x \in \mathbb{N}} \frac{Z}{x^3} \ln\left(Z \frac{\max(2^x, x^3/Z)}{Y_\alpha x^3}\right)$$

Since $2^x Z$ dominates for x large enough, this sum converges. Let $\delta > 0$ be given. Choose M large enough that $\sum_{x=M}^{\infty} \frac{Z}{x^3} \ln\left(Z \frac{\max(2^x, x^3/Z)}{Y_\alpha x^3}\right) < \delta/2$. Note that M is independent of α – this is possible because $Y_\alpha \geq 1$. Then we also have

$$\sum_{x=M}^{\infty} \frac{Z}{x^3} \ln\left(Z \frac{\alpha 2^x + (1-\alpha) x^3/Z}{Y_\alpha x^3}\right) < \delta/2$$

Now each term of the sum goes to zero as $\alpha \to 0$, since $Y_0 = 1$ and $\alpha \mapsto Y_\alpha$ is continuous. Hence we can choose α small enough that the sum of the first $M - 1$ terms is $< \delta/2$. This proves that the whole sum goes to 0 as $\alpha \to 0$.

For the case of $D(r_\alpha|q)$, we have

$$D(r_\alpha|q) = \sum_{x\in\mathbb{N}} \big(\alpha r(x) + (1-\alpha)q(x)\big)\ln\left(\frac{\alpha r(x)+(1-\alpha)q(x)}{q(x)}\right)$$

We can use an analogous argument. We can use monotonicity and the fact that $\frac{Z'}{x^2}$ is eventually bigger than $\frac{Z}{x^3}$ to prove convergence, with convergence speed independent of α, so that the tail is eventually $< \delta/2$ independently of α, then choose α to bound the head.

Now consider $D(r_\alpha|p_\alpha)$, which we can rewrite as

$$D(r_\alpha|p_\alpha) = \sum_{x\in\mathbb{N}} \left(\frac{\alpha Z'}{x^2} + \frac{(1-\alpha)Z}{x^3}\right)\ln\left(\frac{\frac{\alpha Z'}{x^2}+\frac{(1-\alpha)Z}{x^3}}{\frac{Y_\alpha}{\alpha 2^x+(1-\alpha)x^3/Z}}\right).$$

First, we see that this is larger than $\sum_x \frac{\alpha Z'}{x^2}\ln\left(\frac{\frac{\alpha Z'}{x^2}}{\frac{Y_\alpha}{\alpha 2^x+(1-\alpha)x^3}}\right)$.

We can rewrite this as

$$\sum_{x\in\mathbb{N}} \frac{\alpha Z'}{x^2}\ln(aZ'\frac{\alpha 2^x+(1-\alpha)x^3}{Yx^2})$$

$$\geq \sum_{x\mathbb{N}} \frac{\alpha Z'}{x^2}\ln\left(\alpha Z'\frac{\alpha 2^x}{Y_\alpha x^2}\right)$$

As above, this diverges independently of the value of α (as long as $\alpha > 0$). Hence $D(r_\alpha|p_\alpha) = \infty$ for all $\alpha > 0$.

Hence for sufficiently small but positive α, the distributions p_α, q, r_α satisfy the desired properties.

□

One might hope that this rather stark result is just a quirk of the infinities involved, disappearing when we restrict to finite probability measures. Since KL-divergences on finite sets are always *finite* (as long as the measures involved have the same support), we cannot reproduce the infinity in the above result on finite sets. However, we can come arbitrarily close, in the following sense:

Proposition A.3 (Instance of KL-divergence failing). *For any $\epsilon, K > 0$, there exist a finite set $\{1,\dots N\}$ and probability measures p, q, r on it so that $D(q|p), D(r|q) < \epsilon, D(r|p) > K$.*

Proof. Let p_α, q, r_α be as before. Let $p_\alpha^N, q^N, r_\alpha^N$ be the distributions on $\{1,\dots N\}$ so that $p_\alpha^N(x) = p_\alpha(x)$ for $x \in \{1,\dots N-1\}$, etc. Then $D(p_\alpha^N|r_\alpha^N) \to \infty$ as $N \to \infty$ – the sum computing this divergence is $N-1$ terms of the sum computing the overall (infinite) divergence, and a remainder term which is certainly positive for N large enough.

On the other hand, we now show that $D(p_\alpha^N|q^N) \to D(p_\alpha|q)$ as $N \to \infty$. To see this, consider the sum in question:

$$D(p_\alpha^N|q^N) = \sum_{x=1}^{N-1} p_\alpha(x)\ln(p_\alpha(x)/q(x)) + \left(\sum_{x=N}^{\infty} p_\alpha(x)\right)\ln\left(\frac{\sum_{x=N}^\infty p_\alpha(x)}{\sum_{x=N}^\infty q(x)}\right)$$

The first term here converges to $D(p_\alpha|q)$, so it suffices to show that the remainder converges to zero.

We can write this remainder as $(\sum_{x=N}^\infty p_\alpha(x))(\ln(\sum_{x=N}^\infty p_{\alpha(x)}) - \ln(\sum_{x=N}^\infty q(x)))$

Observe that we can write

$$D(p_\alpha|q) = \sum_x p_\alpha(x)\ln(p_\alpha(x)) - \sum_x p_\alpha(x)\ln(q(x))$$

Since the left-hand side is finite, and the first sum is convergent, so is the second sum.

Hence we can use convexity of $\ln$ to obtain the following inequality:

$$\leq \left(\sum_{x=N}^{\infty} p_\alpha(x)\right) \ln \left(\sum_{x=N}^{\infty} p_\alpha(x)\right) - \sum_{x=N}^{\infty} p_\alpha(x) \ln(q(x))$$

Now convergence means that as $N \to \infty$ the second term goes to zero. And since $\sum_{x=N}^{\infty} p_\alpha(x) \to 0$, and $x \ln(x) \to 0$ when $x \to 0$, so does the first term.

The analogous argument verifies the same statement for $D(q^N | r_\alpha^N)$.

Thus, by choosing α small enough and N big enough, we obtain the desired measures. □

B STRING-DIAGRAMMATIC CONSTRUCTION OF INTERVENTIONAL DISTRIBUTIONS

In Definition 2.3, we claim that an interventional model has a well-defined interventional distribution, a kernel $I_S : \prod_{v \in S} \mathbf{X}_v^M \to \prod_{v \in V(M)} \mathbf{X}_v^M$ for any subset $S \subseteq V(M)$.

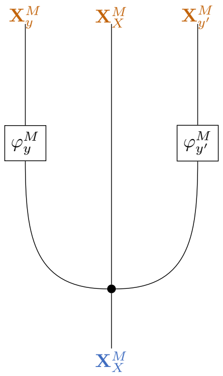
The intuition is as follows. To sample $(y_v)_{v \in V(M)}$ according to $I_S(x_v)_{v \in S}$, we do the following:

1. If $v \in S$, $y_v = x_v$ with probability 1.
2. Identify a variable v so that y_v has not been determined yet, but all its parents have been determined, and sample y_v according to $\varphi_v^M((y_{v'})_{v' \to v})$.

The question is whether this gives a well-defined distribution. We here prove that it does.

For convenience, we write $\mathbf{X}_X^M = \prod_{v \in X} \mathbf{X}_v^M$ when $X \subset V(M)$ is a subset of variables of M and $\mathrm{pa}(v) \subseteq V(M)$ for the set of parents of $v \in V(M)$.

We use the graphical notation known as *string diagrams*. These are widely used in category theory to depict constructions in *monoidal categories*. A full discussion of the technical details behind their use is beyond the scope of this appendix (see, for example, Selinger [2011]). Their meaning in the special case under consideration, kernels between finite sets, can be intuitively understood. For example, the following diagram (read bottom-to-top)



depicts a kernel $\mathbf{X}_X^M \to \mathbf{X}_y^M \times \mathbf{X}_X^M \times \mathbf{X}_{y'}^M$, informally described as "given $x \in \mathbf{X}_X^M$, sample $y \in \mathbf{X}_y$ from the distribution $\varphi_y^M(x)$ and independently sample y' from the distribution $\varphi_{y'}^M(x)$, then return the tuple (y, x, y')".

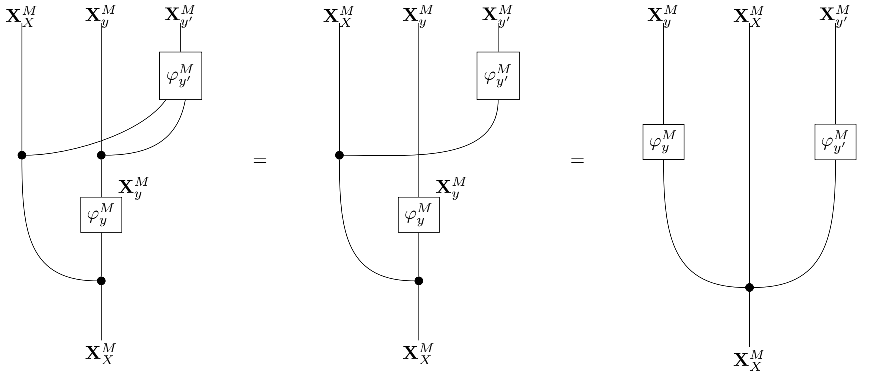
Proposition B.1. *The interventional distribution in Definition 2.3 is well-defined.*

For this proposition, we need the following lemma.

Lemma B.2. *Let S be a finite partially ordered set. Let $A : s_1 \ldots s_n$ and $B : s'_1 \ldots s'_n$ be two totalizations of the ordering on S — in other words, two ways of arranging the elements of S in a non-decreasing sequence. Then one can turn A into B by a finite sequence of transpositions, where each transposition exchanges two adjacent, incomparable elements.*

Proof. Let's show that any nondecreasing sequence can be turned into B by such a sequence of transpositions — this is really the content of the lemma. Define the error of a sequence $s_1 \ldots s_n$ as the total number of pairs i, j so that s_i and s_j are not in the same order as in B. If the error is zero, we must already be in sequence B. Suppose the error is greater than zero. Then there must be a pair of consecutive elements, s_i, s_{i+1}, that are in the wrong order compared to the ordering B. The elements must also be incomparable:
1) we cannot have $s_{i+1} \leq s_i$, since it is a non-decreasing sequence,
2) we cannot have $s_i \leq s_{i+1}$ — if this was true, the pair would already be in the same order as in B.
Hence we can swap s_i and s_{i+1}, which decreases the error by 1. After a finite number of steps the error must be zero and we have obtained B. □

Proof of Proposition B.1. By applying Lemma B.2 to the vertices of the DAG G^M, partially ordered by causal dependence, we see that we can move between any two constructions of the interventional distribution by swapping two consecutive variables at a time. Hence it suffices to show that we may swap the order of two consecutive y_i, neither dependent on the other, without changing the final distribution. Consider the following diagram manipulation:



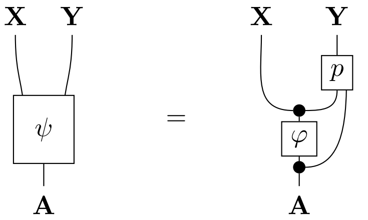
(B.1)

In the first step we use the fact that y' does not depend on y, so we may delete the $\mathbf{X}_y^M$ input to $\varphi_{y'}^M$. Then rearrange the wires.

This shows that the composition $\mathbf{X}_X^M \to \mathbf{X}_{X\cup\{y\}}^M \to \mathbf{X}_{X\cup\{y,y'\}}^M$ is equal to another map $\mathbf{X}_X \to \mathbf{X}_{X\cup\{y,y'\}}$. A similar argument shows that the composition $\mathbf{X}_X \to \mathbf{X}_{X\cup\{y'\}} \to \mathbf{X}_{X\cup\{y,y'\}}^M$ is equal to the same map. This concludes our proof. □

We also prove our claim that this distribution is "the right one", in the sense that the mechanisms are the conditional distributions. We introduce the following diagram-theoretic definition of conditionals:

Definition B.3. *Let $\psi : \mathbf{A} \to \mathbf{X} \times \mathbf{Y}$ be a Markov kernel. We say that a kernel $p : \mathbf{A} \times \mathbf{X} \to \mathbf{Y}$ is a* conditional distribution of $Y \in \mathbf{Y}$ given $A \in \mathbf{A}$ and $X \in \mathbf{X}$, *if there exists $\varphi : \mathbf{A} \to \mathbf{X}$ so that we have the following identity:*



(B.2)

Remark B.4. Definition B.3 is a definition of conditional distributions suitable for parameterized joint distributions. Dealing with such distributions is necessary if we want to combine conditional and interventional distributions. In the case $\mathbf{A} = \{*\}$, we recover the usual situation of a joint distribution on a product set.

Let us spell out the connection with the normal definition of conditional distribution: a map $p : \mathbf{A} \times \mathbf{X} \to \mathbf{Y}$ is a conditional distribution for $\psi : \mathbf{A} \to \mathbf{X} \times \mathbf{Y}$ if and only if, for all $a \in \mathbf{A}$, and for all $x \in \mathbf{X}$ with nonzero probability given a, the distribution $p(a, x)$ is the conditional distribution of Y given $X = x$ and $(X, Y) \sim \psi(a)$. This is also the reason we say *a* conditional distribution and not *the* conditional distribution.

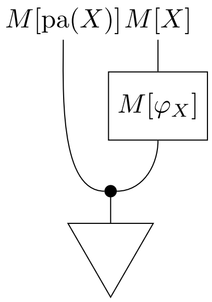
For a more thorough discussion of this point from a categorical point of view, see, for example, Fritz [2020, Section 11 (in particular Definition 11.5 and Remark 11.6)].

Proposition B.5. *Each mechanism $\varphi_X^M : \mathbf{X}_{\mathrm{pa}(v)}^M \to \mathbf{X}_v^M$ is a conditional distribution for the observational distribution $* \to \mathbf{X}_{\mathrm{pa}(V)}^M \times \mathbf{X}_v^M$.*

Proof. Recall that given a distribution $* \to \mathbf{X} \otimes \mathbf{Y}$, a kernel $\mathbf{X} \to \mathbf{Y}$ is a conditional distribution if and only if we have the identity in Equation (B.2). After marginalizing out the other variables, the observational distribution on $\mathbf{X}_{\mathrm{pa}(v)}^M \times \mathbf{X}_v^M$ factors as

$$* \to \mathbf{X}_{\mathrm{pa}(v)}^M \xrightarrow{(1,\varphi_v^M)} \mathbf{X}_{\mathrm{pa}(v)}^M \times \mathbf{X}_v^M.$$

Diagrammatically, this looks like



Which is the statement we wanted. □

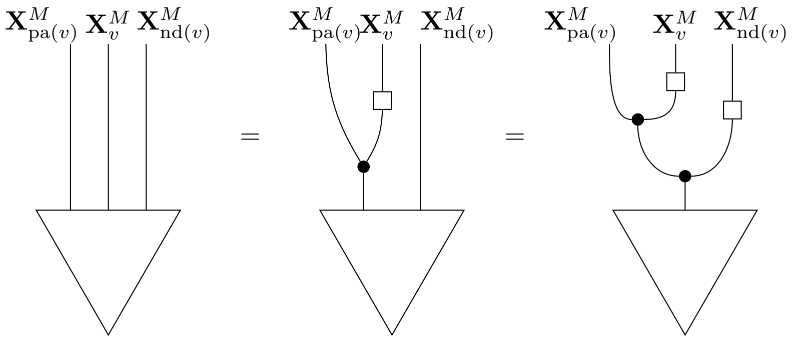
Here the triangle denotes the observational distribution on $\mathbf{X}^M_{\mathrm{pa}(v)}$.

There is nothing surprising about this proposition: just as in the classical theory of graphical models, it holds by construction. In classical treatments, this is usually not made into a theorem, although it is implicit in most treatments of the Markov property for structure causal models, see, for example, Peters et al. [2017, Proposition 6.31].

Proposition B.6. *The observational distribution satisfies the directed Markov property with respect to the DAG G.*

Proof. We must prove that any variable v is independent of its non-descendants given its parents. Let us introduce the notation $\mathrm{nd}(v)$ for the non-descendants of v excluding its parents. Then we are trying to show $\mathrm{nd}(v) \perp v \mid \mathrm{pa}(v)$.

Observe this diagram manipulation:


$\mathbf{X}^M_{\mathrm{pa}(v)}$
$\mathbf{X}^M_v$
$\mathbf{X}^M_{\mathrm{nd}(v)}$
$=$
$\mathbf{X}^M_{\mathrm{pa}(v)}$
$\mathbf{X}^M_v$
$\mathbf{X}^M_{\mathrm{nd}(v)}$
$=$
$\mathbf{X}^M_{\mathrm{pa}(v)}$
$\mathbf{X}^M_v$
$\mathbf{X}^M_{\mathrm{nd}(v)}$

In the first step, we are factoring the observational distribution on $v \cup \mathrm{pa}(v) \cup \mathrm{nd}(v)$ as "sample the parents and non-descendants of v, then sample v conditional on the parents" — according to the definition, this is a possible choice for how to construct the observational distribution.

In the second step, we are factoring the distribution on $\mathrm{pa}(v) \cup \mathrm{nd}(v)$ as "sample the parents of v, then sample the non-descendants of v *according to the conditional distribution*". This is always possible, and yields the diagram on the right.

By Fritz [2020, Remark 12.2], this implies the conditional independence $\mathrm{nd}(v) \perp v \mid \mathrm{pa}(v)$. □

Remark B.7. In fact, Propositions B.5 and B.6 characterize the observational distribution uniquely. This can be proven diagrammatically by using a diagrammatic formulation of Proposition B.6 to show that the observational distribution factorizes as a certain diagram, and then using Proposition B.5 to show that the morphisms in this diagram may be replaced by the mechanisms.

In classical terms, this argument corresponds to arguing that the probability factorizes according to the graph, and that the factors must be precisely the mechanisms.

C JENSEN-SHANNON DIVERGENCE IS COMPOSITIONAL

Proposition 2.8 (kernel composition is short)**.** *The composition of kernels*

$$\mathrm{FinStoch}(\mathbf{X},\mathbf{Y}) \otimes \mathrm{FinStoch}(\mathbf{Y},\mathbf{Z}) \to \mathrm{FinStoch}(\mathbf{X},\mathbf{Z})$$

is a short map, that is,

$$d_{\mathrm{JSD}}(f_1 \circ g_1, f_2 \circ g_2) \leq d_{\mathrm{JSD}}(f_1, f_2) + d_{\mathrm{JSD}}(g_1, g_2)$$

for any $f_1, f_2 \in \mathrm{FinStoch}(\mathbf{Y},\mathbf{Z})$*,* $g_1, g_2 \in \mathrm{FinStoch}(\mathbf{X},\mathbf{Y})$*.*

Proof. Since d_{JSD} is a metric, this is equivalent to the two statements

$$d_{\mathrm{JSD}}(fg_0, fg_1) \leq d_{\mathrm{JSD}}(g_0, g_1)$$

$$d_{\mathrm{JSD}}(fg, f'g) \leq d_{\mathrm{JSD}}(f, f')$$

In each case it suffices to show the given inequality at each $x \in \mathbf{X}$, so we can assume that $\mathbf{X} = *$. Since $x \mapsto \sqrt{x}$ is a monotone map, it suffices to show that

$$\begin{aligned} &\mathrm{JSD}(fp_0, fp_1) \leq \mathrm{JSD}(p_0, p_1) \\ \text{and} \quad &\mathrm{JSD}(f_0 p, f_1 p) \leq \sup_x \mathrm{JSD}(f_0(x), f_1(x)), \end{aligned}$$

where p_0, p_1 are distributions on $\mathbf{Y}$.

The first case follows, since "postprocessing" the random variable X can only reduce its mutual information with B.

In the second case, we are comparing the following two situations:

1. If you learn the value of a random variable sampled from $f_i(x)$, how much do you learn about i, if x is chosen to maximize this amount of information.
2. If you learn the value of a random variable sampled from $f_i(x)$, where x is chosen at random, how much do you learn about i?

To see that the first number is bigger, consider the following: In the second case, even if you were additionally told what x was (giving you *more* information), you would still, at best, be in the first situation. □

Lemma 2.10 (JSD is compositional)**.** *Consider a diagram (not necessarily commutative) in* FinStoch *of the following form:*

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[d, "a"] & A' \\ B \arrow[r, "g"] \arrow[d, "b"] & B' \arrow[u, "b'"'] \\ C \arrow[r, "h"] & C' \arrow[u, "c'"'] \end{tikzcd} \tag{C.1}$$

Then $d_{\mathrm{JSD}}(f, b'c'hba) \leq d_{\mathrm{JSD}}(f, b'ga) + d_{\mathrm{JSD}}(g, c'hb)$*.*

Proof.

$$\begin{aligned} d(f, b'c'hba) &\leq d(f, b'ga) + d(b'ga, b'c'hba) \\ &\leq d(f, b'ga) + d(g, c'hb), \end{aligned}$$

where the first inequality is the triangle inequality and the last one uses Proposition 2.8. □

D ENRICHED CATEGORY THEORY AND ERROR CATEGORIES

An *enriched* category is a generalization of the concept of category, where the *set* of maps $x \to y$, $\mathcal{C}(x, y)$, has been replaced by an object $\mathcal{C}(x, y) \in \mathcal{V}$ in some other, *enriching*, category. For example, the set of linear maps $V \to W$ between two vector spaces can itself be equipped with the structure of a vector space in a canonical way – this defines an *enrichment* of the category of vector spaces in itself.

A full discussion of enriched categories is beyond the scope of the present article; see Kelly [2005] for a comprehensive introduction. The present paper contains two examples of enriched categories.

First, Proposition 2.8 shows that the Jensen-Shannon distance defines an enrichment of $\mathrm{FinStoch}$ in the category Met of metric spaces:

Definition D.1. *The category* Met *of metric spaces is defined as follows:*

1. *The objects are metric spaces.*
2. *A morphism $(X, d_X) \to (Y, d_Y)$ is a function $f : X \to Y$ which is distance nonincreasing (or "short"), meaning that $d_Y(f(x), f(x')) \leq d_X(x, x')$.*
3. *Composition is function composition and the identity morphisms are the identity functions.*
4. *The tensor product of two metric spaces is $(X, d_X) \otimes (Y, d_Y) = (X \times Y, d_X \otimes d_Y)$, defined by $(d_X \otimes d_Y)((x, y), (x', y')) = d_X(x, x') + d_Y(y, y')$.*

Then a category enriched in metric spaces consists of the following data:

1. A category $\mathcal{C}$
2. with a metric $d_{\mathcal{C}(X,Y)}$ on each set of morphisms $\mathcal{C}(X, Y)$ (for each pair of objects $X, Y \in \mathcal{C}$)
3. so that, if $f, f' : X \to Y, h : Y \to Z, g : A \to X$, we have
$$d_{\mathcal{C}(X,Z)}(h \circ f, h \circ f'), d_{\mathcal{C}(A,Y)}(f \circ g, f' \circ g) \leq d_{\mathcal{C}(X,Y)}(f, f').$$
In other words, post- and precomposition are distance nonincreasing maps.

Lemma 2.10 is essentially a lemma about enriched categories – the statement and the proof make sense for any category enriched over Met.

Second, the compositional property of our error measure, Proposition 2.12, can be phrased to say that error defines an enrichment of FinMod in the following category of *error spaces*.

Definition D.2 (Category of error spaces)**.** *An* error space (X, e) *consists of a set X and a function $e : X \to [0, \infty]$. A morphism of error spaces $f : (X, e_X) \to (Y, e_Y)$ is a function $f : X \to Y$ so that $e_Y(f(x)) \leq e_X(x)$ for all $x \in X$. The tensor product $(X, e_X) \otimes (Y, e_Y)$ of two error spaces is $(X \times Y, e_X \otimes e_Y)$, where $e_X \otimes e_Y(x, y) = e_X(x) + e_Y(y)$. This data defines a symmetric monoidal category* Err *of error spaces.*

A category enriched in error spaces, or an Err-category, then consists of the following data:

1. A category $\mathcal{C}$,
2. with an error $e(f)$ for each morphism f in $\mathcal{C}$
3. such that, when f, g are composable, $e(fg) \leq e(f) + e(g)$.

There is an error category of error spaces, $\underline{\mathrm{Err}}$, where the maps $f : (X, e_X) \to (Y, e_Y)$ are *all* functions $f : X \to Y$, and the error of such a function is $e(f) := \max(0, \sup_x e_Y(f(x)) - e_X(x))$; that is, the error is the maximal increase in error created by the function.

In Section 3, we mentioned the notion of a *functor*. A functor is a mapping between categories preserving the compositional structure:

Definition D.3 (Functor)**.** *Given categories $\mathcal{C}, \mathcal{D}$, a functor $F : \mathcal{C} \to \mathcal{D}$ consists of an object $F(X)$ in $\mathcal{D}$ for each object X in $\mathcal{C}$, as well as a morphism $F(f) : F(X) \to F(Y)$ for each $f : X \to Y$, so that $F(f \circ g) = F(f) \circ F(g)$ and $F(1_X) = 1_{F(X)}$.*

There is also a notion of functor between enriched categories. Here, we spell this out for the case of Err-categories: A functor of Err-categories (or Err-functor) $\mathcal{C} \to \mathcal{D}$ then consists of the following data:

1. A functor $F : \mathcal{C} \to \mathcal{D}$
2. such that $e(F(f)) \leq e(f)$.

In particular, an Err-functor $F : \mathcal{C} \to \underline{\mathrm{Err}}$ consists of the following:

1. For each $C \in \mathcal{C}$, an error space $F(C)$,
2. for each map $f : C \to D$, a function $F(C) \to F(D)$
3. such that $F(f \circ g) = F(f) \circ F(g)$ and
4. such that $e_{F(D)}(F(f)(x)) \leq e_{F(C)}(x) + e(f)$.

Thus, this captures the desired properties of the collection of implemented models discussed in Section 3.

References

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

Gregory M. Kelly. *Basic Concepts of Enriched Category Theory*. Number 10. 2005.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

Peter Selinger. *A Survey of Graphical Languages for Monoidal Categories*, pages 289–355. Springer, 2011.