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**.

1 Introduction

Many machine learning algorithms contain an irreducible aspect of randomness. Using category theory to reason about how this randomness breaks down into compositional and functorial structure helps us build a high-level picture of probabilistic learning and its connections to other fields. There are two kinds of uncertainty that most probabilistic reasoning aims to capture:

Definition 1.1. *Epistemic uncertainty is uncertainty due to limited data or knowledge.*

For example, if we have a very small amount of data then we need to cope with high epistemic uncertainty. Cho et al. [5] and Culbertson et al. [8; 7] explore how new data points affect their models' epistemic uncertainty.

Definition 1.2. *Aleatoric uncertainty is inherent uncertainty in a system that can cause results to differ each time we run the same experiment*

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. *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 distr over the possible outputs conditioned on some input [21].

Even models that produce point estimates, such as the ones described by Fong et al. [10], 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 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

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

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

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 (MSE 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.'s [10], 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. [9] 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.

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 NNs that have stochastic processes as layers. To do this, we replace the domain of Fong et al.'s [10] Backpropagation functor **Para** 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.

specific contributions:

- Develop a strategy for composing stochastic processes that is compatible with both subordination [17] and parametric function composition [10].
- Introduce two categories with this compositional structure, one based on **Para** [10] and one based on the co-Kleisli category of the product comonad, 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 wrt 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 [10].

2 Preliminaries

2.1 Probability

2.1.1 Probability Measures, RVs and Markov Kernels

We can also form measurable spaces from topological spaces.

Definition 2.5. *Given the topological space Ω the Borel σ-algebra $\mathcal{B}(\Omega)$ of Ω is the σ-algebra generated by the collection of open subsets of Ω. We call the measurable space $(\Omega, \mathcal{B}(\Omega))$ a Borel measurablespace.*

Definition 2.6. *A standard Borel space is a Borel measurable space associated with a* ***Polish space*** *. Standard Borel spaces are closed under countable products.*

The fundamental objects in measure-theoretic probability are the probability measure and probability space:

Definition 2.7. *A probability space is a triplet (Ω, Σ, μ) where (Ω, Σ) is a measurable space (Definition 2.2) 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.*

Probability spaces are closed under products. This is, when (Ω, Σ, μ) and (Ω', Σ', μ') are probability spaces the product space $(\Omega \times \Omega', \Sigma \times \Sigma', \mu\mu')$ where $\mu\mu'(\omega) = \mu(\omega)\mu'(\omega)$ is also a probability space.

In practice, we will generally work with parameterized probability measures, which we call Markov kernels.

Definition 2.8. *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).

Another foundational object in measure-theoretic probability is the random variable, which is paradoxically neither random nor a variable.

Definition 2.9. *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.

Rvs and probability measures are closely related:

Definition 2.10. *Given a probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ and a rv $f : \Omega \to \mathbb{R}$, thepushforward $f_*\mu$ of μ along f is a probability measure over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined to be:*

$$f_*\mu : \mathcal{B}(\mathbb{R}) \to [0,1]$$
$$f_*\mu(\sigma_{\mathbb{R}}) = \mu(f^{-1}(\sigma_{\mathbb{R}})).$$

Like probability measures, rvs have a parameterized extension.

Definition 2.11. *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 $f : \Omega \times T \to \mathbb{R}$. We 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 following Markov Kernel:

$$f_*\mu : T \times \mathcal{B}(\mathbb{R}) \to [0,1]$$
$$f_*\mu(x_t, \sigma_\mathbb{R}) = f(\ , x_t)_*\mu(\sigma_\mathbb{R}) = \mu(f(\ , x_t)^{-1}(\sigma_\mathbb{R})).$$

2.1.2 Categories in Probability

Measurable spaces and measurable functions form a symmetric monoidal category:

Definition 2.12. *The objects in* **Meas** *are pairs (A, Σ_A), where Σ_A is a σ-algebra over A, and morphisms are measurable functions. The tensor product of the measurable spaces (A, Σ_A) and (B, Σ_B) in* **Meas** *is the product measurable space $(A \times B, \Sigma_A \times \Sigma_B)$ (Definition 2.4) and the tensor product of the measurable functions g, f is the function $(g \otimes f)(x, y) = (g(x), f(y))$.*

A notable subcategory of **Meas** is the following:

Definition 2.13. *The category* **BorelMeas** *is the symmetric monoidal subcategory of* **Meas** *in which objects are limited to* ***standard Borel measurable spaces****.*

Since standard Borel spaces are closed under countable products this subcategory is symmetric monoidal.

Proposition 1. *We can form a strict symmetric monoidal subcategory* **EucMeas** *of* **BorelMeas** *in which objects are restricted to be $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ for some $n \in \mathbb{N}$, the tensor of the objects $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ is $(\mathbb{R}^{n+m}, \mathcal{B}(\mathbb{R}^{n+m}))$ and morphisms are restricted to be infinitely differentiable.*

Proof. We first need to show that **EucMeas** is a symmetric monoidal subcategory of **Meas**. **EucMeas** contains all identities since the identity function $f(x_n) = x_n$ is infinitely differentiable for all n. Next, given two infinitely differentiable measurable maps g, f the composition $g \circ f$ and tensor $(g \otimes f)(x, y) = (g(x), f(y))$ are infinitely differentiable and measurable as well. Therefore morphisms in **EucMeas** are closed under composition and tensor. Next, since the tensor product of the standard Borel spaces $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ in **Meas** is the standard Borel space $(\mathbb{R}^{n+m}, \mathcal{B}(\mathbb{R}^{n+m}))$ we have that objects in **EucMeas** are closed under tensor as well.

Next, in order to show that **EucMeas** is strict monoidal we need to show that the associators and unitors in **EucMeas** are identities. First, since:

$$(\mathbb{R}^{n+(m+k)}, \mathcal{B}(\mathbb{R}^{n+(m+k)})) = (\mathbb{R}^{(n+m)+k}, \mathcal{B}(\mathbb{R}^{(n+m)+k}))$$

the associators in **EucMeas** are identities. Next, since:

$$(\mathbb{R}^{0+n}, \mathcal{B}(\mathbb{R}^{0+n})) = (\mathbb{R}^{n+0}, \mathcal{B}(\mathbb{R}^{n+0})) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$$

the unitors in **EucMeas** are identities.

□

Definition 2.14. [18; 15] *In the symmetric monoidal category* **Stoch** *objects are measurable spaces and morphisms are Markov kernels. 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 $\sigma_c \in \Sigma_C$:*

$$(\mu' \circ \mu) : A \times \Sigma_C \to [0,1]$$
$$(\mu' \circ \mu)(x_a, \sigma_c) = \int_{x_b \in B} \mu'(x_b, \sigma_c) d\mu(x_a, \_).$$

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

$$\delta : A \times \Sigma_A \to [0,1]$$
$$\delta(x_a, \sigma_a) = \begin{cases} 1 & x_a \in \sigma_a \\ 0 & x_a \notin \sigma_a \end{cases}.$$

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:*

$$(\mu' \otimes \mu) : (A \times C) \times (\Sigma_B \times \Sigma_D) \to [0,1]$$

where $\Sigma_B \times \Sigma_D$ *is the product sigma-algebra and for* $x_a \in A, x_c \in C, \sigma_b \in \Sigma_B, \sigma_d \in \Sigma_D$*:*

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

The objects in **Stoch** are also equipped with a commutative comonoidal structure that is compatible with the monoidal product in **Stoch**. That is, each metric space $X \in$ **Stoch** is equipped with a comultiplication map $\mathsf{cp} : X \to X \otimes X$ and a counit map $\mathsf{del} : X \to 1$ that satisfy the commutative comonoid equations, naturality of del and:

$$\mathsf{cp}_{X \otimes Y} = (id_X \otimes \sigma_{Y,X} \otimes id_Y)(\mathsf{cp}_X \otimes \mathsf{cp}_Y),$$

where $\sigma_{Y,X} : X \times Y \to Y \times X$ is the symmetric monoidal swap map in **Stoch**.

Stoch has many notable subcategories. For example, if we limit to countably generated measurable spaces as objects and Markov kernels over perfect probability measures as morphisms we get the following category:

Definition 2.15. *The category* **CGStoch** *is the subcategory of* **Stoch** *in which objects are limited to countably generated measurable spaces.*

If we add an additional condition of separability we form the following:

Definition 2.16. *The category* **BorelStoch** *is the subcategory of* **Stoch** *in which objects are limited to standard Borel spaces (the Borel spaces associated with Polish spaces).*

Limiting further to finite spaces yields the following category:

Definition 2.17. *The category* **FinStoch** *is the subcategory of* **Stoch** *in which objects are limited to finite measurable spaces.*

2.2 Parameterized Morphisms

A categorical tool that has risen in prominence to represent parameters in machine learning models is the **Para** operator. This operator has several presentations [4; 10; 13; 6]. A simple definition is as follows:

Definition 2.18. *Let* **C** *be a strict symmetric monoidal category. Then* **Para**(**C**) *is a category with the same objects as* **C**. *A morphism* $A \to B$ *in* **Para**(**C**) *is a pair* (P, f) *where* P *is an object of* **C** *and:*

$$f : P \otimes A \to B$$

is a morphism in **C**. *The composition of morphisms:*

$$f : P \otimes A \to B \qquad g : Q \otimes B \to C$$

is given by $(Q \otimes P, g \circ (id_Q \otimes f))$*:*

$$(Q \otimes P) \otimes A \xrightarrow{=} Q \otimes (P \otimes A) \xrightarrow{id_Q \otimes f} Q \otimes B \xrightarrow{g} C$$

3 Random Variables and Independence

3.1 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 rvs [14; 11; 12].

Since **BorelStoch** (Definition 2.16) is the Kleisli category of the restriction of the Giry monad [15] over **BorelMeas** (Definition 2.13), we can define an embedding functor from **BorelMeas** 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 rvs 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 rvs defined on this space f, f'. We can think of these rvs as morphisms in **BorelMeas** from (Ω, Σ) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We can represent this probability space as a morphism in **BorelStoch** between the monoidal unit $*$ and (Ω, Σ): that is, a Markov kernel $\mu : * \times \Sigma \to [0, 1]$. Going forward we will write the type signature $* \times \Sigma \to [0, 1]$ 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 rvs 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:

$$(\delta_f \otimes \delta_{f'}) \circ \mathsf{cp} \circ \mu : \mathcal{B}(\mathbb{R} \times \mathbb{R}) \to [0, 1]$$

where $\mathsf{cp} : \Omega \to \Omega \otimes \Omega$ is the comonoidal copy map at Ω in **BorelStoch** [12]. We can see that for $\sigma \times \sigma' \in \mathcal{B}(\mathbb{R} \times \mathbb{R})$:

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

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 rvs f and f' over (Ω, Σ, μ).

Therefore, the rvs f and f' are independent over the probability space (Ω, Σ, μ) iff the probability measures $(\delta_f \circ \mu) \otimes (\delta_{f'} \circ \mu)$ and $(\delta_f \otimes \delta_f{}')\circ \mathsf{cp} \circ \mu$ are equal.

4 The co-Kleisli Construction

Fong et al. and Cruttwell et al. [10; 6] build their characterization of machine learning optimization problems on top of the category of Euclidean spaces and parameterized infinitely 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 infinitely differentiable maps in Fong et al.'s [10] construction with parameterized random variables.

A quick note on the category of Euclidean spaces and parameterized infinitely differentiable maps between them: Fong et al. [10] calls this category **Para** whereas Cruttwell et al. [6] call it **Para**(**Euc**) (Definition 2.18). We will use Cruttwell et al.'s [6] notation, but we will work with the category **EucMeas** (Proposition 1) instead of **Euc** to make it easier to talk about probabilistic constructions.

Before moving to **Para**(**EucMeas**), we will start with the category **EucMeas** (Proposition 1) of Euclidean spaces and infinitely differentiable maps between them. Our first step will be to replace the morphisms in **EucMeas** with stochastic processes, or indexed families of random variables.

To start, note that $(O \times \_) : \mathbf{C} \to \mathbf{C}$ is an endofunctor that maps the object $A \in Ob(\mathbf{C})$ to to $O \times A$ and maps the morphism $f : A \to B$ to the morphism $id_O \otimes f : O \times A \to O \times B$. We can now introduce the following definition:

Definition 4.1. *For some Cartesian monoidal category* $\mathbf{C}$ *and object* O *in* $\mathbf{C}$*,* $\mathbf{CoKl}_O(\mathbf{C})$ *is the co-Kleisli category of* $\mathbf{C}$ *under the product comonad* $(O \times \_)$*.*

The category $\mathbf{CoKl}_O(\mathbf{C})$ has the same objects as $\mathbf{C}$ and the morphisms in $\mathbf{CoKl}_O(\mathbf{C})[A, B]$ are morphisms in $\mathbf{C}[O \times A, B]$. The identity map at the object A is:

$$(\mathsf{del}_O \otimes id_A) : O \times A \to A$$

where $\mathsf{del}_O : O \to *$ is the unique map from O to the terminal object $*$. The $\mathbf{CoKl}_O(\mathbf{C})$-composition of the morphisms $f : O \times A \to B$ and $g : O \times B \to C$ is:

$$(g \circ_{\mathbf{CoKl}_O(\mathbf{C})} f) : O \times A \to C$$
$$g \circ_{\mathbf{CoKl}_O(\mathbf{C})} f = g \circ_{\mathbf{C}} (id_O \otimes f) \circ_{\mathbf{C}} (\mathsf{cp}_O \otimes id_A)$$

where $\mathsf{cp}_O : O \to O \times O$ is the copy map (aka diagonal) in $\mathbf{C}$.

For example, if Ω is $\mathbb{R}^n$ for some $n \in \mathbb{N}$, the category $\mathbf{CoKl}_{(\Omega, \mathcal{B}(\Omega))}(\mathbf{EucMeas})$ (which we will hereafter abbreviate **CEucMeas**, see Table 4.1) has the same objects as **EucMeas**, 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 **CEucMeas**, the identity arrow at $\mathbb{R}^a$ is the function $f(\omega, x_a) = x_a$ and 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 $\omega \in \Omega, x_a \in \mathbb{R}^a$:

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

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

Proposition 2. *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* **CEucMeas** *to the function* $f(\omega, \_) : \mathbb{R}^a \to \mathbb{R}^b$ *in* **EucMeas** *is a functor* $R_\omega :$ **CEucMeas** $\to$ **EucMeas***, which we call the realization functor.*

Proof. First, if f is the identity map in **CEucMeas** 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 **CEucMeas** and any $x_a \in \mathbb{R}^a$. Then:

$$(R_\omega f' \circ R_\omega f) : \mathbb{R}^a \to \mathbb{R}^c$$
$$(R_\omega f' \circ R_\omega f)(x_a) = (f'(\omega, \_) \circ f(\omega, \_))(x_a) = f'(\omega, f(\omega, x_a)) = R_\omega(f' \circ f)(x_a)$$

so composition is preserved. □

Given a probability measure $\mu : \mathcal{B}(\Omega) \to [0,1]$, we can think of **CEucMeas** as a category of differentiable stochastic processes defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$.

Definition 4.2. *A **Levy Process** is a 1-dim stochastic process $f : \Omega \times \mathbb{R} \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ such that:*

- *$f(\ ,0) = 0$ a.s.*
  - *For $t_d > t_c > t_b > t_a \in \mathbb{R}$, the rvs $f(\ ,t_b) - f(\ ,t_a)$ and $f(\ ,t_d) - f(\ ,t_c)$ are independent.*
  - *For $t_b > t_a \in \mathbb{R}$, the rvs $f(\ ,t_b) - f(\ ,t_a)$ and $f(\ ,t_b - t_a)$ have the same distribution.*
- *For any $\omega \in \Omega$ the function $f(\omega,\ )$ is continuous.*

We can view Levy processes as continuous-time generalizations of random walks, or as Brownian motions with drift.

Definition 4.3. *A subordinator is a non-decreasing Levy Process. That is, if f is a subordinator then for any fixed $\omega \in \Omega$ the function $f(\omega,\ )$ is non-decreasing.*

Since subordinators are closed under composition we can show the following:

Proposition 3. *Continuously differentiable subordinators form a single-object subcategory of* **CEucMeas** *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 et al. [17] 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.

□

4.1 Independence and Dependence in **CEucMeas**

Since all of the stochastic processes in **CEucMeas** are defined over the same probability space $(\Omega, \mathcal{B}(\Omega), \mu)$, there is a major difference between how **CEucMeas** 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 **CEucMeas** and the vectors $x_a \in \mathbb{R}^a, x_c \in \mathbb{R}^c$, the rvs $f(\ ,x_a)$ 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 **CEucMeas** to **BorelStoch** such that for $x_a \in \mathbb{R}^a, \sigma_b \in \mathcal{B}(\mathbb{R}^b)$:

$$f_*\mu : \mathbb{R}^a \times \mathcal{B}(\mathbb{R}^b) \to [0,1]$$

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

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

$$(f' \circ f)_* \mu : \mathbb{R}^a \times \mathcal{B}(\mathbb{R}^c) \to [0,1]$$

$$
\begin{aligned}
(f' \circ f)_* \, \mu(x_a, \sigma_c) &= \\
\mu((f' \circ f)(\_, x_a)^{-1}(\sigma_c)) &= \\
\int_{\omega \in \Omega} \delta((f'(\omega, f(\omega, x_a)), \sigma_c) d\mu &= \\
\int_{\omega \in \Omega} \left(\int_{x_b \in \mathbb{R}^b} \delta(f'(\omega, x_b), \sigma_c) d\delta(f(\omega, x_a), \_) \right) d\mu &= \\
\int_{x_b \in \mathbb{R}^b} \int_{\omega \in \Omega} \delta(f'(\omega, x_b), \sigma_c) \; d\delta(f(\omega, x_a), \_) \; d\mu &
\end{aligned}
$$

whereas:

$$[f'_* \mu \circ f_* \mu] : \mathbb{R}^a \times \mathcal{B}(\mathbb{R}^c) \to [0, 1]$$

$$
\begin{aligned}
[f'_* \mu \circ f_* \mu] \, (x_a, \sigma_c) &= \\
\int_{x_b \in \mathbb{R}^b} \mu(f'(\_, x_b)^{-1}(\sigma_c)) \; d\mu(f(\_, x_a)^{-1}(\_)) &= \\
\int_{x_b \in \mathbb{R}^b} \left(\int_{\omega \in \Omega} \delta(f'(\omega, x_b), \sigma_c) d\mu \right) \left(\int_{\omega \in \Omega} d\delta(f(\omega, x_a), \_) d\mu \right) &.
\end{aligned}
$$

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

The reason for this mismatch comes down to the fact that composition in **BorelStoch** is based on the Markov property, whereas composition in **CEucMeas** is not. In the next Section we will define a new category of stochastic processes that exhibits this independence behavior.

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

Table 1: We introduce several compositional constructions for building new categories in this section. These can produce unwieldy names, so for readability we have abbreviated some of them here.

5 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, over the next few Sections we build a new category of stochastic processes in which the map $f \to f_* \mu$ is functorial. In Section 6.3 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 **CEucMeas**, 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 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 [19; 2; 1].

5.1 An Subcategory of $\mathbf{Para}(\mathbf{C})$

Consider the small symmetric strict monoidal categories $\mathbf{C}$ and $\mathbf{D}$ such that there exists a faithful identity-on-objects strict monoidal functor $\iota : \mathbf{D} \hookrightarrow \mathbf{C}$. That is, we can think of $\mathbf{D}$ as a subcategory of $\mathbf{C}$.

Proposition 4. *For the small strict symmetric monoidal categories $\mathbf{C}$ and $\mathbf{D}$ equipped with a faithful identity-on-objects strict monoidal functor $\iota : \mathbf{D} \hookrightarrow \mathbf{C}$ we can form a subcategory $\mathbf{Para_D}(\mathbf{C})$ of $\mathbf{Para}(\mathbf{C})$ (Definition 2.18) in which the morphisms in $\mathbf{Para_D}(\mathbf{C})[A, B]$ are pairs (P, f) where P is an object in the image of ι and $f : P \otimes A \to B$ is a morphism in $\mathbf{C}$.*

Proof. In order to show that $\mathbf{Para_D}(\mathbf{C})$ is a subcategory of $\mathbf{Para}(\mathbf{C})$ we simply need to show that $\mathbf{Para_D}(\mathbf{C})$ is closed under composition and contains all identities.

To start, note that $\mathbf{Para_D}(\mathbf{C})$ contains all identities. Since $\mathbf{D}$ is a strict monoidal category we can write the signature of the identity arrow $id_A : A \to A$ in $\mathbf{D}$ as $id_A : *_{\mathbf{D}} \otimes A \to A$ where $*_{\mathbf{D}}$ is the monoidal unit in $\mathbf{D}$. Therefore id_A is an arrow in $\mathbf{Para_D}(\mathbf{C})$. This arrow is trivially the $\mathbf{Para_D}(\mathbf{C})$-identity at A.

Next, note that $\mathbf{Para_D}(\mathbf{C})$ is closed under composition. Consider the morphisms:

$$f_1 : P_1 \otimes A \to B \qquad f_2 : P_2 \otimes A \to B$$

where P_1, P_2 are objects in the image of ι. Since ι is identity on objects and strict monoidal it must be that $P_1 \otimes P_2$ is an object in $\mathbf{D}$. Therefore:

$$f_2 \circ_{\mathbf{Para}(\mathbf{C})} f_1 : P_2 \otimes P_1 \otimes A \to B$$

is a morphism in $\mathbf{Para_D}(\mathbf{C})$. □

5.2 A Category of Parametric Measurable Maps

In this Section, we will use the $\mathbf{Para_D}$ construction (Proposition 4) to build a new category of stochastic processes over which the mapping $f \to f_*\mu$ is functorial. In this category composition will have the same independence structure that it has in $\mathbf{Stoch}$.

5.2.1 Lawvere Parameterization

We begin with the following definition:

Definition 5.1. *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 strict monoidal 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 $O \times O \times \cdots \times O$. When the tensor is repeated n times we will write this as O^n. We also write O^0 for the monoidal unit $*$. 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 $\mathbf{CoKl}_O(\mathbf{C})$:

$$f \circ_{\mathbf{C}} (\mathsf{cp}_O(n) \otimes id_A^{\mathbf{C}}) : O \times A \to B.$$

Where $id_A^{\mathbf{C}}$ is the identity arrow on A in $\mathbf{C}$ and $\mathsf{cp}_O(n)$ is the $n-1$ repeated application of the copy map $O \to O \times O$ in $\mathbf{C}$. That is:

- $\mathsf{cp}_O(3) : O \to O \times O \times O$ is the double application of the copy map $(id_O \otimes \mathsf{cp}_O) \circ \mathsf{cp}_O$.
- $\mathsf{cp}_O(2) : O \to O \times O$ is just the copy map cp_O.
- $\mathsf{cp}_O(1) : O \to O$ is the identity map id_O in $\mathbf{C}$.

- $\mathsf{cp}_O(0) : O \to *$ is del_O, the unique map from O to the terminal object $*$.

,

Proposition 5. *Suppose* $\mathbf{C}$ *is a Cartesian monoidal category and* O *is an object in* $\mathbf{C}$*. Then* $Copy : \mathbf{Para}_{O^*}(\mathbf{C}) \to \mathbf{CoKl}_O(\mathbf{C})$ *is a full identity-on-objects 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 $\mathbf{CoKl}_O(\mathbf{C})$ (Definition 4.1). Then f is also an arrow in $\mathbf{Para}_{O^*}(\mathbf{C})$ and *Copy* maps f to:

$$f \circ_{\mathbf{C}} (\mathsf{cp}_O(1) \otimes id_A^{\mathbf{C}}) = f.$$

Therefore *Copy* is full.

Next, since $\mathbf{C}$ is strict monoidal we have

$$A = * \times A = O^0 \times A$$

which implies that *Copy* maps the $id_A : A \to A$ arrow in $\mathbf{Para}_{O^*}(\mathbf{C})$ to the arrow:

$$id_A \circ_{\mathbf{C}} (\mathsf{cp}_O(0) \otimes id_A^{\mathbf{C}}) : O \times A \to A$$

which is the identity arrow in $\mathbf{CoKl}_O(\mathbf{C})$. 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 $\mathbf{Para}_{O^*}(\mathbf{C})$:

$$(Copy f' \circ Copy f) : O \otimes A \to C$$

$$\begin{aligned}
(Copy f' \circ_{\mathbf{CoKl}_O(\mathbf{C})} Copy f) &= \\
\big(f' \circ_{\mathbf{C}} (\mathsf{cp}_O(n) \otimes id_B^{\mathbf{C}})\big) \circ_{\mathbf{CoKl}_O(\mathbf{C})} \big(f \circ_{\mathbf{C}} (\mathsf{cp}_O(m) \otimes id_A^{\mathbf{C}})\big) &= \\
\big(f' \circ_{\mathbf{C}} (\mathsf{cp}_O(n) \otimes id_B^{\mathbf{C}})\big) \circ_{\mathbf{C}} (id_O^{\mathbf{C}} \otimes (f \circ_{\mathbf{C}} (\mathsf{cp}_O(m) \otimes id_A^{\mathbf{C}}))) \circ_{\mathbf{C}} (\mathsf{cp}_O \otimes id_A^{\mathbf{C}}) &= \\
f' \circ_{\mathbf{C}} (\mathsf{cp}_O(n) \otimes (f \circ_{\mathbf{C}} (\mathsf{cp}_O(m) \otimes id_A^{\mathbf{C}}))) \circ_{\mathbf{C}} (\mathsf{cp}_O \otimes id_A^{\mathbf{C}}) &= \\
f' \circ_{\mathbf{C}} (id_{O^n} \otimes (f \circ_{\mathbf{C}} (\mathsf{cp}_O(m) \otimes id_A^{\mathbf{C}}))) \circ_{\mathbf{C}} (\mathsf{cp}_O(n+1) \otimes id_A^{\mathbf{C}}) &= \\
f' \circ_{\mathbf{C}} (id_{O^n} \otimes f) \circ_{\mathbf{C}} (\mathsf{cp}_O(n+m) \otimes id_A^{\mathbf{C}}) &= \\
(f' \circ_{\mathbf{Para}_{O^*}(\mathbf{C})} f) \circ_{\mathbf{C}} (\mathsf{cp}_O(n+m) \otimes id_A^{\mathbf{C}}) &= \\
Copy(f' \circ_{\mathbf{Para}_{O^*}(\mathbf{C})} f).
\end{aligned}$$

□

5.3 Applying Para to EucMeas

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 strict monoidal functor as the inclusion:

$$\iota : (\Omega, \mathcal{B}(\Omega))^* \hookrightarrow \mathbf{EucMeas}$$

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:

$$\mu^n : \mathcal{B}(\Omega^n) \to [0,1]$$
$$\mu^n(\sigma_1 \times \sigma_2 \times \cdots \times \sigma_n) = \mu(\sigma_1)\mu(\sigma_2)\cdots\mu(\sigma_n).$$

Definition 5.2. *We can apply* $\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}$ *to* **EucMeas** *to form the Lawvere parameterization* $\mathbf{Para}_{(\Omega,\mathcal{B}(\Omega))^*}(\mathbf{EucMeas})$*, which we will hereafter abbreviate* **PEuc***.*

Intuitively, **PEuc** allows us to reason about probabilistic relationships in terms of measurable functions rather than probability measures.

Next, by Proposition 5, we have an identity-on-objects functor, $Copy$, from **PEuc** to **CEucMeas**. 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 **CEucMeas**, if we compose 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 $f(\_, x_a)_* \mu^n(\_) : \mathcal{B}(\mathbb{R}) \to [0, 1]$ 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 functor from* **PEuc** *to* **BorelStoch***.*

Proof. In this proof we rely on the following property of product measures, which holds by Fubini's theorem when $\Omega = \mathbb{R}^k$:

$$\int_{(\omega_n, \omega_m) \in \Omega^n \times \Omega^m} f(\omega_n, \omega_m) d\mu^{n+m} = \int_{\omega_n \in \Omega^n} \int_{\omega_m \in \Omega^m} f(\omega_n, \omega_m) d\mu^n d\mu^m$$

We first note that since the objects in $Ob(\mathbf{PEuc}) = Ob(\mathbf{EucMeas})$ are the standard Borel measurable spaces $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ it must be that $Push_\mu$ maps objects in **PEuc** to objects in **BorelStoch**.

Next, note that for any $\mathbb{R}^a$, $Push_\mu$ maps the identity at $\mathbb{R}^a$ in **PEuc** to the identity at $\mathbb{R}^a$ in **BorelStoch** since:

$$\begin{aligned} Push_\mu id_{\mathbb{R}^a}(x_a, \sigma_a) &= \\ \int_{\omega_n \in \Omega^n} \delta(id_{\mathbb{R}^a}(\omega_n, x_a), \sigma_a) d\mu^n &= \\ \int_{\omega_n \in \Omega^n} \delta(x_a, \sigma_a) d\mu^n &= \\ \delta(x_a, \sigma_a) & \end{aligned}$$

Next, we will demonstrate that $Push_\mu$ preserves composition. Suppose we have some:

$$\begin{matrix} f : \Omega^n \times \mathbb{R}^a \to \mathbb{R}^b & f' : \Omega^m \times \mathbb{R}^b \to \mathbb{R}^c \\ x_a \in \mathbb{R}^a & \sigma_c \in \mathcal{B}(\mathbb{R}^c) \end{matrix}$$

Then we can write:

$$Push_\mu (f' \circ f) : \mathbb{R}^a \times \mathcal{B}(\mathbb{R}^c) \to [0, 1]$$

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

□

5.4 Composition Experiments

We can express the difference between composition in **PEuc** and **CEucMeas** with a simple experiment using the numpy [16] and scipy [20] libraries.

Consider the probability space $(\Omega,\mathcal{B}(\Omega),\mu)$ where $\Omega=[0,1]$ and μ is the uniform measure. We can represent samples from this with the following:

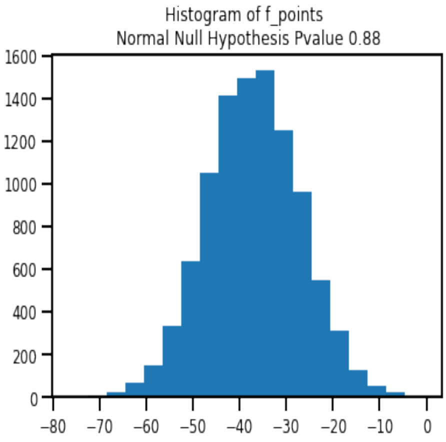
```
import numpy as np
omega_samples = np.random.random(10000)
```

Now consider the following stochastic process $f:\Omega\times\mathbb{R}\to\mathbb{R}$ over $(\Omega,\mathcal{B}(\Omega),\mu)$:

```
from scipy.stats import norm
def f(omega, x):
    normal_points = norm(loc=5 - x, scale=10).ppf(omega)
    return normal_points
```

Note that for any $x\in\mathbb{R}$, the random variable $f(\_,x)$ over has a normal distribution:

```
input_x = 42
f_points = f(omega_samples, input_x)
```

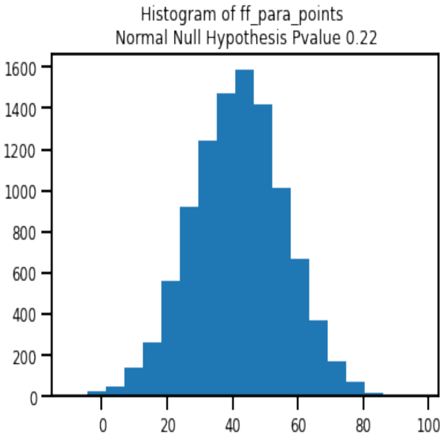


Note that f is an endomorphism on $\mathbb{R}$ in **PEuc**, so we can take the **PEuc**-composition of f with itself to form the arrow $(f \circ f) : \Omega^2 \times \mathbb{R} \to \mathbb{R}$. We write this arrow as:

```
def ff_para(omega1, omega2, x):
    return f(omega2, f(omega1, x))
```

Note that $(f \circ f)$ is a stochastic process over the product probability space $(\Omega^2, \mathcal{B}(\Omega^2), \mu^2)$, and that for any $x \in \mathbb{R}$, the random variable $(f \circ f)(\_, \_, x)$ over $(\Omega^2, \mathcal{B}(\Omega^2), \mu^2)$ has a normal distribution as well:

```
input_x = 42
omega2 = np.random.random((10000, 2))
ff_para_points = ff_para(omega2[:, 0], omega2[:, 1], input_x)
```



Now recall the functor *Copy* : **PEuc** → **CEucMeas** from Proposition 5. This functor acts as identity-on-objects and sends the arrow $f : \Omega^n \times \mathbb{R} \to \mathbb{R}$ in **PEuc** to the following arrow in **CEucMeas**:

$$f \circ_{\mathbf{EucMeas}} (\mathsf{cp}_{\Omega}(n) \otimes id_{\mathbb{R}}) : \Omega \times \mathbb{R} \to \mathbb{R}.$$

We can implement this functor as follows:

```
import inspect
from functools import partial
def CopyFunctor(f):
    def g(omega, x, f=f):
        while len(inspect.getargspec(f).args) > 1:
            f = partial(f, omega)
        return f(x)
    return g
```

Note that:

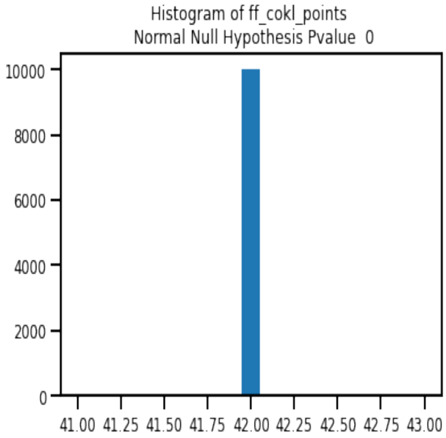
$$Copy(f \circ f) : \Omega \times \mathbb{R} \to \mathbb{R}$$

is a stochastic process over the probability space $(\Omega, \mathcal{B}(\Omega), \mu)$. However, unlike $(f \circ f)(\_, \_, x)$, the random variable:

$$Copy(f \circ f)(\_, x) : \Omega \to \mathbb{R}$$

is not normal for any fixed $x \in \mathbb{R}$. Instead, it is constant:

```
input_x = 42
omega = np.random.random(10000)
ff_cokl_points = CopyFunctor(ff_para)(omega, input_x)
```



6 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 **EucMeas** 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 [10] Backpropagation functor 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**(**EucMeas**) 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).

6.1 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 mean 0 and 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.

Proposition 7. *Suppose we have a probability space $(\Omega, \mathcal{B}(\Omega), \mu)$ where Ω is $\mathbb{R}^k, k \in \mathbb{N}$. We can define a category* **DF** *that has the same objects as* **EucMeas** *(Definition 1) such that the morphisms between $\mathbb{R}^a$ and $\mathbb{R}^b$ are* **EucMeas***-morphisms of the form:*

$$f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$$

The composition of the morphisms:

$$f_1 : \Omega^{n_1} \times \mathbb{R}^{p_1} \times \mathbb{R}^a \to \mathbb{R}^b \qquad f_2 : \Omega^{n_2} \times \mathbb{R}^{p_2} \times \mathbb{R}^b \to \mathbb{R}^c$$

is the morphism:

$$f_2 \circ f_1 : \Omega^{n_2+n_1} \times \mathbb{R}^{p_2+p_1} \times \mathbb{R}^a \to \mathbb{R}^c$$
$$(f_2 \circ f_1)(\omega_{n_2}, \omega_{n_1}, x_{p_2}, x_{p_1}, x_a) = f_2(\omega_{n_2}, x_{p_2}, f_1(\omega_{n_1}, x_{p_1}, x_a))$$

Proof. We need to show that **DF** contains all identities, is closed under composition, and that composition is associative.

To start, note that the identity arrow at the object $\mathbb{R}^a$ in **DF** is the function:

$$id : \Omega^0 \times \mathbb{R}^0 \times \mathbb{R}^a \to \mathbb{R}^a$$
$$id(x_a) = x_a$$

and therefore **DF** contains all identities.

Next, note that the composition of the morphisms:

$$f_1 : \Omega^{n_1} \times \mathbb{R}^{p_1} \times \mathbb{R}^a \to \mathbb{R}^b \qquad f_2 : \Omega^{n_2} \times \mathbb{R}^{p_2} \times \mathbb{R}^b \to \mathbb{R}^c$$
$$f_2 \circ f_1 : \Omega^{n_2+n_1} \times \mathbb{R}^{p_2+p_1} \times \mathbb{R}^a \to \mathbb{R}^c$$

is in $\mathbf{DF}[\mathbb{R}^a, \mathbb{R}^c]$ and therefore **DF** is closed under composition.

Next, consider the morphisms

$$f_1 : \Omega^{n_1} \times \mathbb{R}^{p_1} \times \mathbb{R}^a \to \mathbb{R}^b \qquad f_2 : \Omega^{n_2} \times \mathbb{R}^{p_2} \times \mathbb{R}^b \to \mathbb{R}^c \qquad f_3 : \Omega^{n_3} \times \mathbb{R}^{p_3} \times \mathbb{R}^c \to \mathbb{R}^d$$

We have that:

$$f_3 \circ (f_2 \circ f_1) : \Omega^{n_3+(n_2+n_1)} \times \mathbb{R}^{p_3+(p_2+p_1)} \times \mathbb{R}^a \to \mathbb{R}^d$$

$$\begin{aligned}(f_3 \circ (f_2 \circ f_1))((\omega_{n_3}, (\omega_{n_2}, \omega_{n_1})), (x_{p_3}, (x_{p_2}, x_{p_1})), x_a) &= \\ f_3(\omega_{n_3}, x_{p_3}, (f_2 \circ f_1)((\omega_{n_2}, \omega_{n_1}), (x_{p_2}, x_{p_1}), x_a)) &= \\ f_3(\omega_{n_3}, x_{p_3}, f_2(\omega_{n_2}, x_{p_2}, f_1(\omega_{n_1}, x_{p_1}, x_a)))\end{aligned}$$

which is equal to:

$$(f_3 \circ f_2) \circ f_1 : \Omega^{(n_3+n_2)+n_1} \times \mathbb{R}^{(p_3+p_2)+p_1} \times \mathbb{R}^a \to \mathbb{R}^d$$

$$\begin{aligned}((f_3 \circ f_2) \circ f_1)(((\omega_{n_3}, \omega_{n_2}), \omega_{n_1}), ((x_{p_3}, x_{p_2}), x_{p_1}), x_a) &= \\ (f_3 \circ f_2)((\omega_{n_3}, \omega_{n_2}), (x_{p_3}, x_{p_2}), f_1(\omega_{n_1}, x_{p_1}, x_a)) &= \\ f_3(\omega_{n_3}, x_{p_3}, f_2(\omega_{n_2}, x_{p_2}, f_1(\omega_{n_1}, x_{p_1}, x_a)))\end{aligned}$$

and therefore composition in **DF** is associative. □

The name **DF** derives from the fact that the arrows in this category are **D**iscriminative and **F**requentist statistical models (see Table 1 for a list of all such abbreviations). 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.

6.2 Gaussian-Preserving Transformations

6.2.1 A Subcategory of Gaussian-Preserving Transformations

Definition 6.1. *A Gaussian-preserving transformation* $T : \mathbb{R}^a \to \mathbb{R}^b$ *is a Borel measurable function such that for any multivariate normal random variable* $f : \Omega^n \to \mathbb{R}^a$ *defined on the probability space* $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$*, the random variable* $(T \circ f) : \Omega^n \to \mathbb{R}^b$ *is multivariate normal and we have:*

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

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:

$$f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$$

there exists some map $T : \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and multivariate normal random variable $G : \Omega^n \to \mathbb{R}^b$ defined on the probability space $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$ such that for all $\omega_n \in \Omega^n, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ the map $T(x_p, \_) : \mathbb{R}^a \to \mathbb{R}^b$ is a Gaussian-preserving transformation and:

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

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.

Since $\mathcal{N}_\mu$ contains the identity arrows we can construct a useful subcategory of **DF**.

Definition 6.2. $\mathbf{DF}_{\mathcal{N}_\mu}$ *is the category with the same objects as* **DF** *and arrows generated by the* **DF**-*composition of arrows 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* $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$, $f(\_, x_p, x_a) : \Omega^n \to \mathbb{R}^b$ *is a multivariate normal random variable defined on the probability space* $(\Omega^n, \mathcal{B}(\Omega^n), \mu^n)$.

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 following random variable:

$$f^l : \Omega^{m+n} \to \mathbb{R}^a$$
$$f^l(\omega_m, \omega_n) = f(\omega_m)$$

We can see this as follows. For any $\sigma_a \in \mathcal{B}(\mathbb{R}^a)$:

$$f^l_* \mu^{m+n} : \mathcal{B}(\mathbb{R}^a) \to [0, 1]$$

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

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:

$$f^r(\_, x_p, x_a) : \Omega^{m+n} \to \mathbb{R}^b$$
$$f^r((\omega_m, \omega_n), x_p, x_a) = f(\omega_n, x_p, x_a)$$

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:

$$G'^l : \Omega^{m+n} \to \mathbb{R}^b$$
$$G'^l(\omega_m, \omega_n) = G'(\omega_m)$$

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:

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

□

As an aside, note that $\mathcal{N}_\mu$ itself is not closed under composition. Suppose

$$f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$$
$$f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$$

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:

$$(f' \circ f) : \Omega^{m+n} \times \mathbb{R}^{q+p} \times \mathbb{R}^a \to \mathbb{R}^c$$
$$(f' \circ f)((\omega_m, \omega_n), (x_q, x_p), x_a) = \|x_q\|_1 T(x_p, x_a) + \|x_q\|_1 G(\omega_n) + G'(\omega_m),$$

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 multivariate normal random variable defined on $(\Omega^{n+m}, \mathcal{B}(\Omega^{n+m}), \mu^{n+m})$. $(f' \circ f)$ is therefore not in $\mathcal{N}_\mu$. However, for any choice of $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ the random variable:

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

is a linear function of multivariate normal random variables and is therefore itself multivariate normal.

6.2.2 Relationship to **Gauss**

$\mathbf{DF}_{\mathcal{N}_\mu}$ is similar to the category **Gauss** from Section 6 of Fritz [12], with a few key differences.

Definition 6.3. *In the category* **Gauss** *[12] 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, this arrow is dependent on the choice of $\xi_{s,C}$.

6.3 Expectation Composition

Definition 6.4. *A subcategory* **C** *of* **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* **C** *and $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:*

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

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

Proof. Consider some $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ in $\mathbf{DF}_{\mathcal{N}_\mu}$ and $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$. We will prove by induction that Definition 6.4 holds.

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 $x_{q''} \in \mathbb{R}^{q''}, x_{q'} \in \mathbb{R}^{q'}, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:

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

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

We can now define the following functor:

Proposition 10. *Suppose* $\mathbf{C} \subseteq \mathbf{DF}$ *is an Expectation Composition category. We can define a map* $Exp : \mathbf{C} \to \mathbf{Para}(\mathbf{EucMeas})$ *that acts as the identity on objects and sends the arrow* $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ *in* $\mathbf{C}$ *to the following function:*

$$f_E : \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$$
$$f_E(x_p, x_a) = E_{\mu^n}[f(\_, x_p, x_a)] = \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n.$$

$Exp : \mathbf{C} \to \mathbf{Para}(\mathbf{EucMeas})$ *is a functor.*

Proof. To start, note that Exp trivially sends objects in $\mathbf{C}$ to objects in $\mathbf{Para}(\mathbf{EucMeas})$. Next, note that for any morphism in $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ in $\mathbf{C}$ the Leibniz integration rule implies that the following function is differentiable and therefore also Borel measurable:

$$f_E : \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$$
$$f_E(x_p, x_a) = E_{\mu^n}[f(\_, x_p, x_a)] = \int_{\omega_n\in\Omega^n} f(\omega_n, x_p, x_a)d\mu^n.$$

Therefore Exp sends morphisms in $\mathbf{C}$ to morphisms in $\mathbf{Para}(\mathbf{EucMeas})$. Next, we can see that Exp preserves identities since $Exp(id)$ is the identity function in $\mathbf{Para}(\mathbf{EucMeas})$

$$Exp(id)(x_a) = E_{\mu^n}[id(\_, x_a)] = E_{\mu^n}[x_a] = x_a$$

Finally, consider the morphisms $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ and $f' : \Omega^m \times \mathbb{R}^q \times \mathbb{R}^b \to \mathbb{R}^c$ in $\mathbf{C}$. We have that for $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$:

$$\begin{aligned}
&Exp(f' \circ f)(x_q, x_p, x_a) = \\
&\int_{(\omega_m, \omega_n) \in \Omega^{m+n}} (f' \circ f)((\omega_m, \omega_n), (x_q, x_p), x_a) d\mu^{m+n} = \\
&\int_{(\omega_m, \omega_n) \in \Omega^{m+n}} f'(\omega_m, x_q, f(\omega_n, x_p, x_a)) d\mu^{m+n} =^* \\
&\int_{\omega_m \in \Omega^m} f' \left(\omega_m, x_q, \int_{\omega_n \in \Omega^n} f(\omega_n, x_p, x_a) d\mu^n \right) d\mu^m = \\
&Exp(f')(x_q, Exp(f)(x_p, x_a))
\end{aligned}$$

where the step marked with $*$ is by the definition of an Expectation Composition category. This implies that Exp preserves composition. □

7 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 [10] 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 a backpropagation functor for any subcategory in this class.

7.1 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 the probability measure:

$$f(\_, x_p, x_a)_* \mu^n : \mathcal{B}(\mathbb{R}^b) \to [0, 1]$$
$$f(\_, x_p, x_a)_* \mu^n = \mu^n (f(\_, x_p, x_a)^{-1})$$

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 $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ 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:

$$l : \Omega^n \times \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$$

that we introduced in Section 6.1 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 7.1. *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 a semicategory **CondLikelihood** of abstract conditional likelihoods.

Proposition 11. *We can define a semicategory* **CondLikelihood** *in which objects are spaces of the form $\mathbb{R}^n$ for some $n \in \mathbb{N}$ and 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 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:*

$$(L'_j \circ L_i) : \mathbb{R}^{q+p} \times \mathbb{R}^a \times \mathbb{R}^c \to \mathbb{R}$$
$$(L'_j \circ L_i)((x_q, x_p), x_a, x_c) = \int_{x_b \in \mathbb{R}^b} L'_j(x_q, x_b, x_c) L_i(x_p, x_a, x_b) dx_b.$$

for $L_i \in \mathbf{L}, L'_j \in \mathbf{L}'$.

Proof. We need to show that **CondLikelihood** is closed under composition and that composition in **CondLikelihood** is associative.

To start, note that if $L_i : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ and $L'_j : \mathbb{R}^q \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R}$ are abstract conditional likelihood functions then their composition is also an abstract conditional likelihood function since:

$$(L'_j \circ L_i) : \mathbb{R}^{q+p} \times \mathbb{R}^a \times \mathbb{R}^c \to \mathbb{R}$$
$$(L'_j \circ L_i)((x_q, x_p), x_a, x_c) = \int_{x_b \in \mathbb{R}^b} L'_j(x_q, x_b, x_c) L_i(x_p, x_a, x_b) dx_b.$$

is also Borel-measurable and Lebesgue integrable.

Next, we need to show that for any pair of equivalence classes:

$$\mathbf{L} : \mathbb{R}^a \to \mathbb{R}^b \qquad \mathbf{L}' : \mathbb{R}^b \to \mathbb{R}^c$$

and choice of representatives:

$$L_1 : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R} \qquad L_2 : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$$

in $\mathbf{L}$ and:

$$L'_1 : \mathbb{R}^q \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R} \qquad L'_2 : \mathbb{R}^q \times \mathbb{R}^b \times \mathbb{R}^c \to \mathbb{R}$$

in $\mathbf{L}'$ we have that for any $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ the functions:

$$(L'_1 \circ L_1)((x_q, x_p), x_a, \_) : \mathbb{R}^c \to \mathbb{R}$$
$$(L'_2 \circ L_2)((x_q, x_p), x_a, \_) : \mathbb{R}^c \to \mathbb{R}$$

are λ^c-a.e. equivalent.

Define σ_c to be the set of all $x_c \in \mathbb{R}^c$ where:

$$(L'_1 \circ L_1)((x_q, x_p), x_a, x_c) \neq (L'_2 \circ L_2)((x_q, x_p), x_a, x_c)$$

We need to show that σ_c has Lebesgue measure 0. For any $x_c \in \mathbb{R}^c$, define $\sigma_b(x_c)$ to be the union of the following subsets of $\mathbb{R}^b$:

$$\sigma_b(x_c) = \{x_b \mid L_1(x_p, x_a, x_b) \neq L_2(x_p, x_a, x_b)\} \cup \{x_b \mid L'_1(x_q, x_b, x_c) \neq L'_2(x_q, x_b, x_c)\}$$

Now for any x_c where the Lebesgue measure of $\sigma_b(x_c)$ is 0 we have the following:

$$\begin{aligned}
&(L'_1 \circ L_1)((x_q, x_p), x_a, x_c) = \\
&\int_{x_b \in \mathbb{R}^b} L'_1(x_q, x_b, x_c) L_1(x_p, x_a, x_b) dx_b = \\
\int_{x_b \in \mathbb{R}^b - \sigma_b(x_c)} L'_1(x_q, x_b, x_c) L_1(x_p, x_a, x_b) dx_b + &\int_{x_b \in \sigma_b(x_c)} L'_1(x_q, x_b, x_c) L_1(x_p, x_a, x_b) dx_b = \\
&\int_{x_b \in \mathbb{R}^b - \sigma_b(x_c)} L'_1(x_q, x_b, x_c) L_1(x_p, x_a, x_b) dx_b = \\
&\int_{x_b \in \mathbb{R}^b - \sigma_b(x_c)} L'_2(x_q, x_b, x_c) L_2(x_p, x_a, x_b) dx_b = \\
\int_{x_b \in \mathbb{R}^b - \sigma_b(x_c)} L'_2(x_q, x_b, x_c) L_2(x_p, x_a, x_b) dx_b + &\int_{x_b \in \sigma_b(x_c)} L'_2(x_q, x_b, x_c) L_2(x_p, x_a, x_b) dx_b = \\
&\int_{x_b \in \mathbb{R}^b} L'_2(x_q, x_b, x_c) L_2(x_p, x_a, x_b) dx_b = \\
&(L'_2 \circ L_2)((x_q, x_p), x_a, x_c)
\end{aligned}$$

Therefore, for any $x_c \in \sigma_c$ it must be that the Lebesgue measure of $\sigma_b(x_c)$ is greater than 0.

Since L_1, L_2 are representatives of the same equivalence class it must be that the set:

$$\{x_b \mid L_1(x_p, x_a, x_b) \neq L_2(x_p, x_a, x_b)\} \subseteq \mathbb{R}^b$$

always has Lebesgue measure 0, and therefore σ_c is equal to the set of all x_c for which the set:

$$\{x_b \mid L'_1(x_q, x_b, x_c) \neq L'_2(x_q, x_b, x_c)\} \subseteq \mathbb{R}^b$$

has Lebesgue measure greater than 0.

Now suppose for contradiction that the set σ_c has Lebesgue measure greater than 0. Then the set:

$$\{(x_b, x_c) \mid L'_1(x_q, x_b, x_c) \neq L'_2(x_q, x_b, x_c)\} \subseteq \mathbb{R}^{b+c}$$

must have Lebesgue measure greater than 0 as well. However, this is impossible since L'_1, L'_2 are representatives of the same equivalence class and therefore for any fixed $x_b \in \mathbb{R}^b$ the set:

$$\{x_c \mid L'_1(x_q, x_b, x_c) \neq L'_2(x_q, x_b, x_c)\} \subseteq \mathbb{R}^c$$

must have Lebesgue measure equal to 0. Therefore σ_c has Lebesgue measure equal to 0.

Therefore since σ_c has Lebesgue measure 0 we can conclude that:

$$\begin{gathered}
(L'_1 \circ L_1)((x_q, x_p), x_a, \_) : \mathbb{R}^c \to \mathbb{R} \\
(L'_2 \circ L_2)((x_q, x_p), x_a, \_) : \mathbb{R}^c \to \mathbb{R}
\end{gathered}$$

are λ^c-a.e. equivalent and **CondLikelihood** is closed under composition.

Next, we need to show that composition is associative. Suppose the following are representatives of three arrows in **CondLikelihood**:

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

Now consider the representatives of their composition:

$$\begin{gathered}
f_3 \circ (f_2 \circ f_1) : \mathbb{R}^a \to \mathbb{R}^d \\
(f_3 \circ f_2) \circ f_1 : \mathbb{R}^a \to \mathbb{R}^d
\end{gathered}$$

For $x_{p_3} \in \mathbb{R}^{p_3}, x_{p_2} \in \mathbb{R}^{p_2}, x_{p_1} \in \mathbb{R}^{p_1}, x_a \in \mathbb{R}^a, x_d \in \mathbb{R}^d$ we then have:

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

Therefore, composition in **CondLikelihood** is associative, so **CondLikelihood** is a semicategory. □

Note that **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 $x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a, x_b \in \mathbb{R}^b$:

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

If we extend from functions to generalized functions (distributions) we can form a category similar to **CondLikelihood**. For example, Blute [3] defines 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 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:

$$L_f : \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$$
$$L_f(x_p, x_a, x_b) = \frac{d\ f(\_, x_p, x_a)_* \mu^n}{d\lambda^b}(x_b)$$

That is, we have:

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

Proposition 12. $\mathbf{DF}_{\mathcal{R}_\mu}$ *is a semicategory.*

Proof. Since composition in $\mathbf{DF}_{\mathcal{R}_\mu}$ is the same as in **DF** we simply need to show that $\mathbf{DF}_{\mathcal{R}_\mu}$ is closed under **DF**-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 $x_q \in \mathbb{R}^q, x_p \in \mathbb{R}^p, x_a \in \mathbb{R}^a$ there exists some Borel-measurable and Lebesgue integrable $g : \mathbb{R}^c \to \mathbb{R}$ such that for $\sigma_c \in \mathcal{B}(\mathbb{R}^c)$:

$$(f' \circ f)(\_, (x_q, x_p), x_a)_* \mu^{m+n} : \mathcal{B}(\mathbb{R}^c) \to [0, 1]$$
$$(f' \circ f)(\_, (x_q, x_p), x_a)_* \mu^{m+n}(\sigma_c) = \int_{x_c \in \sigma_c} g(x_c) d\lambda^c$$

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

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

Therefore we have that:

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

and $L_{f'\circ f}$ is therefore Lebesgue integrable and Borel measurable since $L_{f'}$ and L_f are Lebesgue integrable and Borel measurable.

□

Proposition 13. *We can define a semifunctor $\mathcal{RN}_\mu : \mathbf{DF}_{\mathcal{R}_\mu} \to$ **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:*

$$
\begin{aligned}
L_f : \mathbb{R}^p\times\mathbb{R}^a\times\mathbb{R}^b\to\mathbb{R} \\
L_f(x_p,x_a,x_b) = \frac{df(\_,x_p,x_a)_*\mu^n}{d\lambda^b}(x_b).
\end{aligned}
$$

Proof. To start, note that $\mathcal{RN}_\mu$ maps objects in $\mathbf{DF}_{\mathcal{R}_\mu}$ to objects in **CondLikelihood** by definition. Next, Proposition 12 implies that for each morphism $f\in\mathbf{DF}_{\mathcal{R}_\mu}$ the function

$$
\begin{aligned}
L_f : \mathbb{R}^p\times\mathbb{R}^a\times\mathbb{R}^b\to\mathbb{R} \\
L_f(x_p,x_a,x_b) = \frac{df(\_,x_p,x_a)_*\mu^n}{d\lambda^b}(x_b).
\end{aligned}
$$

exists and therefore $\mathcal{RN}_\mu$ maps morphisms in $\mathbf{DF}_{\mathcal{R}_\mu}$ to morphisms in **CondLikelihood**.

Now we will show that $\mathcal{RN}_\mu$ preserves composition. Suppose

$$
\begin{aligned}
f : \Omega^n\times\mathbb{R}^p\times\mathbb{R}^a\to\mathbb{R}^b \\
f' : \Omega^m\times\mathbb{R}^q\times\mathbb{R}^b\to\mathbb{R}^c
\end{aligned}
$$

are arrows in $\mathbf{DF}_{\mathcal{R}_\mu}$. Then for any:

$$
\begin{aligned}
x_q\in\mathbb{R}^q \qquad x_p\in\mathbb{R}^p \\
x_a\in\mathbb{R}^a \qquad x_c\in\mathbb{R}^c
\end{aligned}
$$

we have:

$$\mathcal{RN}_{\mu}(f' \circ f) : \mathbb{R}^{q+p} \times \mathbb{R}^{a} \times \mathbb{R}^{c} \to \mathbb{R}$$

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

□

7.2 Maximum Likelihood

Suppose we have a probability space $(\mathbb{R}^a \times \mathbb{R}^b, \mathcal{B}(\mathbb{R}^a \times \mathbb{R}^b), \tau)$. 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 (note that log is a monotonic transformation and we just use it to make the math easier - the optimal value of x_p is the same with or without it):

$$L_{\tau} : \mathbb{R}^p \to \mathbb{R}$$

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

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.

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:

$$S_n = \{(x_{a_1}, x_{b_1}), (x_{a_2}, x_{b_2}), \cdots, (x_{a_n}, x_{b_n})\}$$

in $\mathbb{R}^a \times \mathbb{R}^b$.

Definition 7.2. *The maximum log likelihood estimator for f with respect to the samples:*

$$(x_{a_1}, x_{b_1}), (x_{a_2}, x_{b_2}), \cdots, (x_{a_n}, x_{b_n}) \in \mathbb{R}^a \times \mathbb{R}^b$$

is the vector $x_p \in \mathbb{R}^p$ that maximizes the function:

$$L_{S_n}(x_p) : \mathbb{R}^p \to \mathbb{R}$$

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

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 [10] 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 $j \leq b$, the jth component of $f : \Omega^n \times \mathbb{R}^p \times \mathbb{R}^a \to \mathbb{R}^b$ 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:

$$\begin{aligned} l_{ij} &: \mathbb{R}^p \to \mathbb{R} \\ l_{ij}(x_p) &= \frac{df(\_, x_p, x_{a_i})[j]_* \mu^n}{d\lambda}(x_{b_i}[j]) \end{aligned}$$

where we write $x_{b_i}[j] \in \mathbb{R}$ for the jth component of the vector $x_{b_i} \in \mathbb{R}^b$. We can now define the following:

Definition 7.3. *The maximum log-marginal likelihood estimator for f with respect to the samples:*

$$(x_{a_1}, x_{b_1}), (x_{a_2}, x_{b_2}), \cdots, (x_{a_n}, x_{b_n}) \in \mathbb{R}^a \times \mathbb{R}^b$$

is the vector $x_p \in \mathbb{R}^p$ that maximizes the function:

$$\begin{aligned} M_{S_n} &: \mathbb{R}^p \to \mathbb{R} \\ M_{S_n}(x_p) &= \sum_{i=1}^{n} \sum_{j=1}^{b} log\ l_{ij}(x_p). \end{aligned}$$

where $l_{ij}(x_p)$ is the marginal likelihood at $x_p \in \mathbb{R}^p$ of the jth component of f for $(x_{a_i}, x_{b_i}) \in S_n$.

Note that $M_{S_n}(x_p) = L_{S_n}(x_p)$ when the real-valued random variables:

$$f(\_, x_p, x_{a_i})[j] : \Omega^n \to \mathbb{R}$$

are mutually independent for all x_{a_i} and $j \leq b$.

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 [10] 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 $j \leq b$:

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

We will now make this formal.

7.3 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 7.4. *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***:*

$$U(\mathbf{C}) \overset{inc}{\hookrightarrow} U(\mathbf{DF}) \overset{inc'}{\hookleftarrow} \mathbf{DF}_{\mathcal{R}_\mu}$$

(where inc and inc′ are respectively the inclusion maps of $U(\mathbf{C})$ *and* $\mathbf{DF}_{\mathcal{R}_\mu}$ *into* $U(\mathbf{DF})$*) 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:*

$$log \frac{df(\_, x_p, x_a)[j]_*\mu^n}{d\lambda} : \mathbb{R} \to \mathbb{R}$$

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

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

Proposition 14. $\mathbf{DF}_{\mathcal{N}_\mu}$ *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:

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

Since $\mathbf{DF}_{\mathcal{N}_\mu}$ and $\mathbf{DF}_{\mathcal{R}_\mu}$ are small this intersection is well-defined. Note also that if we have:

$$f_1 \in \mathbf{DF}_{\mathcal{N}_\mu}[\mathbb{R}^a, \mathbb{R}^b], f_1 \in \mathbf{DF}_{\mathcal{R}_\mu}[\mathbb{R}^a, \mathbb{R}^b]$$
$$f_2 \in \mathbf{DF}_{\mathcal{N}_\mu}[\mathbb{R}^b, \mathbb{R}^c], f_2 \in \mathbf{DF}_{\mathcal{R}_\mu}[\mathbb{R}^b, \mathbb{R}^c]$$

Then since $\mathbf{DF}_{\mathcal{N}_\mu}$ and $\mathbf{DF}_{\mathcal{R}_\mu}$ are closed under composition it must be that:

$$f_2 \circ f_1 \in \mathbf{DF}_{\mathcal{N}_\mu}[\mathbb{R}^a, \mathbb{R}^c] \cap \mathbf{DF}_{\mathcal{R}_\mu}[\mathbb{R}^a, \mathbb{R}^c]$$

Therefore $\mathbf{C}_{\mathcal{R}_\mu}$ is a semicategory.

Now note that there exist identity-on-objects and identity-on-morphisms inclusion semifunctors:

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

such that the following diagram commutes:

$$\begin{tikzcd}
\mathbf{C}_{\mathcal{R}_\mu} \arrow[r, hook, "id_r"] \arrow[d, hook, "id_l"] & \mathbf{DF}_{\mathcal{R}_\mu} \arrow[d, hook, "inc'"] \\
U(\mathbf{DF}_{\mathcal{N}_\mu}) \arrow[r, hook, "inc"] & U(\mathbf{DF})
\end{tikzcd}$$

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

$$\begin{tikzcd}
\mathbf{C}' \arrow[r, "r"] \arrow[d, "l"] & \mathbf{DF}_{\mathcal{R}_\mu} \arrow[d, hook, "inc'"] \\
U(\mathbf{DF}_{\mathcal{N}_\mu}) \arrow[r, hook, "inc"] & U(\mathbf{DF})
\end{tikzcd}$$

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 must also be in $\mathbf{DF}_{\mathcal{R}_\mu}$ and any object or morphism in the image of r must also be in $U(\mathbf{DF}_{\mathcal{N}_\mu})$. Therefore, any object or morphism in the image of either l or r must also be in $\mathbf{C}_{\mathcal{R}_\mu}$.

We can therefore define a 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

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

Since h must have the same actions on objects and morphisms as l and r to satisfy these equations it must be unique, and therefore $(\mathbf{C}_{\mathcal{R}_\mu}, id_l, id_r)$ is the pullback of the diagram:

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

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 $f(\_, x_p, x_a)[j] : \Omega. \to \mathbb{R}$ is univariate normal (Proposition 8). 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:

$$log \ \frac{df(\_, x_p, x_a)[j]_* \mu^n}{d\lambda} : \mathbb{R} \to \mathbb{R}$$

$$\begin{aligned} log \left(\frac{df(\_, x_p, x_a)[j]_* \mu^n}{d\lambda}(y) \right) &= \\ log \left(\left(\frac{1}{s_n(f(\_, x_p, x_a)[j])\sqrt{2\pi}} \right) \exp \left(- \frac{(y - E_{\mu^n}[f(\_, x_p, x_a)[j]])^2}{2s_n(f(\_, x_p, x_a)[j])^2} \right) \right) &= \\ - \frac{log(2\pi s_n(f(\_, x_p, x_a)[j])^2)}{2} - \frac{1}{2s_n(f(\_, x_p, x_a)[j])^2} \left(y - E_{\mu^n}[f(\_, x_p, x_a)[j]] \right)^2 &. \end{aligned}$$

Therefore:

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

□

7.4 Backpropagation

The arrows in a Marginal Likelihood Factorization Category **C** are equipped with the structure that we need to derive both an optimization objective and a learning procedure. Therefore, for any Marginal Likelihood Factorization Category **C** and choice of learning rate ϵ we can define a backpropagation functor into Fong et al.'s [10] **Learn** category.

Definition 7.5. *Write F_{er} for Fong et al.'s [10] Backpropagation functor with learning rate ϵ under the marginal error function er of* **C**. *We define the functor E_{er} to map a parametric statistical model in* **C** *to a learning algorithm:*

$$\begin{aligned} E_{er} &: \mathbf{C} \to \mathbf{Learn} \\ E_{er} &= F_{er} \circ Exp \end{aligned}$$

Where Exp is defined in Proposition 10. For example, E_{er} 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 **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 **C** are then used to define the loss function er.

8 Closing Thoughts on Categorical Stochastic Processes and Likelihood

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 directly, 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. [9] use to assess how the visual cortex performs a rapid stimulus categorization task.

Once we build such a model, we can use E_{er} to derive a Learner with a structure that incorporates this combined uncertainty. This functor will convert the model to a point estimator and bundle the combined uncertainty into a loss function.

One of the largest differences between this construction and those of Cho et al. [5] and Culbertson et al. [7] 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 functor above aims to find an optimal parameter value given the data we have seen, it makes 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.

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] Matteo Capucci, Bruno Gavranović, Jules Hedges, and Eigil Fjeldgren Rischel. Towards foundations of categorical cybernetics, 2021.

[5] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

[6] G. S. H. Cruttwell, Bruno Gavranović, Neil Ghani, Paul Wilson, and Fabio Zanasi. Categorical foundations of gradient-based learning. *arXiv e-prints arXiv:2103.01931*, 2021.

[7] Jared Culbertson and Kirk Sturtz. Bayesian machine learning via category theory. *arXiv preprint arXiv:1312.1445*, 2013.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. DOI: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

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

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

[19] 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.

[20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[21] 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.