Volume 3 Issue 1 ISSN 2631-4444

Categorical Stochastic Processes and Likelihood

Dan Shiebler

Department for Continuing Education and Department of Computer Science
University of Oxford, Oxford, United Kingdom

We take a category-theoretic perspective on the relationship between probabilistic modeling and gradient based optimization. We define two extensions of function composition to stochastic process subordination: one based on a co-Kleisli category and one based on the parameterization of a category with a Lawvere theory. We show how these extensions relate to the category of Markov kernels **Stoch** through a pushforward procedure.

We extend stochastic processes to parametric statistical models and define a way to compose the likelihood functions of these models. We demonstrate how the maximum likelihood estimation procedure defines a family of identity-on-objects functors from categories of statistical models to the category of supervised learning algorithms **Learn**.

Code to accompany this paper can be found on GitHub[1].

1 Introduction

The explosive success of machine learning over the last two decades has inspired theoretical work aimed at developing rigorous frameworks for reasoning about and extending machine learning algorithms. For example, inspired by the inherent compositional structure at the heart of gradient based optimization, several authors have developed category theoretic frameworks for reasoning about neural networks and automatic differentiation [5; 9; 11; 12]. Separately, one of the most active areas of applied category theory focuses on building a categorical framework for probability theory and statistics. Researchers like Fritz [14], Cho and Jacobs [4], and Culbertson and Sturtz [6; 7] have developed strategies for describing the construction of probabilistic models from data in categorical terms. We aim to bridge these streams of research by using a probabilistic construction to define an optimization objective.

Cho and Jacobs [4] and Culbertson and Sturtz [6; 7] explore how new data points affect their models' **epistemic uncertainty**, or uncertainty due to limited data or knowledge. For example, a simple model of a complex nonlinear system is likely to have high epistemic uncertainty. Another form of uncertainty is **aleatoric uncertainty**, or inherent uncertainty in a system that will cause results to differ each time we run the same experiment. For example, if we aim to predict the output of a system that includes a non-deterministic stage (such as a coin toss), we will need to cope with aleatoric uncertainty.

Aleatoric uncertainty is common in physical systems. For example, many biological processes will produce slightly different results based on randomness in turbulent fluid flows. For this reason, models that approximate physical systems often implicitly or explicitly produce a probability distribution over the possible outputs conditioned on some input [25].

Even models that produce point estimates, such as the ones described by Fong et al. [12], can be viewed as predicting the expected value of some unknown probability distribution. For example, suppose we have some system $X \to y$ that contains a degree of aleatoric uncertainty such that $P(y|X)$ is Gaussian. Now suppose we train a point estimate model that predicts y from X such that the mean square error between the model's predictions and the observations from the execution of this system is minimized. This is approximately equivalent to minimizing the

Dan Shiebler: daniel.shiebler@kellogg.ox.ac.uk, danshiebler.com

[1] https://github.com/dshieble/Categorical_Stochastic_Processes_and_Likelihood

Kullback-Leibler (KL) divergence (which measures how one probability distribution is different from a second, reference probability distribution) between a distribution with expected value given by the model's output and $P(y|X)$. In this way the structure of the model's aleatoric uncertainty is captured in its loss function (mean square error in this case).

Now consider a physical system which has several components, each of which has some degree of aleatoric uncertainty. Suppose we want to build a compositional model for this system. If we use the neural network-like composition of Fong et al. [12], then we can only represent the full model's uncertainty with the loss function that parameterizes the backpropagation functor. As a result, we cannot characterize the interactions between the uncertainty in the different parts of the system.

For example, Eberhardt et al. [8] build a convolutional neural network model to assess how the visual cortex performs a rapid stimulus categorization task. Their model includes multiple layers which represent the hierarchy within the central nervous system from photorecepters in the eye, to edge-detecting neurons in the primary visual cortex, to higher-order feature detectors in the later stages of visual cortex. Although there is aleatoric uncertainty at each layer of this biological system, Eberhardt et al. use a standard composition of neural network layers and therefore can only represent this uncertainty with a cross-entropy loss over the model's final output.

In this paper we describe an alternative strategy for constructing and composing parametric models such that we can explicitly characterize how different subsystems' uncertainties interact. We use this strategy to build a generalized framework for training neural networks that have stochastic processes as layers. To do this, we replace the domain of Fong et al.'s [12] Backpropagation functor (**Para**, also written as **Para**(**Euc**) [16]) with a probabilistically motivated category over which we can define the error function $er : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ through the maximum likelihood procedure. Our specific contributions are to:

- Develop a strategy for composing stochastic processes that is compatible with both subordination [20] and parametric function composition [12].
- Introduce two categories with this compositional structure, one based on **Para**(**Euc**) [16] and one based on the co-Kleisli category of the co-monad $(\Omega \otimes \_)$, and explore their relationships with each other and with the category **Stoch** of Markov kernels.
- Extend the category of stochastic processes to a category of parametric statistical models.
- Demonstrate that the Radon-Nikodym derivative with respect to the Lebesgue measure acts as a semifunctor from a sub-semicategory of parametric statistical models into a semicategory of likelihood functions.
- Define a family of subcategories of parametric statistical models over which we can use the maximum likelihood procedure to define a backpropagation functor into the category **Learn** of learning algorithms [12].

2 Preliminaries

2.1 Probability Measures, Random Variables and Markov Kernels

A **probability space** is a triplet (Ω, Σ, μ) where (Ω, Σ) is a measurable space and μ is a **probability measure** over (Ω, Σ). That is, μ is a countably additive function over the σ-algebra Σ that returns results in the unit interval $[0, 1]$ such that $\mu(\Omega) = 1, \mu(\varnothing) = 0$. Recall that Σ is a set of subsets of Ω. For some topological space Ω, we will write $\mathcal{B}(\Omega)$ for the **Borel algebra** of Ω, or the smallest σ-algebra that contains all open sets.

A **random variable** defined on the probability space (Ω, Σ, μ) is a measurable function from (Ω, Σ) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We will sometimes use the term "random variable" to refer to measurable functions into $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ as well. These are also called multivariate random variables or random vectors. While some authors use uppercase letters like X to denote random variables, we will use lowercase letters like f, g to emphasize that random variables are functions. Given a probability

space $(\Omega, \mathcal{B}(\Omega), \mu)$ and a random variable $f : \Omega \to \mathbb{R}$, the **pushforward** $f_*\mu$ of μ along f is a probability measure over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined to be:

$$f_*\mu(\sigma_\mathbb{R}) = \mu(f^{-1}(\sigma_\mathbb{R})).$$

A **Markov kernel** between the measurable space (A, Σ_A) and the measurable space (B, Σ_B) is a function $\mu : A \times \Sigma_B \to [0, 1]$ such that:

- For all $\sigma_b \in \Sigma_B$, the function $\mu(\_, \sigma_b) : A \to [0, 1]$ is measurable.
- For all $x_a \in A$, $\mu(x_a, \_) : \Sigma_B \to [0, 1]$ is a probability measure on (B, Σ_B). In particular:

$$\mu(x_a, B) = 1 \qquad \mu(x_a, \varnothing) = 0.$$

For example, a Markov Kernel between the one-point set and the measurable space (A, Σ_A) is just a probability measure over (A, Σ_A).

A **stochastic process** defined in the probability space (Ω, Σ, μ) is a family of random variables indexed by some set T. That is, we can write a stochastic process as a function $f : \Omega \times T \to \mathbb{R}$. In this paper we will limit our study to stochastic processes that are jointly Borel-measurable. We can define the pushforward of μ along such a stochastic process f to be the Markov Kernel $f_*\mu : T \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ where for $x_t \in T, \sigma_\mathbb{R} \in \mathcal{B}(\mathbb{R})$:

$$f(x_t, \_)_*\mu(\sigma_\mathbb{R}) = \mu(f(x_t, \_)^{-1}(\sigma_\mathbb{R})).$$

2.2 Categories

A central category that we will work in is the symmetric monoidal category **Meas** of measurable spaces and measurable functions. The objects in **Meas** are pairs (A, Σ_A), where Σ_A is a σ-algebra over A. A morphism from (A, Σ_A) to (B, Σ_B) in **Meas** is a measurable function f such that for any $\sigma_B \in \Sigma_B$, $f^{-1}(\sigma_B) \in \Sigma_A$. The tensor product of the measurable spaces (A, Σ_A) and (B, Σ_B) in **Meas** is the space $(A \times B, \Sigma_A \otimes \Sigma_B)$, where $\Sigma_A \otimes \Sigma_B$ is the product σ-algebra of Σ_A and Σ_B. Note that **Meas** is not cartesian closed. Staton et al. [19] introduce a similar category **QBS** that is cartesian closed. The objects in **QBS** are **quasi-Borel spaces**, or tuples (X, M_X) where X is a set and M_X is a set of functions from $\mathbb{R}$ into X such that:

- If $f \in M_X$ and $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then $f \circ g \in M_X$.
- If $f : \mathbb{R} \to X$ is constant then $f \in M_X$.
- If $\mathbb{R} = \biguplus_{i \in \mathbb{N}} S_i$ such that each set S_i is Borel and $\forall_{i \in \mathbb{N}} f_i : \mathbb{R} \to X \in M_X$, then g is in M_X, where $g(r) = f_i(r)$ for $r \in S_i$.

We will generally work in the following subcategory of **Meas**:

Definition 2.1. Euc *is the strict Cartesian monoidal subcategory of* **Meas** *where objects are restricted to be* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *for some* $n \in \mathbb{N}$ *and morphisms are restricted to be continuously differentiable.*

Note that in **Euc** the tensor product of the objects $(\mathbb{R}^a, \mathcal{B}(\mathbb{R}^a))$ and $(\mathbb{R}^b, \mathcal{B}(\mathbb{R}^b))$ is $(\mathbb{R}^{a+b}, \mathcal{B}(\mathbb{R}^{a+b}))$.

Another important category that we will consider is **Stoch** [18; 21], which has measurable spaces as objects and Markov kernels as morphisms. We define the composition of the Markov kernels $\mu : A \times \Sigma_B \to [0, 1]$ and $\mu' : B \times \Sigma_C \to [0, 1]$ to be the following, where $x_a \in A$ and $\sigma_c \in \Sigma_C$:

$$(\mu' \circ \mu)(x_a, \sigma_c) = \int_{x_b \in B} \mu'(x_b, \sigma_c) d\mu(x_a, \_).$$

The identity morphism at (A, Σ_A) is δ where for $x_a \in A, \sigma_a \in \Sigma_A$:

$$\delta(x_a, \sigma_a) = \begin{cases} 1 & x_a \in \sigma_a \\ 0 & x_a \notin \sigma_a \end{cases}.$$

The tensor product of the Markov Kernels $\mu : A \times \Sigma_B \to [0,1]$ and $\mu' : C \times \Sigma_D \to [0,1]$ in **Stoch** is the Markov Kernel $(\mu' \otimes \mu) : (A \times C) \times (\Sigma_B \otimes \Sigma_D) \to [0,1]$ where $\Sigma_B \otimes \Sigma_D$ is the product sigma-algebra and for $x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c, \sigma_b \in \Sigma_B, \sigma_d \in \Sigma_D$:

$$(\mu' \otimes \mu)((x_a, x_c), \sigma_b \times \sigma_d) = \mu(x_a, \sigma_b)\mu(x_c, \sigma_d).$$

The objects in **Stoch** are also equipped with a commutative comonoidal structure that is compatible with the monoidal product in **Stoch**. Fritz et al. [14] dub categories with this structure **Markov Categories**.

Definition 2.2. *A Markov category is a semicartesian symmetric monoidal category* $(\mathbf{C}, \otimes, 1)$ *in which every object* X *is equipped with a comultiplication map* $cp : X \to X \otimes X$ *and a counit map* $del : X \to 1$ *that satisfy the commutative comonoid equations, naturality of del and:*

$$cp_{X\otimes Y} = (id_X \otimes \sigma_{Y,X} \otimes id_Y)(cp_X \otimes cp_Y),$$

where $\sigma_{Y,X}$ *is the symmetric monoidal swap map in* $\mathbf{C}$.

Stoch naturally arises as the Kleisli category of the Giry Monad, which is an affine symmetric monoidal monad that sends a measurable space to the space of probability measures over that space [18].

Stoch has many notable subcategories based on restrictions of these measurable spaces. For example, the category **FinStoch** consists of finite measurable spaces and Markov Kernels between them. In order to be able to define regular conditional probabilities, Fong [10] and Culbertson et al. [7] restrict to countably generated measurable spaces (**CGStoch**), whereas Fritz et al. [15] restrict to standard Borel spaces (**BorelStoch**), which are the Borel spaces associated with Polish spaces.

2.2.1 Random Variables and Independence in **BorelStoch**

In any categorical presentation of probability, a natural question is how to reason about the notion of independence of random variables [13; 14; 17].

Since **BorelStoch** is the Kleisli category of the restriction of the Giry monad [18] over the **Meas**-subcategory of standard Borel spaces, we can define an embedding functor from this subcategory into **BorelStoch** that acts as an identity on objects and sends the measurable function $f : (A, \Sigma_A) \to (B, \Sigma_B)$ to the Dirac Markov kernel $\delta_f : A \times \Sigma_B \to [0,1]$ where for $x_a \in A, \sigma_b \in \Sigma_B$:

$$\delta_f(x_a, \sigma_b) = \begin{cases} 1 & f(x_a) \in \sigma_b \\ 0 & f(x_a) \notin \sigma_b \end{cases}.$$

This formalizes the intuition that Markov Kernels are a generalization of both measurable functions and probability measures, and provides an avenue to directly study random variables and their independence in **BorelStoch**.

Now suppose we have a probability space (Ω, Σ, μ) such that (Ω, Σ) is standard Borel, and two real-valued random variables defined on this space f, f'. We can think of these random variables as morphisms in **Meas** from (Ω, Σ) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We can represent this probability space as a morphism in **BorelStoch** between 1 and (Ω, Σ): that is, a Markov kernel $\mu : 1 \times \Sigma \to [0,1]$. Going forward we will write the type signature $1 \times \Sigma \to [0,1]$ as $\Sigma \to [0,1]$ for convenience.

We can then represent f and f' with their embeddings into **BorelStoch**: the Dirac Markov kernels $\delta_f, \delta_{f'}$. If we compose δ_f and μ in **BorelStoch**, we form a new probability measure $(\delta_f \circ \mu) : \mathcal{B}(\mathbb{R}) \to [0,1]$, which is the pushforward measure $f_*\mu$ of μ along f.

We now have a hint of how we can reason about the independence or dependence of random variables in **BorelStoch**. First, consider the probability measure $(\delta_f \circ \mu) \otimes (\delta_{f'} \circ \mu) : \mathcal{B}(\mathbb{R} \times \mathbb{R}) \to [0,1]$ where for $\sigma \times \sigma' \in \mathcal{B}(\mathbb{R} \times \mathbb{R})$:

$$\begin{aligned} [(\delta_f \circ \mu) \otimes (\delta_{f'} \circ \mu)]\,(\sigma \times \sigma') = \\ \left[\int_{\omega\in\Omega} \delta_f(\omega, \sigma) d\mu\right] \left[\int_{\omega\in\Omega} \delta_{f'}(\omega, \sigma') d\mu\right] = \\ f_*\mu(\sigma) f'_*\mu(\sigma'). \end{aligned}$$

This is simply the product measure over $(\mathbb{R}\times\mathbb{R}, \mathcal{B}(\mathbb{R}\times\mathbb{R}))$ of the probability measures $(\delta_f \circ \mu)$ and $(\delta_{f'} \circ \mu)$ over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It is completely determined by the marginal distributions of f and f' over the probability space (Ω, Σ, μ), and it is agnostic to the independence or dependence structure of f and f'. The reason for this is that the measure μ is essentially "duplicated", and the random variables f and f' are not actually compared over the same probability space.

In contrast, consider instead the probability measure $(\delta_f \otimes \delta_{f'}) \circ cp \circ \mu : \mathcal{B}(\mathbb{R}\times\mathbb{R}) \to [0,1]$, where $cp : \Omega \to \Omega \otimes \Omega$ is the comonoidal copy map at Ω in **BorelStoch** [14]. We can see that for $\sigma \times \sigma' \in \mathcal{B}(\mathbb{R}\times\mathbb{R})$:

$$\left[(\delta_f \otimes \delta_{f'}) \circ cp \circ \mu\right](\sigma \times \sigma') = \left[\int_{\omega\in\Omega} \delta_f(\omega, \sigma)\delta_{f'}(\omega, \sigma')d\mu\right].$$

This is the probability measure over $(\mathbb{R}\times\mathbb{R}, \mathcal{B}(\mathbb{R}\times\mathbb{R}))$ associated with the joint distribution of the random variables f and f' over (Ω, Σ, μ).

Therefore, the random variables f and f' are independent over the probability space (Ω, Σ, μ) if and only if the probability measures $(\delta_f \circ \mu) \otimes (\delta_{f'} \circ \mu)$ and $(\delta_f \otimes \delta'_f) \circ cp \circ \mu$ are equal.

3 The co-Kleisli Construction

Fong et al. [12] and Gavranović [16] build their characterization of machine learning optimization problems on top of the category **Para**(**Euc**) of Euclidean spaces and parameterized differentiable maps between them. Rather than represent the loss function itself categorically, the authors treat it as an externally-provided hyperparameter.

However, in practice the loss function is usually implied by the problem. A common problem statement is as follows: given some parameterized random variable, derive the parameters that maximize the likelihood of some observed data being drawn from the distribution of this random variable. A natural question is therefore whether it is possible to replace the parameterized differentiable maps in **Para**(**Euc**) with parameterized random variables.

Before moving to **Para**(**Euc**), we will start with the category **Euc** of Euclidean spaces and differentiable maps between them. Our first step will be to replace the morphisms in **Euc** with stochastic processes, or indexed families of random variables. We start with the following definition:

Definition 3.1. *For some Cartesian monoidal category* $\mathbf{C}$ *and object* A *in* $\mathbf{C}$, $\mathbf{CoKl}_A(\mathbf{C})$ *is the co-Kleisli category of* $\mathbf{C}$ *under the co-monad* $(A \otimes \_)$.

For example, if Ω is $\mathbb{R}^n$ for some $n \in \mathbb{N}$, the category $\mathbf{CoKl}_{(\Omega,\mathcal{B}(\Omega))}(\mathbf{Euc})$ (which we will hereafter abbreviate **CEuc**, see Table 3.1) has the same objects as **Euc**, and the morphisms between $\mathbb{R}^a$ and $\mathbb{R}^b$ are continuously differentiable (and therefore Borel-measurable) functions of the form $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$. In **CEuc**, the composition of $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega \times \mathbb{R}^b \to \mathbb{R}^c$ is $(f' \circ f) : \Omega \times \mathbb{R}^a \to \mathbb{R}^c$ where for $\omega \in \Omega, x_a \in \mathbb{R}^a$:

$$(f' \circ f)(\omega, x_a) = f'(\omega, f(\omega, x_a)).$$

And the tensor of $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega \times \mathbb{R}^c \to \mathbb{R}^d$ is $(f' \otimes f) : \Omega \times \mathbb{R}^a \times \mathbb{R}^c \to \mathbb{R}^b \times \mathbb{R}^d$ where for $\omega \in \Omega, x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$:

$$(f' \otimes f)(\omega, (x_a, x_c)) = (f(\omega, x_a), f'(\omega, x_c)).$$

One important thing to note is that ω is reused when we compose or tensor f and f'. This allows us to make the following claim:

Proposition 1. *For any* $\omega \in \Omega$, *the identity-on-objects map that sends the function* $f : \Omega\times\mathbb{R}^a \to \mathbb{R}^b$ *in* **CEuc** *to the function* $f(\omega, \_) : \mathbb{R}^a \to \mathbb{R}^b$ *in* **Euc** *is a strict monoidal functor* $R_\omega : \mathbf{CEuc} \to \mathbf{Euc}$, *which we call the* ***realization functor****.*

Proof. First, if f is the identity map in **CEuc** then $f(\omega, \_)$ is by definition the identity function. Next, consider $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b, f' : \Omega \times \mathbb{R}^b \to \mathbb{R}^c$ in **CEuc** and any $x_a \in \mathbb{R}^a$. Then:

$$(R_\omega f' \circ R_\omega f)(x_a) = (f'(\omega, \_) \circ f(\omega, \_))(x_a) = f'(\omega, f(\omega, x_a)) = R_\omega(f' \circ f)(x_a)$$

so composition is preserved. Finally, consider $g : \Omega \times \mathbb{R}^a \to \mathbb{R}^b, g' : \Omega \times \mathbb{R}^c \to \mathbb{R}^d$ in **CEuc** and any $x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$. Then:

$$(R_\omega g \otimes R_\omega g')(x_a, x_c) = (g(\omega, x_a), g'(\omega, x_c)) = R_\omega(g \otimes g')(x_a, x_c)$$

so the monoidal tensor is preserved. □

Given a probability measure $\mu : \mathcal{B}(\Omega) \to [0, 1]$, we can think of **CEuc** as a category of differentiable stochastic processes defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$. One particularly important kind of stochastic process is a **Levy Process**. We can view Levy Processes as continuous-time generalizations of random walks, or as Brownian motions with drift. Formally, a Levy Process is a one-dimensional stochastic process $f : \Omega \times \mathbb{R} \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ such that:

- $f(\_, 0) = 0$ almost surely.
- For $t_d > t_c > t_b > t_a \in \mathbb{R}$, the random variables $f(\_, t_b) - f(\_, t_a)$ and $f(\_, t_d) - f(\_, t_c)$ are independent.
- For $t_b > t_a \in \mathbb{R}$, the random variables $f(\_, t_b) - f(\_, t_a)$ and $f(\_, t_b - t_a)$ have the same distribution.
- For any $\omega \in \Omega$ the function $f(\omega, \_)$ is continuous.

A **subordinator** is a non-decreasing Levy Process. That is, for any fixed $\omega \in \Omega$ the function $f(\omega, \_)$ is non-decreasing.

Proposition 2. *Continuously differentiable subordinators form a single-object subcategory of* **CEuc** *at* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof. First, note that the identity arrow on $\mathbb{R}$ is trivially a subordinator. Next, suppose f and g are subordinators. By Lalley [20] we have that $g \circ f$ is a Levy Process. Since both f and g are non-decreasing, for $t_2 > t_1$ we have for any fixed $\omega \in \Omega$ that:

$$g(\omega, f(\omega, t_2)) > g(\omega, f(\omega, t_1)).$$

Therefore, $g \circ f$ is a subordinator as well. □

3.1 Independence and Dependence in CEuc

Since all of the stochastic processes in **CEuc** are defined over the same probability space $(\Omega, \mathcal{B}(\Omega), \mu)$, there is a major difference between how **CEuc** and **BorelStoch** represent independence and dependence. Given the arrows $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega \times \mathbb{R}^c \to \mathbb{R}^d$ in **CEuc** and the vectors $x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$, the random variables $f(\_, x_a)$ and $f'(\_, x_c)$ may be either dependent or independent.

In order to see how this differs from the situation in **BorelStoch**, recall that the pushforward of μ along the stochastic process $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$ is a mapping from **CEuc** to **BorelStoch** such that for $x_a \in \mathbb{R}^a, \sigma_b \in \mathcal{B}(\mathbb{R}^b)$:

$$f_*\mu(x_a, \sigma_b) = f(\_, x_a)_*\mu(\sigma_b) = \mu(f(\_, x_a)^{-1}(\sigma_b)) = \int_{\omega \in \Omega} \delta(f(\omega, x_a), \sigma_b) d\mu.$$

However, this mapping does not form a functor. We see that for $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$, $f' : \Omega \times \mathbb{R}^b \to \mathbb{R}^c$, $x_a \in \mathbb{R}^a, \sigma_c \in \mathcal{B}(\mathbb{R}^c)$:

$$(f' \circ f)_* \mu(x_a, \sigma_c) = \int_{x_b \in \mathbb{R}^b} \int_{\omega \in \Omega} \delta(f'(\omega, x_b), \sigma_c) \, d\delta(f(\omega, x_a), \_) \, d\mu.$$

whereas:

$$[f'_*\mu \circ f_*\mu](x_a, \sigma_c) = \int_{x_b \in \mathbb{R}^b} \left(\int_{\omega \in \Omega} \delta(f'(\omega, x_b), \sigma_c) d\mu \right) \left(\int_{\omega \in \Omega} d\delta(f(\omega, x_a), \_) d\mu \right).$$

These are not necessarily equivalent if the random variables $f'(\_, x_b), x_b \in \mathbb{R}^b$ are not independent of the random variable $f(\_, x_a)$.

The reason for this mismatch comes down to the fact that tensor and composition in **BorelStoch** are based on the Markov property. We can slightly modify **CEuc** to define a new category of stochastic processes that exhibit this independence behavior.

Shorthand Name	Full Name
CEuc	$\mathbf{CoKl}_{(\Omega,\mathcal{B}(\Omega))}(\mathbf{Euc})$
PEuc	$\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{Euc})$
DF	$\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{Para}_{\mathbf{Euc}}(\mathbf{Euc}))$

Table 1: This paper introduces several compositional constructions for building new categories. These can produce unwieldy names, so for readability we have abbreviated some of them here.

4 The Parameterization Construction

In order to reason about the behavior of a system of stochastic processes, it is useful to study them in a simpler setting. There are two simple ways to do this: take pushforwards and study stochastic processes as Markov Kernels, or take expectations and study stochastic processes as functions. In order to make these lines of study rigorous, we first need to establish the functoriality of these transformations. To this end, in this section we build a new category of stochastic processes such that the map $f \to f_*\mu$ described in Section 3.1 is functorial. In Sections 5.2 and 6 we will explore the functoriality of the expectation.

In order to elevate the pushforward to a functor, we need to modify the definition of how stochastic processes compose. Unlike in **CEuc**, where we treat all stochastic processes as if they were defined over the same probability space, the category in this section will consist of stochastic processes defined over different, non-interacting probability spaces. The composition or tensor of two stochastic processes in this new category will produce a stochastic process over the product of those processes' associated probability spaces. This will allow us to treat all of the stochastic processes in this category as if they were mutually independent.

We note that this strategy of expanding the probability space each time we introduce a new source of randomness is commonly used by probability theorists [1; 2; 24].

4.1 An extension of **Para**

We will begin by slightly modifying Gavranović's [16] **Para** construction, which is itself a generalization of **Para** from Fong et al. [12].

Consider the small symmetric strict monoidal categories $\mathbf{C}$ and $\mathbf{D}$ such that there exists a faithful identity-on-objects monoidal functor $\iota : \mathbf{D} \hookrightarrow \mathbf{C}$. That is, we can think of $\mathbf{D}$ as a subcategory of $\mathbf{C}$. Then write $(\_ \otimes A) \circ \iota : \mathbf{D} \hookrightarrow \mathbf{C}$ to denote the functor that sends the object A' in $\mathbf{D}$ to $A' \otimes A$ in $\mathbf{C}$ and write $c_B : \mathbf{D} \to \mathbf{C}$ for the constant functor that sends all objects in $\mathbf{D}$ to B.

Definition 4.1. *For the small symmetric strict monoidal categories $\mathbf{C}$ and $\mathbf{D}$ equipped with a faithful identity-on-objects monoidal functor $\iota : \mathbf{D} \hookrightarrow \mathbf{C}$, the category $\mathbf{Para}_{\mathbf{D}}(\mathbf{C})$ has the same objects as $\mathbf{C}$ with homset $\mathbf{Para}_{\mathbf{D}}(\mathbf{C})[A, B]$ equal to the set of objects in the comma category $(\_ \otimes A) \circ \iota \downarrow c_B$. That is, the morphisms between A and B in $\mathbf{Para}_{\mathbf{D}}(\mathbf{C})$ are morphisms of the form $P \otimes A \to B$ in $\mathbf{C}$, where P is an object in $\mathbf{D}$. The composition of the arrows:*

$$f : P \otimes A \to B \in \mathbf{Para}_{\mathbf{D}}(\mathbf{C})[A, B] \qquad f' : Q \otimes B \to C \in \mathbf{Para}_{\mathbf{D}}(\mathbf{C})[B, C]$$

in $\mathbf{Para}_{\mathbf{D}}(\mathbf{C})$ is then as follows, where we write $\circ_{\mathbf{C}}$ and $\otimes_{\mathbf{C}}$ for the composition and tensor of arrows in $\mathbf{C}$ respectively:

$$f' \circ f : Q \otimes P \otimes A \to C \qquad f' \circ f = f' \circ_{\mathbf{C}} (id_Q \otimes_{\mathbf{C}} f).$$

And the tensor of arrows $g : P \otimes A \to B$ *and* $g' : Q \otimes C \to D$ *in* $\mathbf{Para_D(C)}$ *is:*

$$g \otimes g' : P \otimes Q \otimes A \otimes C \to B \otimes D$$
$$g \otimes g' = (g \otimes_{\mathbf{C}} g') \circ_{\mathbf{C}} (id_P \otimes_{\mathbf{C}} \sigma_{(Q,A)} \otimes_{\mathbf{C}} id_C)$$

where $\sigma_{(Q,A)} : Q \otimes A \to A \otimes Q$ *is the symmetric monoidal swap map in* $\mathbf{C}$*. The monoidal unit* 1 *is the same in* $\mathbf{Para_D(C)}$ *as in* $\mathbf{C}$ *and the identity arrow at* A *in* $\mathbf{Para_D(C)}$ *is* $- \otimes_{\mathbf{C}} id_A : 1 \otimes A \to A$*, where* $id_A : A \to A$ *is the identity arrow at* A *in* $\mathbf{C}$*.*

Note that unlike Gavranović [16], we require $\mathbf{C}$ to be strict monoidal in order to ensure that composition is associative without resorting to equivalence classes.

Proposition 3. *Suppose* $\mathbf{C}$ *and* $\mathbf{C}'$ *are small symmetric strict monoidal categories with a strict monoidal functor* $F : \mathbf{C} \to \mathbf{C}'$ *between them. Suppose* $\mathbf{D}$ *is a small symmetric strict monoidal category equipped with a faithful identity-on-objects strict monoidal functor* $\iota : \mathbf{D} \hookrightarrow \mathbf{C}$ *and that the image of* $F \circ \iota$ *is a subcategory* $\mathbf{D}'$ *of* $\mathbf{C}'$*. Then the map* $F_p : \mathbf{Para_D(C)} \to \mathbf{Para_{D'}(C')}$ *that applies the same actions on objects and arrows as* F *is a strict monoidal functor.*

Proof. We will first show that F_p is a functor, and then we will show that it is strict monoidal. Like above, we write $\circ_{\mathbf{C}}$, $\otimes_{\mathbf{C}}$, and $\sigma_{(Q,A)}$ for the composition, tensor, and symmetric monoidal swap of arrows in $\mathbf{C}$.

First note that since $F_p : \mathbf{Para_D(C)} \to \mathbf{Para_{D'}(C')}$ applies the same actions on objects and arrows as $F : \mathbf{C} \to \mathbf{C}'$, it trivially preserves identity morphisms. Next, we will show that F_p preserves composition. Suppose $f : P \otimes A \to B, g : Q \otimes B \to C$ are arrows in $\mathbf{Para_D(C)}$. Then we have that:

$$F_p(g \circ f) = F(g \circ_{\mathbf{C}} (id_Q \otimes_{\mathbf{C}} f)) = Fg \circ_{\mathbf{C}'} (id_Q \otimes_{\mathbf{C}'} Ff) = F_p g \circ F_p f.$$

Next, we will show that F_p is strict monoidal. We first note that F_p trivially preserves the monoidal unit, since the monoidal unit is the same in $\mathbf{C}$ and $\mathbf{Para_D(C)}$. Next, suppose $f : P \otimes A \to B$ and $g : Q \otimes C \to D$ are arrows in $\mathbf{Para_D(C)}$. Then we have that:

$$\begin{aligned} F_p(f \otimes g) = \\ F((f \otimes_{\mathbf{C}} g) \circ_{\mathbf{C}} (id_P \otimes_{\mathbf{C}} \sigma_{(Q,A)} \otimes_{\mathbf{C}} id_C)) = \\ (Ff \otimes_{\mathbf{C}'} Fg) \circ_{\mathbf{C}'} (id_P \otimes_{\mathbf{C}'} \sigma_{(Q,A)} \otimes_{\mathbf{C}'} id_C) = \\ F_p f \otimes F_p g. \end{aligned}$$

□

4.2 A Category of Parametric Measurable Maps

In this Section, we will use the **Para** construction to build a new category of stochastic processes over which the mapping $f \to f_* \mu$ is functorial. In this category the tensor and composition will have the same independence structure that they have in **Stoch**.

4.2.1 Lawvere Parameterization

We begin with the following definition:

Definition 4.2. *Suppose* $\mathbf{C}$ *is a strict Cartesian monoidal category,* O^* *is a Lawvere theory with generating object* O*, and* ι *is a faithful identity-on-objects functor* $\iota : O^* \hookrightarrow \mathbf{C}$*. Then* $\mathbf{Para}_{O^*}(\mathbf{C})$ *is a* ***Lawvere parameterization*** *of* $\mathbf{C}$*.*

Note that the objects in O^* are of the form $O \otimes O \otimes \cdots \otimes O$. When the tensor is repeated n times we will write this as O^n. For any strict Cartesian monoidal category $\mathbf{C}$ with a Lawvere parameterization we can define a mapping $Copy : \mathbf{Para}_{O^*}(\mathbf{C}) \to \mathbf{CoKl}_O(\mathbf{C})$. This mapping acts as identity-on-objects and sends the arrow $f : O^n \times A \to B$ in $\mathbf{Para}_{O^*}(\mathbf{C})$ to the following arrow in $\mathbf{CoKl}_O(\mathbf{C})$:

$$f \circ_{\mathbf{C}} (\Delta_O^{n-1} \otimes_{\mathbf{C}} id_A^{\mathbf{C}}) : O \times A \to B.$$

For clarity, $id^{\mathbf{C}}_A$ is the identity arrow on A in $\mathbf{C}$, $\Delta_O : O \to O \otimes O$ is the copy (aka diagonal) map in $\mathbf{C}$, $\Delta^{n-1}_O : O \to O \otimes O \otimes \cdots \otimes O$ is the repeated application of this map $n-1$ times and Δ^0_O is the identity on O.

Proposition 4. *Copy is a full identity-on-objects strict monoidal functor.*

Proof. First, we note that *Copy* is identity-on-objects by definition.

Next, consider any objects A, B in $\mathbf{C}$ and any arrow $f : O \times A \to B$ in the co-Kleisli category of $\mathbf{C}$ under $(O \times \_)$. Then f is also an arrow in $\mathbf{Para}_{O^*}(\mathbf{C})$ and *Copy* maps f to f. Therefore *Copy* is full.

Next, since the id_A arrow in $\mathbf{Para}_{O^*}(\mathbf{C})$ is of the form $1 \otimes A \to A$, *Copy* maps it to the arrow $id_A \circ_{\mathbf{C}} (\Delta^0_O \otimes id^{\mathbf{C}}_A) = id_A$. Therefore, *Copy* preserves identity morphisms.

Next, we will show *Copy* preserves composition. Suppose $f : O^m \times A \to B$ and $f' : O^n \times B \to C$ are arrows in $\mathbf{Para}_{O^*}(\mathbf{C})$:

$$\begin{aligned}(Copy f' \circ Copy f) = \\ \left(f' \circ_{\mathbf{C}} (\Delta^{n-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_B)\right) \circ \left(f \circ_{\mathbf{C}} (\Delta^{m-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_A)\right) = \\ (f' \circ f) \circ_{\mathbf{C}} (\Delta^{n+m-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_A) = \\ Copy(f' \circ f).\end{aligned}$$

Finally, we will show that *Copy* preserves tensor. Suppose $f : O^m \times A \to B$ and $f' : O^n \times C \to D$ are arrows in $\mathbf{Para}_{O^*}(\mathbf{C})$:

$$\begin{aligned}(Copy f' \otimes Copy f) = \\ \left[f' \circ_{\mathbf{C}} (\Delta^{n-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_C)\right] \otimes \left[f \circ_{\mathbf{C}} (\Delta^{m-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_A)\right] = \\ (f' \otimes f) \circ_{\mathbf{C}} (\Delta^{n+m-1}_O \otimes_{\mathbf{C}} id^{\mathbf{C}}_{C \otimes A}) = \\ Copy(f' \otimes f).\end{aligned}$$

□

4.2.2 Applying **Para** to **Euc**

Now suppose we have a probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where Ω is $\mathbb{R}^k, k \in \mathbb{N}$. We can form the Lawvere theory $(\Omega, \mathcal{B}(\Omega))^*$ with generating object $(\Omega, \mathcal{B}(\Omega))$ and tuples $(\Omega, \mathcal{B}(\Omega))^n = (\Omega^n, \mathcal{B}(\Omega^n))$ as objects. We can also form the faithful identity-on-objects functor $\iota : (\Omega, \mathcal{B}(\Omega))^* \hookrightarrow \mathbf{Euc}$. Then for any $(\Omega^n, \mathcal{B}(\Omega^n)) \in (\Omega, \mathcal{B}(\Omega))^*$, we can create the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$ where μ^n is the product measure:

$$\mu^n(\sigma_1 \times \sigma_2 \times \cdots \times \sigma_n) = \mu(\sigma_1)\mu(\sigma_2) \cdots \mu(\sigma_n).$$

Now consider the Lawvere parameterization $\mathbf{Para}_{(\Omega, \mathcal{B}(\Omega))^*}(\mathbf{Euc})$ (which we will hereafter abbreviate **PEuc**). Intuitively, **PEuc** allows us to reason about probabilistic relationships in terms of measurable functions rather than probability measures. We can make this probabilistic intuition more formal. First, **PEuc** behaves similarly to a category of Markov Kernels and we can show the following:

Proposition 5. *We can construct a Markov Category [14] on top of* **PEuc** *by equipping each object with the comultiplication map cp and and the counit map dc defined as follows:*

$$\begin{aligned} cp : 1 \times \mathbb{R}^a \to \mathbb{R}^a \times \mathbb{R}^a \qquad & dc : 1 \times \mathbb{R}^a \to 1 \\ cp(-, x_a) = (x_a, x_a) \qquad & dc(-, x_a) = - \end{aligned}$$

Proof in Appendix

Next, by Proposition 4, we have an identity-on-objects functor, *Copy*, from **PEuc** to **CEuc**. Let's drill deeper into this relationship. We can view an arrow of the form $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$ in

PEuc as a stochastic process over $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. However, unlike in **CEuc**, if we compose or tensor f with another arrow in **PEuc**, we do not get another stochastic process over $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. Instead, we get a stochastic process over some other probability space. Intuitively, we can think of the stochastic processes in **PEuc** as being defined over different, non-interacting probability spaces.

Now given some arrow $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}$ in **PEuc** and $x_a \in \mathbb{R}^a$, the measurable function $f(\_, x_a)$ is a real-valued random variable over the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. The pushforward of μ^n along this random variable $f(\_, x_a)_* \mu^n(\_)$ is then a probability measure over the space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

In general, we can extend this pushforward procedure to define a mapping between parametric families of measurable maps and Markov Kernels. Given some $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$ we can define $Push_\mu f : \mathbb{R}^a \times \mathcal{B}(\mathbb{R}^b) \to [0, 1]$ where for $x_a \in \mathbb{R}^a, \sigma_b \in \mathcal{B}(\mathbb{R}^b)$:

$$Push_\mu f(x_a, \sigma_b) = f(\_, x_a)_* \mu^n(\sigma_b) = \int_{\omega_n \in \Omega^n} \delta(f(\omega_n, x_a), \sigma_b) d\mu^n.$$

Proposition 6. *The mapping $Push_\mu$ that takes a parametric family $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$ of measurable maps to the Markov Kernel $f_* \mu^n$ is an identity-on-objects strict monoidal functor from* **PEuc** *to* **BorelStoch**.

Proof. We first note that for any $\mathbb{R}^a$, $Push_\mu$ trivially maps the identity at $\mathbb{R}^a$ in **PEuc** to its identity in **BorelStoch**. Next, we will demonstrate that $Push_\mu$ preserves composition. Suppose we have some $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$, $f' : \Omega^m \times \mathbb{R}^b \to \mathbb{R}^c$, $x_a \in \mathbb{R}^a, \sigma_c \in \mathcal{B}(\mathbb{R}^c)$:

$$\begin{aligned}
Push_\mu\left(f' \circ f\right)(x_a, \sigma_c) = \\
\int_{(\omega_m, \omega_n) \in \Omega^m \times \Omega^n} \delta((f' \circ f)((\omega_m, \omega_n), x_a), \sigma_c) d\mu^{n+m} = \\
\int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \delta((f'(\omega_m, f(\omega_n, x_a)), \sigma_c) d\mu^n d\mu^m = \\
\int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \int_{x_b \in \mathbb{R}^b} \delta(f'(\omega_m, x_b), \sigma_c) d\delta(f(\omega_n, x_a), _) d\mu^n d\mu^m = \\
\int_{x_b \in \mathbb{R}^b} \int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \delta(f'(\omega_m, x_b), \sigma_c) d\delta(f(\omega_n, x_a), _) d\mu^n d\mu^m = \\
\int_{x_b \in \mathbb{R}^b} \left[\int_{\omega_m \in \Omega^m} \delta(f'(\omega_m, x_b), \sigma_c) d\mu^n\right] \left[\int_{\omega_n \in \Omega^n} d\delta(f(\omega_n, x_a), _) d\mu^m\right] = \\
\int_{x_b \in \mathbb{R}^b} [Push_\mu f'](x_b, \sigma_c)\ d[Push_\mu f](x_a, _) = \\
(Push_\mu f' \circ Push_\mu f)(x_a, \sigma_c).
\end{aligned}$$

Finally, we will demonstrate that $Push_\mu$ preserves tensor. Suppose we have some $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$, $f' : \Omega^m \times \mathbb{R}^c \to \mathbb{R}^d$, $x_c \in \mathbb{R}^c$, $x_a \in \mathbb{R}^a$, $\sigma_d \in \mathcal{B}(\mathbb{R}^d)$, and $\sigma_b \in \mathcal{B}(\mathbb{R}^b)$:

$$\begin{aligned}
Push_\mu\left(f' \otimes f\right)((x_c, x_a), (\sigma_d \times \sigma_b)) = \\
\int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \delta((f' \otimes f)((\omega_m, \omega_n), (x_c, x_a)), (\sigma_d \times \sigma_b)) d\mu^m d\mu^n = \\
\int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \delta(f'(\omega_m, x_c), \sigma_d) \delta(f(\omega_n, x_a), \sigma_b) d\mu^m d\mu^n = \\
\int_{\omega_m \in \Omega^m} \delta(f'(\omega_m, x_c), \sigma_d) d\mu^m \int_{\omega_n \in \Omega^n} \delta(f(\omega_n, x_a)), \sigma_b) d\mu^n = \\
(Push_\mu f')(x_c, \sigma_d)(Push_\mu f)(x_a, \sigma_b) = \\
(Push_\mu f' \otimes Push_\mu f)((x_a, x_c), (\sigma_d \otimes \sigma_b)).
\end{aligned}$$

□

5 Parameterized Statistical Models

We have been discussing the arrows in **PEuc** as parameterized random variables, or stochastic processes, but we can also think of them as **Euc** arrows with an element of randomness that is dictated by the probability measure μ. One of the primary goals of this work is to replace the domain of Fong et al.'s [12] Backpropagation functor, **Para**(**Euc**), with a probabilistically motivated category over which we can define the error function $er : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ through maximum likelihood. Therefore, a natural next step is to extend **PEuc** to a category in which we can instead think of the arrows as **Para**(**Euc**) arrows with an element of randomness added.

In order to do this, we will replace the stochastic processes in **PEuc** with parameterized stochastic processes, which we will also refer to as parametric statistical models. That is, the arrows in this category will consist of families of random variables that have two layers of parameterization: one layer acts as the model input (e.g. the independent variable in a linear regression model) and one layer acts as the model parameters (e.g. the slope, intercept and variance terms).

5.1 The Category **DF**

Given a probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where $\Omega = \mathbb{R}^k, k \in \mathbb{N}$, any stochastic process $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$ in **PEuc** defines a stochastic relationship between values in $\mathbb{R}^a$ and $\mathbb{R}^b$. A parametric statistical model is a parameterized family of such relationships. For example, consider a univariate linear regression model $l : \Omega^n \times \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ where for $\omega_n \in \Omega^n, [a, b, s] \in \mathbb{R}^3, x \in \mathbb{R}$:

$$l(\omega_n, [a, b, s], x) = ax + b + f_{\mathcal{N}(0,s^2)}(\omega_n)$$

and $f_{\mathcal{N}(0,s^2)}$ is a normally distributed random variable with variance s^2. Any value $[a, b, s] \in \mathbb{R}^3$ defines the stochastic process, or **PEuc** arrow:

$$l(\_, [a, b, s], \_) : \Omega^n \times \mathbb{R} \to \mathbb{R}.$$

For any model input value $x \in \mathbb{R}$, the function $l(\_, [a, b, s], x)$ is then a random variable defined on the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. Like with any ordinary univariate linear regression model, this random variable is normally distributed on the real line.

We can define a category of such models by applying $\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}$ to $\mathbf{Para}_{\mathbf{Euc}}(\mathbf{Euc})$ to form the category $\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{Para}_{\mathbf{Euc}}(\mathbf{Euc}))$, which we will rename **DF** for brevity (see Table 1 for a list of all such abbreviations). This naming derives from the fact that the arrows in this category are **D**iscriminative and **F**requentist statistical models. That is, each arrow operates as if both the parameters and input values are fixed and only the output value is probabilistic. For example, the homset $\mathbf{DF}[\mathbb{R}, \mathbb{R}]$ includes the linear regression model above. In contrast, generative models and Bayesian models assume a probability distribution over the input and parameter values respectively.

5.2 A subcategory of Gaussian-preserving transformations

Definition 5.1. *A **Gaussian-preserving transformation** $T : \mathbb{R}^a \to \mathbb{R}^b$ is a function such that for any multivariate normal random variable $f : \Omega \to \mathbb{R}^a$ defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$, the random variable $T(f(\_)) : \Omega \to \mathbb{R}^b$ is multivariate normal and we have:*

$$\int_{\omega \in \Omega} T(f(\omega))d\mu = T\left(\int_{\omega \in \Omega} f(\omega)d\mu\right).$$

For example, any linear function is Gaussian-preserving.

Now for some probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where $\Omega = \mathbb{R}^k, k \in \mathbb{N}$, we can construct a set of **DF**-arrows $\mathcal{N}_\mu$ such that for any $f \in \mathcal{N}_\mu$ with the signature $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $\omega_n \in \Omega^n, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:

$$f(\omega_n, x_p, x_a) = T(x_p, x_a) + G(\omega_n)$$

where $T(x_p, \_) : \mathbb{R}^a \to \mathbb{R}^b$ is a Gaussian-preserving transformation and $G : \Omega^n \to \mathbb{R}^b$ is a multivariate normal random variable defined on the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. Note that this

includes the univariate linear regression model l, as well as the identity arrow, since constant distributions are multivariate normal with variance 0.

Note that $\mathcal{N}_\mu$ is closed under the tensor in **DF**, since given the maps $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^c \to \mathbb{R}^d$, $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in $\mathcal{N}_\mu$ and $\omega_m \in \Omega^m, \omega_n \in \Omega^n, x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_c \in \mathbb{R}^c, x_a \in \mathbb{R}^a$:

$$\begin{aligned}(f' \otimes f)((\omega_m, \omega_n), (x_q, x_p), (x_c, x_a))) = \\ (T'(x_q, x_c) + G'(\omega_m), T(x_p, x_a) + G(\omega_n)) = \\ (T'(x_q, x_c), T(x_p, x_a)) + (G'(\omega_m), G(\omega_n)).\end{aligned}$$

Next, we will define $\mathbf{DF}_{\mathcal{N}_\mu}$ to be the category with the same objects as **DF** and arrows generated by the composition of arrows in $\mathcal{N}_\mu$.

Proposition 7. $\mathbf{DF}_{\mathcal{N}_\mu}$ *is a strict symmetric monoidal subcategory of* **DF**.

Proof. Since $\mathbf{DF}_{\mathcal{N}_\mu}$ contains the identities and is closed under composition by definition, we only need to demonstrate that $\mathbf{DF}_{\mathcal{N}_\mu}$ is closed under the monoidal product on arrows. We will demonstrate that for any f, g in $Ar(\mathbf{DF}_{\mathcal{N}_\mu})$ we can write $g \otimes f$ as a composition of arrows in $\mathcal{N}_\mu$. First note that:

$$f = (f_n \circ ... \circ f_1) \qquad g = (g_m \circ ... \circ g_1)$$

where for all $i \leq n, j \leq m$, f_i and g_j are arrows in $\mathcal{N}_\mu$. Without loss of generality, we will assume that $n \leq m$, which implies that:

$$f = (id_m \circ id_{m-1} \circ ... \circ id_{n+1} \circ f_n \circ ... \circ f_1).$$

We can now write the following:

$$g \otimes f = (g_m \otimes id_m) \circ (g_{m-1} \otimes id_{m-1}) \circ ... \circ (g_{n+1} \otimes id_{n+1}) \circ (g_n \otimes f_n) \circ ... \circ (g_1 \otimes f_1).$$

Since this is a composition of arrows in $\mathcal{N}_\mu$, $g \otimes f$ is in $\mathcal{N}_\mu$. □

Proposition 8. *Given any arrow* $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ *in* $\mathbf{DF}_{\mathcal{N}_\mu}$ *and* $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$, $f(\_, x_p, x_a)$ *is a multivariate normal random variable defined on the probability space* $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$.

Proof. We will show that this property holds for the arrows in $\mathcal{N}_\mu$ and that it is preserved by composition.

To begin, note that for any n, m, the pushforward of μ^m along $f : \Omega^m \to \mathbb{R}^a$ is equivalent to the pushforward of μ^{m+n} along the random variable $f^l(\omega_m, \omega_n) = f(\omega_m)$ where $\omega_m \in \Omega^m, \omega_n \in \Omega^n$. For $\sigma_a \in \mathcal{B}(\mathbb{R}^a)$:

$$\begin{aligned}f^l_* \mu^{m+n}(\sigma_a) = \\ \int_{(\omega_m, \omega_n) \in \Omega^{m+n}} \delta(f^l(\omega_m, \omega_n), \sigma_a) d\mu^{m+n} = \\ \int_{\omega_m \in \Omega^m} \int_{\omega_n \in \Omega^n} \delta(f^l(\omega_m, \omega_n), \sigma_a) d\mu^m d\mu^n = \\ \int_{\omega_m \in \Omega^m} \delta(f(\omega_m), \sigma_a) d\mu^m \int_{\omega_n \in \Omega^n} d\mu^n = \\ f_* \mu^m(\sigma_a).\end{aligned}$$

By a similar argument we have that the pushforward of μ^m along $f : \Omega^m \to \mathbb{R}^a$ is equivalent to the pushforward of μ^{n+m} along the random variable $f^r(\omega_n, \omega_m) = f(\omega_m)$.

Next, we note that for any $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ and arrow $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b \in \mathcal{N}_\mu$, the random variable $f(\_, x_p, x_a) : \Omega^n \to \mathbb{R}^b$ is multivariate normal and defined on the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. This follows from the fact that for $\omega_n \in \Omega^n$:

$$f(\omega_n, x_p, x_a) = T(x_p, x_a) + G(\omega_n)$$

where $T(x_p, x_a)$ is a constant and $G : \Omega^n \to \mathbb{R}^b$ is multivariate normal. Next, we show that for any $x_p \in \mathbb{R}^p, x_q \in \mathbb{R}^q, x_a \in \mathbb{R}^a$, arrow $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ in $\mathcal{N}_\mu$ and arrow $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in **DF** such that the random variable $f(\_, x_p, x_a) : \Omega^n \to \mathbb{R}^b$ is multivariate normal, the random variable:

$$(f' \circ f)(\_, (x_q, x_p), x_a) : \Omega^{m+n} \to \mathbb{R}^b$$

is multivariate normal over $(\Omega^{m+n}, \mathcal{B}(\Omega^{m+n}), \mu^{m+n})$ since:

$$\begin{aligned}(f' \circ f)((\omega_m, \omega_n), (x_q, x_p), x_a) = \\ f'(\omega_m, x_q, f(\omega_n, x_p, x_a)) = \\ T'(x_q, f(\omega_n, x_p, x_a)) + G'(\omega_m).\end{aligned}$$

Since the random variable $f(\_, x_p, x_a) : \Omega^n \to \mathbb{R}^b$ is multivariate normal over $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$, by the note above we have that the random variable $f^r((\omega_m, \omega_n), x_p, x_a) = f(\omega_n, x_p, x_a)$ defined over $(\Omega^{m+n}, \mathcal{B}(\Omega^{m+n}), \mu^{m+n})$ is multivariate normal. Since x_q is constant this implies that the following random variable is also multivariate normal:

$$T'(x_q, f^r(\_, x_p, x_a)) : \Omega^{m+n} \to \mathbb{R}^c.$$

Similarly, the random variable $G''^l(\omega_m, \omega_n) = G'(\omega_m)$ is also multivariate normal and independent of $T(x_q, f^r(\_, x_p, x_a))$. Therefore, we can write:

$$\begin{aligned}(f' \circ f)((\omega_m, \omega_n), (x_q, x_p), x_a) = \\ T'(x_q, f(\omega_n, x_p, x_a)) + G'(\omega_m) = \\ T'(x_q, f^r((\omega_m, \omega_n), x_p, x_a)) + G''^l(\omega_m, \omega_n).\end{aligned}$$

Since this is a sum of independent normally distributed random variables, the following random variable is also multivariate normal:

$$(f' \circ f)(\_, (x_q, x_p), x_a) : \Omega^{m+n} \to \mathbb{R}^b.$$

□

As an aside, note that $\mathcal{N}_\mu$ itself is not closed under composition. Suppose $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ and $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ are in $\mathcal{N}_\mu$ and that $f'(\omega_m, x_q, x_b) = T'(x_q, x_b) + G'(\omega_m)$ where $T'(x_q, x_b) = \|x_q\|_1 x_b$. Note that T' is Gaussian preserving since the product of a constant and a Gaussian is Gaussian. Now if we write $f(\omega_n, x_p, x_a) = T(x_p, x_a) + G(\omega_n)$ we see that:

$$(f' \circ f)((\omega_m, \omega_n), (x_q, x_p), x_a) = \|x_q\|_1 T(x_p, x_a) + \|x_q\|_1 G(\omega_n) + G'(\omega_m),$$

which we cannot express as a sum of a Gaussian-preserving transformation over $\mathbb{R}^{q+p} \times \mathbb{R}^a \to \mathbb{R}^b$ and a multivariable normal random variable defined on $(\Omega^{n+m}, \mathcal{B}(\Omega^{n+m}), \mu^{n+m})$.

5.2.1 Relationship to Gauss

$\mathbf{DF}_{\mathcal{N}_\mu}$ is similar to the category **Gauss** from Section 6 of Fritz et al. [14], with a few key differences. In **Gauss**, objects are natural numbers and morphisms $a \to b$ are tuples (M, C, s) where M is a matrix in $\mathbb{R}^{b \times a}$, C is a positive semidefinite matrix in $\mathbb{R}^{b \times b}$ and s is a vector in $\mathbb{R}^b$.

Intuitively, the morphisms in **Gauss** represent transformations of random variables. That is, (M, C, s) implicitly represents the following transformation of random variables:

$$g(f) = Mf + \xi_{s,C}.$$

Where $\xi_{s,C}$ is a multivariate normal random variable with mean s and covariance matrix C that is independent of f. If the random variable f is normally distributed, then $g(f)$ is as well.

A primary difference between **Gauss** and $\mathbf{DF}_{\mathcal{N}_\mu}$ is that the morphisms in $\mathbf{DF}_{\mathcal{N}_\mu}$ explicitly include the functional form of $\xi_{s,C}$ in the morphism itself. For any arrow $(M, C, s) : a \to b$ in **Gauss** and a choice of such an $\xi_{s,C}$ over $(\Omega, \mathcal{B}(\Omega), \mu)$, we can form the $\mathbf{DF}_{\mathcal{N}_\mu}$ arrow $f' : \Omega \times \mathbb{R}^0 \times \mathbb{R}^a \to \mathbb{R}^b$ where for $\omega \in \Omega, x_a \in \mathbb{R}^a$:

$$f'(\omega, x_a) = Mx_a + \xi_{s,C}(\omega).$$

However, since this arrow is dependent on the choice of $\xi_{s,C}$, this mapping is not functorial.

5.2.2 Expectation Composition

Definition 5.2. *A subcategory* $\mathbf{C}$ *of* $\mathbf{DF}$ *is an* ***Expectation Composition*** *category if for any* $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ *and* $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ *in* $\mathbf{C}$ *and* $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$*:*

$$\int_{(\omega_m,\omega_n)\in\Omega^{m+n}} f'(\omega_m, x_q, f(\omega_n, x_p, x_a))d\mu^{m+n} =$$
$$\int_{\omega_m\in\Omega^m} f'\left(\omega_m, x_q, \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n\right) d\mu^m.$$

Proposition 9. $\mathbf{DF}_{\mathcal{N}_\mu}$ *is an Expectation Composition category.*

Proof. We will use a proof by induction. By the definition of $\mathbf{DF}_{\mathcal{N}_\mu}$, there exists some $k \in \mathbb{N}$ such that we can express f' as a composition of k arrows in $\mathcal{N}_\mu$. First note that if $k = 1$, then f' is in $\mathcal{N}_\mu$, and the statement must hold since for $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:

$$\int_{(\omega_m,\omega_n)\in\Omega^{m+n}} f'(\omega_m, x_q, f(\omega_n, x_p, x_a))d\mu^{m+n} =$$
$$\int_{(\omega_m,\omega_n)\in\Omega^{m+n}} T'(x_q, f(\omega_n, x_p, x_a)) + G'(\omega_m)d\mu^{m+n} =$$
$$\int_{\omega_m\in\Omega^m}\int_{\omega_n\in\Omega^n} T'(x_q, f(\omega_n, x_p, x_a))d\mu^n + G'(\omega_m)d\mu^m =$$
$$\int_{\omega_m\in\Omega^m} T'\left(x_q, \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n\right) + G'(\omega_m)d\mu^m =$$
$$\int_{\omega_m\in\Omega^m} f'\left(\omega_m, x_q, \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n\right) d\mu^m.$$

Next, if $k > 1$ then we can express $f' = h \circ f'_{k-1}$, where h is in $\mathcal{N}_\mu$ and f'_{k-1} is the composition of $k-1$ arrows in $\mathcal{N}_\mu$. Without loss of generality we will assume f'_{k-1} and h have the following signatures:

$$f'_{k-1} : \Omega^{m'} \times \mathbb{R}^{q'} \times \mathbb{R}^b \to \mathbb{R}^d \qquad h : \Omega^{m''} \times \mathbb{R}^{q''} \times \mathbb{R}^d \to \mathbb{R}^c.$$

Note that $q' + q'' = q$ and $m' + m'' = m$. Now we can show the following, where the step marked $*$ holds by induction and $x_{q''} \in \mathbb{R}^{q''}, x_{q'} \in \mathbb{R}^{q'}, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:

$$\int_{(\omega_{m''},\omega_{m'},\omega_n)\in\Omega^{m''+m'+n}} f'((\omega_{m''},\omega_{m'}),(x_{q''},x_{q'}), f(\omega_n, x_p, x_a))d\mu^{m''+m'+n} =$$
$$\int_{(\omega_{m''},\omega_{m'},\omega_n)\in\Omega^{m''+m'+n}} T_h(x_{q''}, f'_{k-1}(\omega_{m'}, x_{q'}, f(\omega_n, x_p, x_a)) + G_h(\omega_{m''})d\mu^{m''+m'+n} =$$
$$\int_{(\omega_{m''},\omega_{m'})\in\Omega^{m''+m'}} h\left(\omega_{m''}, x_{q''}, \int_{\omega_n\in\Omega^n} f'_{k-1}(\omega_{m'}, x_{q'}, f(\omega_n, x_p, x_a))d\mu^n\right) d\mu^{m''+m'} =^*$$
$$\int_{(\omega_{m''},\omega_{m'})\in\Omega^{m''+m'}} h\left(\omega_{m''}, x_{q''}, f'_{k-1}(\omega_{m'}, x_{q'}, \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a))d\mu^n\right) d\mu^{m''+m'} =$$
$$\int_{(\omega_{m''},\omega_{m'})\in\Omega^{m''+m'}} f'\left((\omega_{m''},\omega_{m'}),(x_{q''},x_{q'}), \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n\right) d\mu^{m''+m'}.$$

By induction we have that the original statement holds for all $f', f \in \mathbf{DF}_{\mathcal{N}_\mu}$. □

For $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in an Expectation Composition category $\mathbf{C}$ and $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ the following function must be differentiable by the Leibniz integration rule:

$$f_E(x_p, x_a) = E_{\mu^n}[f(\_, x_p, x_a)] = \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n.$$

We can therefore define a functor $Exp : \mathbf{C} \to \mathbf{Para}(\mathbf{Euc})$ that acts as the identity on objects and sends the arrow f to f_E.

6 Likelihood and Learning

In this section we will apply the maximum likelihood procedure to the arrows in **DF** to derive the error function $er : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We will then use this error function to define a modification of Fong et al.'s [12] backpropagation functor. However, since different arrows in **DF** have likelihood functions of different forms, we will not define a single backpropagation functor out of **DF**. Instead, we will define multiple functors from subcategories of **DF** into **Learn**.

To do this, we will first define a substructure of **DF** with well-defined likelihood functions. Then, we will describe a class of subcategories of **DF** derived from this substructure. Finally, we will define two backpropagation functors for any subcategory in this class.

6.1 Conditional Likelihood

The **conditional likelihood** is a general measure of the goodness of fit of a set of parameters and observed data for a given parametric statistical model. We can define the conditional likelihood of a parametric statistical model $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ over the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$ at the points $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$ in terms of the pushforward measure of μ^n along the random variable $f(\_, x_p, x_a)$. To do this, we evaluate the Radon-Nikodym derivative of $f(\_, x_p, x_a)_* \mu^n = \mu^n(f(\_, x_p, x_a)^{-1})$ with respect to a reference measure at the point x_b. In this work we select the Lebesgue measure over $\mathbb{R}^b$, λ^b, as the reference measure. Note that the Radon-Nikodym derivative with respect to the Lebesgue measure is not defined for all measures. For example, no discrete measure has a Radon-Nikodym derivative with respect to the Lebesgue measure, since for any finite collection of points A in $\mathbb{R}^b$, $\lambda^b(A) = 0$. Formally the conditional likelihood function for f is $L_f : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ where for $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$:

$$L_f(x_p, x_a, x_b) = \frac{d\ f(\_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b).$$

For example, the conditional likelihood function for the univariate linear regression model l that we introduced in Section 5.1 is $L_l : \mathbb{R}^3 \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ where for $[a, b, s] \in \mathbb{R}^3, x \in \mathbb{R}, y \in \mathbb{R}$:

$$L_l([a, b, s], x, y) = \frac{1}{s\sqrt{2\pi}} exp\left(-\frac{(y - (ax + b))^2}{2s^2}\right).$$

Definition 6.1. *An **abstract conditional likelihood** from $\mathbb{R}^a$ to $\mathbb{R}^b$ is a Borel-measurable and Lebesgue-integrable function of the form $L : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$.*

We can define the composition of the abstract conditional likelihoods $L : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ and $L' : \mathbb{R}^q \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R}$ to be $(L' \circ L) : \mathbb{R}^{q+p} \times \mathbb{R}^a \times \mathbb{R}^c \to \mathbb{R}$ where for $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$:

$$(L' \circ L)((x_q, x_p), x_a, x_c) = \int_{x_b \in \mathbb{R}^b} L'(x_q, x_b, x_c) L(x_p, x_a, x_b) dx_b.$$

Similarly, we can define a tensor product of abstract conditional likelihoods. The tensor of $L' : \mathbb{R}^q \times \mathbb{R}^c \times \mathbb{R}^d \to \mathbb{R}$ and $L : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ is $(L' \otimes L) : \mathbb{R}^{q+p} \times \mathbb{R}^{c+a} \times \mathbb{R}^{d+b} \to \mathbb{R}$ where for $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_c \in \mathbb{R}^c, x_a \in \mathbb{R}^a, x_d \in \mathbb{R}^d, x_b \in \mathbb{R}^b$:

$$(L' \otimes L)((x_q, x_p), (x_c, x_a), (x_d, x_b)) = L'(x_q, x_c, x_d) L(x_p, x_a, x_b).$$

We can define a **monoidal semicategory** of abstract conditional likelihoods, which we name **CondLikelihood**. Monoidal semicategories are similar to monoidal categories but lack identity morphisms.

Definition 6.2. *A monoidal semicategory is a monoid object in* **SemiCat**, *the monoidal category of semicategories.*

The objects in **CondLikelihood** are spaces of the form $\mathbb{R}^n$ for some $n \in \mathbb{N}$. The tensor of the objects $\mathbb{R}^a$ and $\mathbb{R}^b$ in **CondLikelihood** is defined to be $\mathbb{R}^{a+b}$. The unit of this tensor is $\mathbb{R}^0$.

The morphisms between $\mathbb{R}^a$ and $\mathbb{R}^b$ are equivalence classes of abstract conditional likelihood functions such that for $L, L^* : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ we have $L \sim L^*$ if for all $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$, the functions $L(x_p, x_a, \_) : \mathbb{R}^b \to \mathbb{R}$ and $L^*(x_p, x_a, \_) : \mathbb{R}^b \to \mathbb{R}$ are λ^b-a.e. equivalent.

We define the composition and tensor of these equivalence classes in terms of their representatives. That is, consider the equivalence classes $\mathbf{L}$ and $\mathbf{L}'$ and suppose $L_i : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ is in $\mathbf{L}$ and $L'_j : \mathbb{R}^q \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R}$ is in $\mathbf{L}'$. Then the representatives of $\mathbf{L}' \circ \mathbf{L}$ are $L'_j \circ L_i$ for $L_i \in \mathbf{L}, L'_j \in \mathbf{L}'$. Note that for any $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$, the functions $(L'_j \circ L_i)((x_q, x_p), x_a, \_) : \mathbb{R}^c \to \mathbb{R}$ for all $L_i \in \mathbf{L}, L'_j \in \mathbf{L}'$ are λ^c-a.e. equivalent, so **CondLikelihood** is closed under composition. The tensor of equivalence classes is defined similarly.

However, **CondLikelihood** does not form a category, because objects in **CondLikelihood** do not necessarily have identities. For example, for $b > 0$ there is no function $\delta_b : \mathbb{R}^0 \times \mathbb{R}^b \times \mathbb{R}^b \to \mathbb{R}$ such that the following holds for all $L : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ and $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$:

$$(\delta_b \circ L)(x_p, x_a, x_b) = \int_{x'_b \in \mathbb{R}^b} \delta_b(x_b, x'_b) L(x_p, x_a, x'_b) dx'_b = L(x_p, x_a, x_b).$$

Proposition 10. **CondLikelihood** *is a monoidal semicategory.*

Proof. We will first show that **CondLikelihood** is a semicategory. We have already shown that **CondLikelihood** is closed under composition, so we simply need to show that composition is associative. Suppose the following are representatives of three arrows in **CondLikelihood**:

$$f_1 : \mathbb{R}^{p_1} \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R} \qquad f_2 : \mathbb{R}^{p_2} \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R} \qquad f_3 : \mathbb{R}^{p_3} \times \mathbb{R}^c \times \mathbb{R}^d \to \mathbb{R}$$

Now consider the representatives of their composition $f_3 \circ (f_2 \circ f_1) : \mathbb{R}^a \to \mathbb{R}^d$ and $(f_3 \circ f_2) \circ f_1 : \mathbb{R}^a \to \mathbb{R}^d$. For $x_{p_3} \in \mathbb{R}^{p_3}, x_{p_2} \in \mathbb{R}^{p_2}, x_{p_1} \in \mathbb{R}^{p_1}, x_a \in \mathbb{R}^a, x_d \in \mathbb{R}^d$:

$$\begin{aligned}
&(f_3 \circ (f_2 \circ f_1))((x_{p_3}, x_{p_2}, x_{p_1}), x_a, x_d) = \\
&\int_{x_c \in \mathbb{R}^c} f_3(x_{p_3}, x_c, x_d) \left(\int_{x_b \in \mathbb{R}^b} f_2(x_{p_2}, x_b, x_c) f_1(x_{p_1}, x_a, x_b) dx_b \right) dx_c = \\
&\int_{x_b \in \mathbb{R}^b} \left(\int_{x_c \in \mathbb{R}^c} f_3(x_{p_3}, x_c, x_d) f_2(x_{p_2}, x_b, x_c) dx_c \right) f_1(x_{p_1}, x_a, x_b) dx_b = \\
&((f_3 \circ f_2) \circ f_1)((x_{p_3}, x_{p_2}, x_{p_1}), x_a, x_d).
\end{aligned}$$

Therefore, composition in **CondLikelihood** is associative, so **CondLikelihood** is a semicategory. Next, we will show that **CondLikelihood** is a monoid object in **SemiCat**. Note that:

$$\mathbb{R}^a \otimes (\mathbb{R}^b \otimes \mathbb{R}^c) = (\mathbb{R}^a \otimes \mathbb{R}^b) \otimes \mathbb{R}^c = \mathbb{R}^{a+b+c}$$
$$\mathbb{R}^0 \otimes \mathbb{R}^a = \mathbb{R}^a \otimes \mathbb{R}^0 = \mathbb{R}^a.$$

Now suppose the following are representatives of three arrows in **CondLikelihood**:

$$g_1 : \mathbb{R}^{p_1} \times \mathbb{R}^{b_1} \times \mathbb{R}^{a_1} \to \mathbb{R} \qquad g_2 : \mathbb{R}^{p_2} \times \mathbb{R}^{b_2} \times \mathbb{R}^{a_2} \to \mathbb{R} \qquad g_3 : \mathbb{R}^{p_3} \times \mathbb{R}^{b_3} \times \mathbb{R}^{a_3} \to \mathbb{R}.$$

Consider the representatives of their tensor $((g_3 \otimes g_2) \otimes g_1)$ and $(g_3 \otimes (g_2 \otimes g_1))$. For $x_{p_3} \in \mathbb{R}^{p_3}$, $x_{p_2} \in \mathbb{R}^{p_2}$, $x_{p_1} \in \mathbb{R}^{p_1}$, $x_{b_3} \in \mathbb{R}^{b_3}$, $x_{b_2} \in \mathbb{R}^{b_2}$, $x_{b_1} \in \mathbb{R}^{b_1}$, $x_{a_3} \in \mathbb{R}^{a_3}$, $x_{a_2} \in \mathbb{R}^{a_2}$, and $x_{a_1} \in \mathbb{R}^{a_1}$:

$$\begin{aligned}
&((g_3 \otimes g_2) \otimes g_1)(((x_{p_3}, x_{p_2}), x_{p_1}), ((x_{b_3}, x_{b_2}), x_{b_1}), ((x_{a_3}, x_{a_2}), x_{a_1})) = \\
&(g_3(x_{p_3}, x_{b_3}, x_{a_3}) g_2(x_{p_2}, x_{b_2}, x_{a_2})) g_1(x_{p_1}, x_{b_1}, x_{a_1}) = \\
&g_3(x_{p_3}, x_{b_3}, x_{a_3})(g_2(x_{p_2}, x_{b_2}, x_{a_2}) g_1(x_{p_1}, x_{b_1}, x_{a_1})) = \\
&(g_3 \otimes (g_2 \otimes g_1))((x_{p_3}, (x_{p_2}, x_{p_1})), (x_{b_3}, (x_{b_2}, x_{b_1})), (x_{a_3}, (x_{a_2}, x_{a_1}))).
\end{aligned}$$

Therefore, $\otimes$ satisfies the associative law as well as the left and right unit laws. □

If we extend from functions to generalized functions (distributions) we can form a category similar to **CondLikelihood**. For example, Blute et al. [3] define a category **DRel** of tame distributions in which the Dirac delta δ exists as a singular distribution. The semicategory **CondLikelihood** is similar in spirit to the nuclear ideal of **DRel** that Blute et al. describe. However, we will use conditional likelihood functions to define optimization objectives, and there is no obvious way to do this

with a singular distribution. For this reason we will keep **CondLikelihood** as a monoidal semicategory. Next, given a probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ define $\mathbf{DF}_{\mathcal{R}_\mu}$ to be the substructure of **DF** with the same objects, but with morphisms between $\mathbb{R}^a$ and $\mathbb{R}^b$ limited to $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ such that the following Borel-measurable and Lebesgue-integrable function exists:

$$L_f(x_p, x_a, x_b) = \frac{d\ f(\_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b)$$

where $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$.

Proposition 11. $\mathbf{DF}_{\mathcal{R}_\mu}$ *is a monoidal semicategory.*

Proof. We will first show that $\mathbf{DF}_{\mathcal{R}_\mu}$ is closed under composition. Suppose $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ are arrows in $\mathbf{DF}_{\mathcal{R}_\mu}$. We can show that for all $x_a \in \mathbb{R}^a, x_p \in \mathbb{R}^p, x_q \in \mathbb{R}^q$ there exists some Borel-measurable and Lebesgue integrable $g : \mathbb{R}^c \to \mathbb{R}$ such that for $\sigma_c \in \mathcal{B}(\mathbb{R}^c)$:

$$(f' \circ f)(\_, (x_q, x_p), x_a)_* \mu^{m+n}(\sigma_c) = \int_{x_c \in \sigma_c} g(x_c) d\lambda^c$$

where λ^c is the Lebesgue measure over $\mathbb{R}^c$:

$$\begin{aligned}
&(f' \circ f)(_, (x_q, x_p), x_a)_* \mu^{m+n}(\sigma_c) = \\
&\int_{x_b \in \mathbb{R}^b} f'(_, x_q, x_b)_* \mu^m(\sigma_c)\ \ df(_, x_p, x_a)_* \mu^n = \\
&\int_{x_b \in \mathbb{R}^b} \left[\int_{x_c \in \sigma_c} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x_c) d\lambda^c \right] \left[\frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b) d\lambda^b \right] = \\
&\int_{x_c \in \sigma_c} \left[\left(\int_{x_b \in \mathbb{R}^b} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x_c) \right) \left(\frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b) d\lambda^b \right) \right] d\lambda^c.
\end{aligned}$$

Next, we will show that $\mathbf{DF}_{\mathcal{R}_\mu}$ is closed under tensor. Suppose $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^c \to \mathbb{R}^d$ are arrows in $\mathbf{DF}_{\mathcal{R}_\mu}$. We can show that for all $x_q \in \mathbb{R}^q$, $x_p \in \mathbb{R}^p$, $x_c \in \mathbb{R}^c$, $x_a \in \mathbb{R}^a$ there exists some measurable $g : \mathbb{R}^d \times \mathbb{R}^b \to \mathbb{R}$ such that for $\sigma_d \times \sigma_b \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^b)$:

$$(f' \otimes f)(\_, (x_q, x_p), (x_c, x_a))_* \mu^{m+n}(\sigma_d \times \sigma_b) = \int_{(x_d, x_b) \in \sigma_d \times \sigma_b} g(x_d, x_b) d\lambda^{d+b}$$

where λ^{d+b} is the Lebesgue measure over $\mathbb{R}^{d+b}$:

$$\begin{aligned}
&(f' \otimes f)(_, (x_q, x_p), (x_c, x_a))_* \mu^{m+n}(\sigma_d \times \sigma_b) = \\
&(f'(_, x_q, x_c)_* \mu^m(\sigma_d))\, (f(_, x_p, x_a)_* \mu^n(\sigma_b)) = \\
&\left[\int_{x_d \in \sigma_d} \frac{df'(_, x_q, x_c)_* \mu^m}{d\lambda^d}(x_d) d\lambda^d \right] \left[\int_{x_b \in \sigma_b} \frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b) d\lambda^b \right] = \\
&\int_{(x_d, x_b) \in \sigma_d \times \sigma_b} \left[\left(\frac{df'(_, x_q, x_c)_* \mu^m}{d\lambda^d}(x_d) \right) \left(\frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b) \right) \right] d\lambda^{d+b}.
\end{aligned}$$

□

Next, we can define the mapping $\mathcal{RN}_\mu : \mathbf{DF}_{\mathcal{R}_\mu} \to \mathbf{CondLikelihood}$ that acts as the identity on objects and sends any morphism $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in $\mathbf{DF}_{\mathcal{R}_\mu}$ to the equivalence class that contains the function $\mathcal{RN}_\mu f : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ where for $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$:

$$\mathcal{RN}_\mu f(x_p, x_a, x_b) = \frac{df(\_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b).$$

Note that Proposition 11 implies that this function exists.

Definition 6.3. *A strict monoidal semifunctor is a semifunctor* $F : \mathbf{C} \to \mathbf{D}$ *such that:*

- $F(1_\mathbf{C}) = 1_\mathbf{D}$

- *For* $o_1, o_2 \in Ob(\mathbf{C})$, $F(o_1 \otimes o_2) = F(o_1) \otimes F(o_2)$
- *For* $a_1, a_2 \in Ar(\mathbf{C})$, $F(a_1 \otimes a_2) = F(a_1) \otimes F(a_2)$

Proposition 12. *$\mathcal{RN}_\mu$ is a strict monoidal semifunctor.*

Proof. We will first show that $\mathcal{RN}_\mu$ is a semifunctor. Suppose $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ are arrows in $\mathbf{DF}_{\mathcal{R}_\mu}$. Then for $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$:

$$\begin{aligned}
\mathcal{RN}_\mu(f' \circ f)((x_q, x_p), x_a, x_c)) = \\
\frac{d(f' \circ f)(_, (x_q, x_p), x_a)_* \mu^{m+n}}{d\lambda^c}(x_c) = \\
\frac{d \int_{x_b \in \mathbb{R}^b} f'(_, x_q, x_b)_* \mu^m((_)_c) \;\; df(_, x_p, x_a)_* \mu^n}{d\lambda^c}(x_c) = \\
\frac{d \int_{x_b \in \mathbb{R}^b} \left[\int_{x'_c \in ((_)_c)} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x'_c) d\lambda^c\right] df(_, x_p, x_a)_* \mu^n}{d\lambda^c}(x_c) = \\
\frac{d \int_{x_b \in \mathbb{R}^b} \left[\int_{x'_c \in ((_)_c)} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x'_c) d\lambda^c\right] \left[\frac{df(_, x_p, x_a)_* \mu^n)}{d\lambda^b}(x_b) d\lambda^b\right]}{d\lambda^c}(x_c) = \\
\frac{d \int_{x'_c \in (_)_c} \left[\int_{x_b \in \mathbb{R}^b} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x'_c) \;\; \frac{df(_, x_p, x_a)_* \mu^n)}{d\lambda^b}(x_b) d\lambda^b\right] d\lambda^c}{d\lambda^c}(x_c) = \\
\int_{x_b \in \mathbb{R}^b} \frac{df'(_, x_q, x_b)_* \mu^m}{d\lambda^c}(x_c) \;\; \frac{df(_, x_p, x_a)_* \mu^n)}{d\lambda^b}(x_b) d\lambda^b = \\
(\mathcal{RN}_\mu f' \circ \mathcal{RN}_\mu f)((x_q, x_p), x_a, x_c).
\end{aligned}$$

Next, we will show that $\mathcal{RN}_\mu$ satisfies the conditions in Definition 6.3. Since $\mathcal{RN}_\mu$ is identity-on-objects, the monoidal tensor has the same action on objects in $\mathbf{DF}_{\mathcal{R}_\mu}$ and **CondLikelihood** and the monoidal unit in both $\mathbf{DF}_{\mathcal{R}_\mu}$ and **CondLikelihood** is $\mathbb{R}^0$, the first and second conditions are trivial. To see that $\mathcal{RN}_\mu$ satisfies the third condition, suppose $f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^c \to \mathbb{R}^d$ are arrows in $\mathbf{DF}_{\mathcal{R}_\mu}$. Then for $x_q \in \mathbb{R}^q$, $x_p \in \mathbb{R}^p$, $x_c \in \mathbb{R}^c$, $x_a \in \mathbb{R}^a$, $x_d \in \mathbb{R}^d$, $x_b \in \mathbb{R}^b$:

$$\begin{aligned}
\mathcal{RN}_\mu(f' \otimes f)((x_q, x_p), (x_c, x_a), (x_d, x_b)) = \\
\frac{d\left[(f' \otimes f)(_, (x_q, x_p), (x_c, x_a))_* \mu^{m+n}\right]}{d\lambda^{d+b}}((x_d, x_b)) = \\
\frac{d\left[[f'(_, x_q, x_c)_* \mu^m((_)_d)]\,[f(_, x_p, x_a)_* \mu^n((_)_b)]\right]}{d\lambda^{d+b}}((x_d, x_b)) = \\
\frac{d\left[\left(\int_{x'_d \in (_)_d} \frac{df'(_, x_q, x_c)_* \mu^m}{d\lambda^d}(x'_d) d\lambda^d\right)\left(\int_{x'_b \in (_)_b} \frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x'_b) d\lambda^b\right)\right]}{d\lambda^{d+b}}((x_d, x_b)) = \\
\frac{d\left[\int_{(x'_d, x'_b) \in (_)_d \times (_)_b} \left(\frac{df'(_, x_q, x_c)_* \mu^m}{d\lambda^d}(x'_d)\right)\left(\frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x'_b)\right) d\lambda^{d+b}\right]}{d\lambda^{d+b}}((x_d, x_b)) = \\
\left(\frac{df'(_, x_q, x_c)_* \mu^m}{d\lambda^d}(x_d)\right)\left(\frac{df(_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b)\right) = \\
(\mathcal{RN}_\mu f' \otimes \mathcal{RN}_\mu f)((x_q, x_p), (x_c, x_a), (x_d, x_b)).
\end{aligned}$$

□

6.2 Maximum Likelihood

Suppose we have a probability space $(\mathbb{R}^a \times \mathbb{R}^b, \mathcal{B}(\mathbb{R}^a \times \mathbb{R}^b), \tau)$ such that for each $x_a \in \mathbb{R}^a$, the map $\tau(x_a, \_) : \mathcal{B}(\mathbb{R}^b) \to [0, 1]$ is a probability measure. Suppose that we also have an arrow $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in $\mathbf{DF}_{\mathcal{R}_\mu}$ and we want to find the $x_p \in \mathbb{R}^p$ such that for each $x_a \in \mathbb{R}^a$, the distribution $f(\_, x_p, x_a)_* \mu^n : \mathcal{B}(\mathbb{R}^b) \to [0, 1]$ best approximates $\tau(x_a, \_) : \mathcal{B}(\mathbb{R}^b) \to [0, 1]$. The

maximum expected log-likelihood estimator for f with respect to τ is the vector $x_p \in \mathbb{R}^p$ that maximizes the following function:

$$L_\tau(x_p) = \int_{(x_a,x_b)\in\mathbb{R}^a\times\mathbb{R}^b} log\frac{df(\_,x_p,x_a)_*\mu^n}{d\lambda^b}(x_b)d\tau.$$

That is, the maximum expected log-likelihood estimator for f with respect to τ is the vector x_p that maximizes the expected value of $log\frac{df(\_,x_p,x_a)_*\mu^n}{d\lambda^b}(x_b)$ over τ. Equivalently, x_p minimizes the weighted sum over x_a of the KL-divergences between $f(\_,x_p,x_a)_*\mu^n$ and $\tau(x_a,\_)$, where the weight of each x_a is determined by τ [23].

Now suppose that instead of observing a probability space $(\mathbb{R}^a\times\mathbb{R}^b, \mathcal{B}(\mathbb{R}^a\times\mathbb{R}^b), \tau)$ directly we have a dataset of samples $S_n = \{(x_{a_1},x_{b_1}),(x_{a_2},x_{b_2}),...,(x_{a_n},x_{b_n})\}$ in $\mathbb{R}^a\times\mathbb{R}^b$. The **maximum log likelihood estimator** for f with respect to this dataset is the vector $x_p \in \mathbb{R}^p$ that maximizes the function:

$$L_{S_n}(x_p) = \sum_{i=1}^{n} log\frac{df(\_,x_p,x_{a_i})_*\mu^n}{d\lambda^b}(x_{b_i}).$$

Note that if we assume the samples in S_n are drawn from $(\mathbb{R}^a\times\mathbb{R}^b, \mathcal{B}(\mathbb{R}^a\times\mathbb{R}^b), \tau)$, then by the weak law of large numbers $\frac{1}{n}L_{S_n}$ converges to L_τ in probability as $n\to\infty$.

However, it will be challenging to derive an objective function for Fong et al.'s [12] backpropagation functor from L_{S_n} directly, since their construction assumes that the error function has the signature $er : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and has an invertible derivative. We will slightly modify L_{S_n} to make this easier.

For any $j \le b$, the jth component of f is the function $f[j] : \Omega^n\times\mathbb{R}^p\times\mathbb{R}^a\to\mathbb{R}$ and the marginal likelihood at $x_p\in\mathbb{R}^p$ of this component for some sample $(x_{a_i},x_{b_i})\in S_n$ is:

$$l_{ij}(x_p) = \frac{df(\_,x_p,x_{a_i})[j]_*\mu^n}{d\lambda}(x_{b_i}[j])$$

where we write $x_{b_i}[j]$ for the jth component of x_{b_i}. The **maximum log-marginal likelihood estimator** for f with respect to this dataset is then the vector $x_p\in\mathbb{R}^p$ that maximizes the function:

$$M_{S_n}(x_p) = \sum_{i=1}^{n}\sum_{j=1}^{b} log\ l_{ij}(x_p).$$

Note that $M_{S_n}(x_p) = L_{S_n}(x_p)$ when the real-valued random variables $f(\_,x_p,x_{a_i})[j]$ are mutually independent for all x_{a_i}.

This suggests a criterion for an error function $er : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ over which we can define Fong et al.'s [12] backpropagation functor: we want the following two real-valued functions of $\mathbb{R}^p$ to move in tandem for any fixed $(x_a,y)\in\mathbb{R}^a\times\mathbb{R}$ and $j\le b$:

$$l(x_p) = er\left(E_{\mu^n}[f(\_,x_p,x_a)[j]],y\right) \qquad l'(x_p) = \frac{df(\_,x_p,x_a)[j]_*\mu^n}{d\lambda}(y).$$

We will now make this formal.

6.3 Learning from Likelihoods

Suppose we have a real-valued random variable f over the probability space $(\Omega^n,\mathcal{B}(\Omega^n),\mu^n)$. Write $E_{\mu^n}[f]\in\mathbb{R}$ for the expectation of f over μ^n:

$$E_{\mu^n}[f] = \int_{\omega^n\in\Omega^n} f(\omega^n)\ d\mu^n.$$

And define f^0 to be:

$$f^0(\omega^n) = f(\omega^n) - E_{\mu^n}[f].$$

Next, suppose U : **Cat** $\to$ **SemiCat** is the forgetful functor.

Definition 6.4. *An Expectation Composition category* $\mathbf{C}$ *is a* ***Marginal Likelihood Factorization Category*** *over the measure* $\mu : \mathcal{B}(\Omega) \to [0,1]$ *if the following cospan in* **SemiCat**, *where inc and* inc' *are respectively the inclusion maps of* $U(\mathbf{C})$ *and* $\mathbf{DF}_{\mathcal{R}_\mu}$ *into* $U(\mathbf{DF})$

$$U(\mathbf{C}) \xhookrightarrow{inc} U(\mathbf{DF}) \xhookleftarrow{inc'} \mathbf{DF}_{\mathcal{R}_\mu}$$

has a pullback $U(\mathbf{C}) \xleftarrow{h_l} \mathbf{C}_{\mathcal{R}_\mu} \xrightarrow{h_r} \mathbf{DF}_{\mathcal{R}_\mu}$ *that satisfies the following property. There exists:*

- *A differentiable function with invertible derivative* $er : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
- *For each* $n \in \mathbb{N}$, *a function* $\alpha_n : (\Omega^n \to \mathbb{R}) \to \mathbb{R}$
- *For each* $n \in \mathbb{N}$, *a non-negative function* $\beta_n : (\Omega^n \to \mathbb{R}) \to \mathbb{R}$

such that for any $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, j \leq b$ *and arrow in the semicategory* $\mathbf{C}_{\mathcal{R}_\mu}$ *whose image under* $inc \circ h_l : \mathbf{C}_{\mathcal{R}_\mu} \to U(\mathbf{DF})$ *has the signature* $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$, *we can write:*

$$log \frac{df(\_, x_p, x_a)[j]_* \mu^n}{d\lambda}(y) =$$
$$\alpha_n(f^0(\_, x_p, x_a)[j]) - \beta_n(f^0(\_, x_p, x_a)[j]) er\left(E_{\mu^n}[f(\_, x_p, x_a)[j]], y\right).$$

We will refer to er as a ***marginal error function*** *of* $\mathbf{C}$.

Proposition 13. $\mathbf{DF}_{\mathcal{N}_\mu}$ *is a Marginal Likelihood Factorization Category with a marginal error function* $er(a,b) = (a-b)^2$.

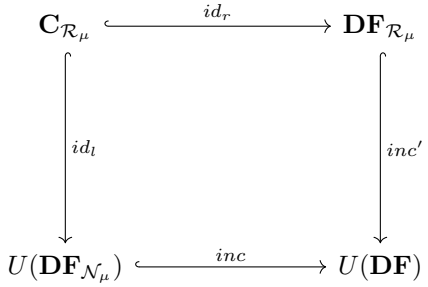
Proof. To begin, consider the structure $\mathbf{C}_{\mathcal{R}_\mu}$ that has the same objects as $\mathbf{DF}_{\mathcal{R}_\mu}$ and:

$$\mathbf{C}_{\mathcal{R}_\mu}[\mathbb{R}^a, \mathbb{R}^b] = U(\mathbf{DF}_{\mathcal{N}_\mu})[\mathbb{R}^a, \mathbb{R}^b] \cap \mathbf{DF}_{\mathcal{R}_\mu}[\mathbb{R}^a, \mathbb{R}^b].$$

Since $U(\mathbf{DF}_{\mathcal{N}_\mu})$ and $\mathbf{DF}_{\mathcal{R}_\mu}$ are small, this intersection is well-defined and $\mathbf{C}_{\mathcal{R}_\mu}$ is a semicategory. Now note that there exist identity-on-objects and identity-on-morphisms inclusion semifunctors

$$id_l : \mathbf{C}_{\mathcal{R}_\mu} \hookrightarrow U(\mathbf{DF}_{\mathcal{N}_\mu}) \qquad id_r : \mathbf{C}_{\mathcal{R}_\mu} \hookrightarrow \mathbf{DF}_{\mathcal{R}_\mu}$$

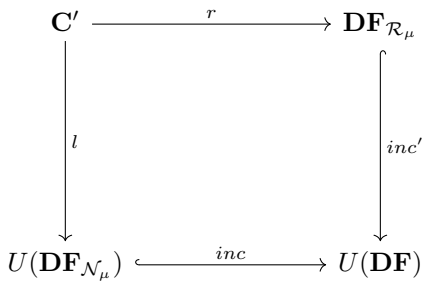
such that the following diagram commutes:



Now consider any other semicategory $\mathbf{C}'$ equipped with monic semifunctors:

$$l : \mathbf{C}' \to U(\mathbf{DF}_{\mathcal{N}_\mu}) \qquad r : \mathbf{C}' \to \mathbf{DF}_{\mathcal{R}_\mu}$$

such that the following diagram commutes:



Since inc and inc' are inclusion maps, l and r must act identically on objects and morphisms. Therefore, any object or morphism in the image of l or r must also be in $\mathbf{C}_{\mathcal{R}_\mu}$, so we can define the unique semifunctor $h : \mathbf{C}' \to \mathbf{C}_{\mathcal{R}_\mu}$ that has the same action on objects and morphisms as l and r. This implies that

$$id_l \circ h = l \qquad id_r \circ h = r.$$

And so $\mathbf{C}_{\mathcal{R}_\mu}$ is the pullback of the diagram:

$$U(\mathbf{DF}_{\mathcal{N}_\mu}) \xrightarrow{inc} U(\mathbf{DF}) \xleftarrow{inc'} \mathbf{DF}_{\mathcal{R}_\mu}$$

Next, consider some $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in $\mathbf{C}_{\mathcal{R}_\mu}$, and note that for any $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, j \leq b$, the random variable $f(\_, x_p, x_a)[j]$ is univariate normal. For each $n \in \mathbb{N}$ we also define the standard deviation function $s_n : (\Omega^n \to \mathbb{R}) \to \mathbb{R}$ where for $g : \Omega^n \to \mathbb{R}$:

$$s_n(g) = \sqrt{E_{\mu_n}[(g - E_{\mu_n}[g])^2]}.$$

Now for any $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, y \in \mathbb{R}, j \leq b$ we can write:

$$log\ \frac{df(\_, x_p, x_a)[j]_* \mu^n}{d\lambda}(y) =$$

$$log\ \frac{1}{s_n(f(\_, x_p, x_a)[j])\sqrt{2\pi}} exp\left(-\left(\frac{y - E_{\mu^n}[f(\_, x_p, x_a)[j]]}{4 s_n(f(\_, x_p, x_a)[j])}\right)^2\right) =$$

$$-\frac{log(2\pi s_n(f(\_, x_p, x_a)[j])^2)}{2} - \frac{1}{2 s_n(f(\_, x_p, x_a)[j])^2}\left(y - E_{\mu^n}[f(\_, x_p, x_a)[j]]\right)^2.$$

Therefore:

$$\alpha_n(g) = -\frac{log(2\pi s_n(g)^2)}{2} \qquad \beta_n(g) = \frac{1}{2 s_n(g)^2} \qquad er(a, b) = (a - b)^2.$$

□

6.4 Backpropagation Functors

For any Marginal Likelihood Factorization Category $\mathbf{C}$ and choice of learning rate ϵ we can define two kinds of backpropagation functors: one into Fong et al.'s **Learn** category [12] and one into a probabilistic analog of **Learn**.

We will first show the functor that maps $\mathbf{C}$ into **Learn**. Write F_{er} for Fong et al.'s Backpropagation functor with learning rate ϵ under the marginal error function er of $\mathbf{C}$. Then we can define the following functor that maps a parametric statistical model in $\mathbf{C}$ to a learning algorithm:

$$E_{er} : \mathbf{C} \to \mathbf{Learn}$$
$$E_{er} = F_{er} \circ Exp.$$

For example, this functor sends parametric statistical models in $\mathbf{DF}_{\mathcal{N}_\mu}$ to learning algorithms that minimize the square error function with gradient descent. We can think of E_{er} as a point estimation functor: it sends an arrow f in $\mathbf{C}$ to a learner whose inference function is formed from f's expectation. The higher order moments of the pushforward distributions of the arrows in $\mathbf{C}$ are only used to define the loss function er.

Next, consider the strict symmetric monoidal subcategory $\mathbf{Learn}_\mathbb{R}$ of **Learn** where objects are restricted to be $\mathbb{R}^n, n \in \mathbb{N}$ and the tensor of objects is $\mathbb{R}^n \otimes \mathbb{R}^m = \mathbb{R}^{n+m}$. Now given the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where $\Omega = \mathbb{R}^k, k \in \mathbb{N}$, we can form the category $\mathbf{Para}_{(\Omega, \mathcal{B}(\Omega))^*}(\mathbf{Learn}_\mathbb{R})$. A morphism between $\mathbb{R}^a$ and $\mathbb{R}^b$ in $\mathbf{Para}_{(\Omega, \mathcal{B}(\Omega))^*}(\mathbf{Learn}_\mathbb{R})$ is a tuple (I, U, r) where I, U, r are functions of types:

$$I : \mathbb{R}^p \times \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b$$
$$U : \mathbb{R}^p \times \Omega^n \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}^p$$
$$r : \mathbb{R}^p \times \Omega^n \times \mathbb{R}^a \times \mathbb{R}^b \to \Omega^n \times \mathbb{R}^a.$$

Intuitively, we can think of such a morphism as a statistical learner in which each of the inference, update and request functions are stochastic processes over $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$.

Now since $\mathbf{DF} = \mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{Para}(\mathbf{Euc}))$, by Proposition 3 the mapping:

$$P_{er} : \mathbf{DF} \to \mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{Learn}_{\mathbb{R}})$$

that applies the same actions on objects and arrows as F_{er} is a strict monoidal functor. Unlike E_{er} however, this functor does not define the gradient update for the statistical model f in terms of its expectation. Instead, given a parameter vector $x_p \in \mathbb{R}^p$, input vector $x_a \in \mathbb{R}^a$ and output vector $x_b \in \mathbb{R}^b$, the update function U in the image of P_{er} will generate different updates for different samples of ω_n from $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$. This is similar to how Tensorflow Probability [22] defines the update step for Distribution layers.

7 Discussion and Future Work

Consider once again a physical system that is composed of several components, each of which has some degree of aleatoric uncertainty. If we construct a neural network model for this system like we describe in Section 1, we cannot characterize the interactions between the uncertainty in the different parts of the system. However, if we model the components of the system as stochastic processes and apply **DF** composition, we can capture how the uncertainty of the component parts combine. For example, given estimates of the kind of uncertainty inherent to the photorecepters in the eye, edge-detecting neurons in primary visual cortex, and higher-order feature detectors in the later stages of visual cortex, we may be able to build a more realistic model of how these sources of uncertainty interact than the one that Eberhardt et al. [8] use to assess how the visual cortex performs a rapid stimulus categorization task.

Once we build such a model, we can use either E_{er} or P_{er} to derive a Learner with a structure that incorporates this combined uncertainty. The functor E_{er} will convert the model to a point estimator and bundle the combined uncertainty into a loss function. In contrast, P_{er} will preserve the uncertainty and produce a learning algorithm where both forward and backward passes are stochastic.

One of the largest differences between this construction and those of Cho and Jacobs [4] and Culbertson and Sturtz [6] is the treatment of model updates in the face of new data. While these authors also describe categorical frameworks in which we can model how a new observation updates the parameters of a statistical model, they primarily study Bayesian algorithms in which the model parameters are represented with a probability distribution.

In contrast, our construction is inherently frequentist. While the backpropagation functors above aim to find an optimal parameter value given the data we have seen, they make no assumptions about what that value may be. Although uncertainty motivates the objective that our parameter estimation procedure aims to optimize, the optimization algorithm does not use it directly. Therefore, a potential future direction for this work is to extend the category **DF** of deterministic and frequentist models to handle generative algorithms that model uncertainty in the input vector and Bayesian algorithms that model uncertainty in the parameter vector.

Furthermore, our current definition of Marginal Likelihood Factorization Categories may be overly restrictive. For example, our definition specifies that each category is characterized by a single marginal error function er. This makes it challenging to build a theory for how we could compose Marginal Likelihood Factorization Categories with different marginal error functions. Another potential future direction would be to relax the restrictions on these categories or prove that they are necessary.

8 Appendix A: Extra Proofs

8.1 Proof of Proposition 5

Proof. First, let's note that **PEuc** is semicartesian because the monoidal unit $(\mathbb{R}^0, \mathcal{B}(\mathbb{R})^0)$ is the terminal object. Next, we will show that cp and dc satisfy the conditions in Definition 2.1 of Fritz et al. [14]. Note that we write the symmetric swap map as $\sigma : \mathbb{R}^a \otimes \mathbb{R}^b \to \mathbb{R}^b \otimes \mathbb{R}^a$.

Commutative Comonoid Condition (Equation 2.2 in Fritz et al. [14])

$$\begin{aligned}(id \otimes cp) \circ cp \circ f = \\ (id \otimes cp) \circ (f \otimes f) = \\ (f \otimes (f \otimes f)) = \\ ((f \otimes f) \otimes f) = \\ (cp \circ f) \otimes (id \circ f) = \\ (cp \otimes id) \circ (f \otimes f) = \\ (cp \otimes id) \circ cp \circ f.\end{aligned}$$

Commutative Comonoid Condition (Equation 2.3a in Fritz et al. [14])

$$(dc \otimes id) \circ cp \circ f = (dc \otimes id) \circ (f \otimes f) = 1 \otimes f = f \otimes 1 = (id \circ f) \otimes (dc \circ f) = (id \otimes dc) \circ cp \circ f.$$

Commutative Comonoid Condition (Equation 2.3b in Fritz et al. [14])

$$\sigma \circ \sigma \circ cp \circ f = cp \circ f.$$

Compatibility with the Monoid Structure (Equation 2.4.a in Fritz et al. [14])

$$dc \circ (f \otimes f') = 1 = 1 \otimes 1 = (dc \circ f) \otimes (dc \circ f').$$

Compatibility with the Monoid Structure (Equation 2.4.b in Fritz et al. [14])

$$\begin{aligned}cp \circ (f \otimes f') = \\ (f \otimes f') \otimes (f \otimes f') = \\ f \otimes (f' \otimes f) \otimes f' = \\ f \otimes (\sigma \circ (f \otimes f')) \otimes f' = \\ (id \otimes \sigma \otimes id) \circ [f \otimes f \otimes f' \otimes f'] = \\ (id \otimes \sigma \otimes id) \circ [(cp \circ f') \otimes (cp \circ f')].\end{aligned}$$

Naturality of dc (Equation 2.5 in Fritz et al. [14])

$$dc \circ f = 1 = dc \circ (f' \circ f).$$

□

References

[1] R.B. Ash, M.F. Gardner, and M.F. Gardner. *Topics in Stochastic Processes*. Probability and Mathematical Statistics: a series of monographs and textbooks. Academic Press, 1975.

[2] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, second edition, 1986. Available at https://www.colorado.edu/amath/sites/default/files/attached-files/billingsley.pdf.

[3] Richard Blute, Prakash Panangaden, and Dorette Pronk. Conformal field theory as a nuclear functor. *Electronic Notes in Theoretical Computer Science*, 172:101–132, 2007. https://doi.org/10.1016/j.entcs.2007.02.005.

[4] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019. https://doi.org/10.1017/S0960129518000488.

[5] Robin Cockett, Geoffrey Cruttwell, Jonathan Gallagher, Jean-Simon Pacaud Lemay, Benjamin MacAdam, Gordon Plotkin, and Dorette Pronk. Reverse derivative categories. *arXiv preprint arXiv*, 2019. https://arxiv.org/abs/1910.07065.

[6] Jared Culbertson and Kirk Sturtz. Bayesian machine learning via category theory. *arXiv preprint*, 2013. https://arxiv.org/abs/1312.1445.

[7] Jared Culbertson and Kirk Sturtz. A categorical foundation for Bayesian probability. *Applied Categorical Structures*, 22(4):647–662, 2014. https://doi.org/10.1007/s10485-013-9324-9.

[8] Sven Eberhardt, Jonah Cader, and Thomas Serre. How deep is the feature analysis underlying rapid visual categorization? *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1108–1116, 2016. https://dl.acm.org/doi/10.5555/3157096.3157220.

[9] Conal Elliott. The simple essence of automatic differentiation. *Proceedings of the ACM on Programming Languages*, 2(ICFP):1–29, 2018. https://doi.org/10.1145/3236765.

[10] Brendan Fong. Causal theories: A categorical perspective on Bayesian networks. *arXiv preprint*, 2013. PhD Thesis, available at https://arxiv.org/abs/1301.6201.

[11] Brendan Fong and Michael Johnson. Lenses and learners. *arXiv preprint*, 2019. https://arxiv.org/abs/1903.03671.

[12] Brendan Fong, David Spivak, and Rémy Tuyéras. Backprop as functor: A compositional perspective on supervised learning. In *2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–13. IEEE, 2019. https://doi.org/10.1109/LICS.2019.8785665.

[13] Uwe Franz. What is stochastic independence? In *Non-commutativity, infinite-dimensionality and probability at the crossroads*, pages 254–274. World Scientific, 2002. Available at https://arxiv.org/abs/math/0206017.

[14] Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. https://doi.org/10.1016/j.aim.2020.107239.

[15] Tobias Fritz and Eigil Fjeldgren Rischel. The zero-one laws of Kolmogorov and Hewitt–Savage in categorical probability. *arXiv preprint arXiv*, 2019. https://arxiv.org/abs/1912.02769.

[16] Bruno Gavranovic. Compositional deep learning. *arXiv preprint*, 2019. https://arxiv.org/abs/1907.08292.

[17] Malte Gerhold, Stephanie Lachs, and Michael Schürmann. Categorial Lévy processes. *arXiv preprint*, 2016. https://arxiv.org/abs/1612.05139.

[18] Michele Giry. A categorical approach to probability theory. In *Categorical aspects of topology and analysis*, pages 68–85. Springer, 1982. https://doi.org/10.1007/BFb0092872.

[19] Chris Heunen, Ohad Kammar, Sam Staton, and Hongseok Yang. A convenient category for higher-order probability theory. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12. IEEE, 2017. https://doi.org/10.1109/LICS.2017.8005137.

[20] Steven P Lalley. Lévy processes, stable processes, and subordinators. 2007. Available at http://galton.uchicago.edu/~lalley/Courses/385/LevyProcesses.pdf.

[21] F William Lawvere. The category of probabilistic mappings. *Unpublished preprint*, 1962.

[22] Abadi Martín. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. https://www.tensorflow.org/.

[23] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. Available at https://www.cs.ubc.ca/~murphyk/MLbook/.

[24] Terence Tao. A review of probability theory, 2010. Blog post, retrieved on 2021/04/04, available at https://terrytao.wordpress.com/2010/01/01/254a-notes-0-a-review-of-probability-theory.

[25] Edgar Y. Walker, R. James Cotton, Wei Ji Ma, and Andreas S. Tolias. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23:122–129, 2020. https://doi.10.1038/s41593-019-0554-5.