A categorical perspective on conditional probability theory

Ilan Ehrlich

December 2022

1 Abstract

Until quite a late stage in history, mathematics only consisted of very basic and concrete fields such as arithmetics or geometry. As it developed, it started getting wider and wider, as new concepts appeared to model nature and as mathematicians were feeling more and more free about inventing new ones just based on constructions structures that did not have to relate directly to nature. This semester we were introduced to the theory of categories in the course of theory of groups, which is quite a fabulous example of this aquisition of freedom, but also an ambitious attempt to formulate some wide generalization of mathematical structures. I had the idea of linking categories to pribability, so I wanted to present in this essay the very basis of categorical conditional probability, mainly based on Jared Culbertson and Kirk Sturtz's paper on the subject.

2 Introduction

The core idea of linking theory of categories to conditional probability, is to perceive it as new probability measure on the probability space. Indeed, if $(\Omega, \mathcal{F}_\Omega, \mathbb{P})$ is a probability space, the probability $\mathbb{P}(\cdot|x) : \mathcal{F} \to [0,1]$ given some $x \in \Omega$ is a probability measure on $(\Omega, \mathcal{F})$[1]. We can imagine a function $f : \Omega \times \mathcal{F} \to [0,1]$ that takes an additional element $A \in \mathcal{F}$ and gives us the probability of A given x:

$$f(x, A) := \mathbb{P}(A|x)$$

This is the main idea of what we call a *Markov kernel* (defined in 3.2). Starting here, we will have all the tools of theory of categories at our disposal to develop a new perspective on conditional probability theory.

[1] In this essay, we will usually assume that $\{x\} \in \mathcal{F}$ to simplify the construction, despite the loss of generality. In fact, if Ω is countable, singletons have to be measurable in its discrete σ-algebra.

3 The category of conditional probability

First, let us recall that a category consists of a class of *objects* and a class of *morphisms*, that can basically be viewed as dots and arrows between them. For example, groups form a categories, where they are the objects, and the morphisms are the group homomorphisms. We are free to create any category with the objects and the morphisms that we want, as long as we define:
• *Composition* between morphisms that validates associativity
• *Identity* morphisms for each object
Before we introduce our main category, we need to introduce a few definitions:

Definition 3.1 (Perfectness)**.**
A measure space $(\Omega, \mathcal{F}, \mu)$ is called *perfect* if for every measurable function $f : \Omega \to \mathbb{R}$ there exists a Borel set $E \subset f(\Omega)$ such that $\mu(f^{-1}(E)) = \mu(\Omega)$.

Definition 3.2 (Markov kernel)**.**
Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be two measure spaces. A *Markov kernel* from $(X, \mathcal{F}_X)$ to $(Y, \mathcal{F}_Y)$ is a map $\kappa : X \times \mathcal{F}_Y \to [0, 1]$ with the following properties:
• For every fixed $B \in \mathcal{F}_Y$ the map $\kappa(\cdot, B) : X \to [0, 1]$ is measurable
• For every fixed $x \in X$ the map $\kappa(x, \cdot) : \mathcal{F}_Y \to [0, 1]$ is a probability measure on $(Y, \mathcal{F}_Y)$

Notation 3.2.1. Let $\kappa : X \times \mathcal{F}_Y \to [0, 1]$ be a Markov kernel, $x \in X$ some element and $B \in \mathcal{F}_Y$ measurable subset. Then the following notation holds:

$$\kappa_B := \kappa(\cdot, B) : X \to [0, 1]$$

$$\kappa_x := \kappa(x, \cdot) : \mathcal{F}_Y \to [0, 1]$$

We can now define our category:

Definition 3.3 (Category of conditional probabilities)**.**
We call $\mathcal{P}$ the category defined by:
• Objects: countably generated measure spaces $(\Omega, \mathcal{F})$
• Morphisms: a morphism $f : (X, \mathcal{F}_X) \to (Y, \mathcal{F}_Y)$ is a Markov kernel of the form

$$f : X \times \mathcal{F}_Y \to [0, 1]$$

that assigns to each $x \in X$ and $B \in \mathcal{F}_Y$ the probability of B given x[2].
• Composition: for a morphism $T : (X, \mathcal{F}_X) \to (Y, \mathcal{F}_Y)$ and a morphism $U : (Y, \mathcal{F}_Y) \to (Z, \mathcal{F}_Z)$ the composition $U \circ T : X \times \mathcal{F}_Z \to [0, 1]$ is called *marginalization over* Y and consists of:

$$(x, C) \mapsto \int_{y \in Y} U(y, C)\, dT_x$$

[2]The definition 3.2 is the most general one of a Markov kernel, but we will be more pricisely focused on morphisms f such that $f(x, \cdot)$ is a perfect probability measure

Remark 3.3.1. Recall that T_x defines a measure on $(Y, \mathcal{F}_Y)$. The above integral is an integral *with respect to a measure*. The process of its definition is too long to be explained in the present essay, but more details of how this is defined can be found **here**. It is not crucial to be perfectly familiar with it to understand this essay, since its properties are very analogous to a regular integral.

Notation 3.3.1. Oftentimes, an object $(X, \mathcal{F}_X)$ will be just denoted X if there is no ambiguity on its σ-algebra.

Remark 3.3.2. An important fact is that the definition 3.3 can encode any probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ as a $\mathcal{P}$-morphism between the object $(1, \{\varnothing, 1\})$ (just written 1) and the object $(\Omega, \mathcal{F})$. Indeed, since 1 is the only element of the space, the morphism f just encodes one probability measure $\mathbb{P} = f_1$. Therefore, we will usually write a probability measure $\mathbb{P}$ as a morphism $P : 1 \to [0, 1]$.

Remark 3.3.3. Technically, it is necessary to prove that the definition 3.3 is really a category, but this would consist more of algebra than probability, and would not be very helpful in this essay.

Remark 3.3.4. Every measurable function $f : X \to Y$ can be associated with a morphism in $\mathcal{P}$ with the dirac measure:

$$\begin{aligned} \delta_f : X \times \mathcal{F}_Y &\to [0, 1] \\ (x, B) &\mapsto \begin{cases} 1 & \text{if } f(x) \in B \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

4 Theorems of the construction

First, more preliminaries...

Definition 4.1 (Absolute continuity)**.** If μ and ν are two measures on the same measurable space $(\Omega, \mathcal{F})$, μ is said to be *absolutely continuous with respect to* ν if $\mu(A) = 0$ for every set A for which $\nu(A) = 0$.

Theorem 4.2 (Radon-Nikodym, admitted)**.** *Let $(\Omega, \mathcal{F})$ be a measure space on which two σ-finite measures are defined, μ and ν. If ν is absolutely continuous with respect to μ, then there exists a σ-measurable function $f : \Omega \to [0, \infty)$ such that for any measurable set $A \subseteq \Omega$*

$$\nu(A) = \int_A f \, d\mu.$$

The function f satisfying the above equality is uniquely defined up to a μ-null set, that is, if g is another function which satisfies the same property, then $f = g$ μ -almost everywhere. The function f is commonly written $\frac{d\nu}{d\mu}$ and is called the Radon–Nikodym derivative.

Definition 4.3 (Categorical weak product)**.** Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be $\mathcal{P}$-objects. A *weak product of* X and Y consists of an object Π and two morphisms $p : \Pi \to X$, $q : \Pi \to Y$ such that for any object Z and morphisms $f : Z \to X$, $g : Z \to Y$ there exists a morphism $h : Z \to \Pi$ such that the following diagram commutes:

$$\begin{tikzcd}
Z \arrow[drr, bend left, "f"] \arrow[dr, dotted, "h"] \arrow[ddr, bend right, "g"'] & & \\
& \Pi \arrow[r, "p"] \arrow[d, "q"] & X \\
& Y &
\end{tikzcd}$$

Lemma 4.4 (Admitted)**.** *Let* $(X, \mathcal{F}_X)$ *and* $(Y, \mathcal{F}_Y)$ *be* $\mathcal{P}$*-objects and* $(X \times Y, \mathcal{F}_{X\times Y})$ *be their product as seen in probability class with projection maps* $\pi_X : X \times Y \to X$ *and* $\pi_Y : X \times Y \to Y$*. Then*

$$((X \times Y, \mathcal{F}_{X\times Y}), \delta_{\pi_X}, \delta_{\pi_Y})$$

defines a weak product in the categorical sense.

We are finally free to discover our main theorem:

Theorem 4.5. *Let* X *and* Y *be countably generated measurable spaces and* $(X\times Y, \mathcal{F}_{X\times Y})$ *their product with projection map* π_Y*. If* J *is a joint distribution on* $X \times Y$ *with marginal* $P_Y = \delta_{\pi_Y} \circ J$ *on* Y*, then there exists a* $\mathcal{P}$ *morphism* f *that makes the following diagram commute:*

$$\begin{tikzcd}
1 \arrow[d, "J"] \arrow[dr, "P_Y"] & \\
X \times Y \arrow[r, "\delta_{\pi_Y}"'] & Y \arrow[l, dotted, bend left, "f"]
\end{tikzcd}$$

and satisfies for all $C \in \mathcal{F}_Y$

$$\int_{A\times B} (\delta_{\pi_Y})_C dJ = \int_C f_{A\times B} dP_Y$$

Moreover, this f *is the unique* $\mathcal{P}$*-morphism with these properties, up to a set of* P_Y*-measure zero.*

This theorem will be the main fundation for our construction. Let us have a look at its interpretation. If we have a joint distribution on a product of two spaces and therefore a marginal distribution on one of them, this directly induces a funtion of conditional probability. This seems quite natural, since having two spaces of events, it should be possible to consider the probability of an event from one of them given an event from the other.

Proof. Since $\mathcal{F}_X$ and $\mathcal{F}_Y$ are countably generated, $\mathcal{F}_{X\times Y}$ is too. Let $\mathcal{G}$ be a countable generating set of $\mathcal{F}_{X\times Y}$, and let $\mathcal{A}$ be the algebra generated by $\mathcal{G}$ [3]. For each $G \in \mathcal{G}$, we define a measure μ_G on Y by

$$\mu_G(B) := J(G \cap \pi_Y^{-1}(B))$$

Then μ_G is absolutely continuous with respect to P_Y and hence we can let $\tilde{f}_G = \frac{d\mu_G}{dP_Y}$ be the Radon-Nikodym derivative. For each $G \in \mathcal{G}$ it is unique up to a set of measure zero, say $\hat{G}$. Now let $N := \cup_{A\in\mathcal{A}} \hat{A}$ and $E_1 := N^c$. Then $\tilde{f}_A|_{E_1}$ is unique for all $A \in \mathcal{A}$. Note that $\tilde{f}_{X\times Y} = 1$ and $f_\varnothing = 0$ on E_1. The condition $\tilde{f}_A \leq 1$ on E_1 for all $A \in \mathcal{A}$ then follows.
For all $B \in \mathcal{F}_Y$ and any countable union $\cup_{i=1}^n A_i$ of disjoint sets of $\mathcal{A}$ we have:

$$\begin{aligned}
\int_{B\cap E_1} \tilde{f}_{\cup_{i=1}^n A_i} dP_Y &= J((\cup_{i=1}^n A_i) \cap \pi_Y^{-1}(B)) \ \text{(by definition of the derivative and the measure)} \\
&= \sum_{i=1}^n J(A_i \cap \pi_Y^{-1}(B)) \\
&= \int_{B\cap E_1} \sum_{i=1}^n \tilde{f}_{A_i} dPY
\end{aligned}$$

with the last equality following from the Monotone convergence Theorem and the fact that all of the $\tilde{f}_{A_i}$ are nonnegative. From the uniqueness of the Radon–Nikodym derivative it follows

$$\tilde{f}_{\cup_{i=1}^n A_i} = \sum_{i=1}^n \tilde{f}_{A_i} \qquad P_Y\text{-almost everywhere}$$

Since there exist only a countable number of finite collection of sets of $\mathcal{A}$ we can find a set $E \subset E_1$ of P_Y-measure one such that the normalized set function $\tilde{f}.(y) : \mathcal{A} \to [0,1]$ is finitely additive on E. These facts altogether show there exists a set $E \in \mathcal{F}_Y$ with P_Y-measure one where for all $y \in E$:

1. $0 \leq f_A(y) \leq 1 \forall A \in \mathcal{A}$
2. $\tilde{f}_\varnothing(y) = 0$ and $\tilde{f}_{X\times Y(y)} = 1$
3. for any finite collection $\{A_i\}_{i=1}^n$ of disjoint sets of $\mathcal{A}$ we have $\tilde{f}_{\cup_{i=1}^n A_i}(y) = \sum_{i=1}^n \tilde{f}_{A_i}(y)$

Thus the set function $\tilde{f} : E \times \mathcal{A} \to [0,1]$ satisfies the condition that $\tilde{f}(y,\cdot)$ is a probability measure on the algebra $\mathcal{A}$. By the Caratheodory extension

[3] Let us be careful not getting confused between $\mathcal{G}, \mathcal{A}$ and $\mathcal{F}_{X\times Y}$. $\mathcal{A}$ is the algebra generated by *finite* unions and intersectons of sets in $\mathcal{G}$, whereas $\mathcal{F}_{X\times Y}$ is the σ-algebra generated by *countable* unions and intersections of sets in $\mathcal{G}$. We therefore have $\mathcal{A} \subset \mathcal{F}_{X\times Y}$. Note that they are however equal if $X \times Y$ is finite.

theorem, there exists a unique extenson of $\tilde{f}(y, \cdot)$ to a probability measure $\hat{f}(y, \cdot) : \mathcal{F}_{X \times Y} \to [0, 1]$. We can now define the expected morphism $f : Y \times \mathcal{F}_{X \times Y} \to [0, 1]$ by:

$$f(y, A) = \begin{cases} \hat{f}(y, A) & \text{if } y \in E \\ J(A) & \text{otherwise} \end{cases}$$

Let us now prove that this construction satisfies the axiom of commutativity. We want to show that $f \circ P_Y(A) = J(A)$ for all $A \in \mathcal{F}_{X \times Y}$. Since $\mathcal{F}_{X \times Y}$ is countably generated, there has to be a sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$ such that we have:

$$\begin{aligned} f \circ P_Y(A) &= \int_Y f_A \, dP_Y = \int_E f_A dP_Y \\ &= \lim_{n \to \infty} \int_E \tilde{f}_{A_n} dP_Y \qquad \text{(permuting the limit and the integral is legal} \\ &= \lim_{n \to \infty} \int_Y \tilde{f}_{A_n} dP_Y \qquad \text{by construction of the integral with respect to a measure)} \\ &= \lim_{n \to \infty} J(A_n) \\ &= J(A) \end{aligned}$$

□

Corollary 4.6 (Reformulation)**.**
Let X and Y be countably generated measurable spaces and J a joint distribution on $X \times Y$ with marginal distributions P_X and P_Y on X and Y, respectively. Then there exist $\mathcal{P}$-morphisms $f : Y \to X \times Y$ and $g : X \to X \times Y$ such that the following diagram commutes:

$$\begin{tikzcd}
& 1 \arrow[d, "J"] \arrow[ddl, "P_X"'] \arrow[ddr, "P_Y"] & \\
& X \times Y \arrow[dl, "\delta_{\pi_X}"] \arrow[dr, "\delta_{\pi_Y}"'] & \\
X \arrow[ur, dashed, bend left, "g"] \arrow[rr, "\delta_{\pi_Y} \circ g"] & & Y \arrow[ul, dashed, bend right, "f"'] \arrow[ll, bend left, "\delta_{\pi_X} \circ f"]
\end{tikzcd}$$

and the equation :

$$\int_U (\delta_{\pi_Y} \circ g)_V dP_X = J(U \times V) = \int_V (\delta_{\pi_X} \circ f)_U dP_Y$$

is satisfied for all $U \in \mathcal{F}_X, V \in \mathcal{F}_Y$.

Proof. From theorem 4.5 there exists a $\mathcal{P}$-morphism $f : Y \to X \times Y$ satisfying $J = f \circ P_Y$. Take the composite $\delta_{\pi_X} \circ f$ and note $(\delta_{\pi_X} \circ f)_U(y) = f_y(U \times V)$ giving:

$$\begin{aligned}
\int_V (\delta_{\pi_X} \circ f)_U dP_Y &= \int_V f_{U\times Y} dP_Y \\
&= J(U \times Y \cap \pi_Y^{-1}(V)) \\
&= J(U \times V)
\end{aligned}$$

Similarily, using a $\mathcal{P}$-morphism $g : X \to X \times Y$ satisfying $J = g \circ P_X$ gives:

$$\int_U (\delta_{\pi_Y} \circ g)_V dP_X = J(U \times V)$$

□

This corollary establishes more clearly the induction of a conditional measure from X to Y and vice versa, when we have a joint distribution on $X \times Y$, which is exactly the result we would expect from a product measure space.

5 Conclusion

As we have seen in this essay, categories can be a strong basis for conditional probabilities, and all of the natural properties we developed in class can be very well represented in the language of objects and arrows. Unfortunately, this essay was to be too short to present more unexpected results that categories allow beyond the usual view point that we have on probability. To go a little bit further, we could link the result from the last corollary to a categorical formulation of Bayes' theorem, which would enrich our sense of the analogy between the regular approach of probability and the categorical one. I will leave free to the reader to imagine the meaning of the following result as an epilogue:

Theorem 5.1 (Bayes'). *If $(X, \mathcal{F}_X, P_X)$ and (Y, F_Y, P_Y) are two probability spaces, then there exists $\mathcal{P}$-morphisms $f : X \to Y$, $g : Y \to X$ such that the following diagram commutes, with the proper definition of the product morphism:*

$$\begin{tikzcd}
X \arrow[d, "\delta_{\Delta_X}"] & 1 \arrow[l, "P_X"'] \arrow[r, "P_Y"] \arrow[d, "J"'] & Y \arrow[d, "\delta_{\Delta_Y}"] \\
X \times X \arrow[r, "id_X \times f"] & X \times Y & Y \times Y \arrow[l, "g \times id_Y"']
\end{tikzcd}$$

where $\Delta : X \to X \times X$ is the diagonal map that simply duplicates a space.

6 Bibliography

[1] This essay was mainly based on Jared Culbertson and Kirk Sturtz's work found **here**, for 3.1,3.2.1,3.3,3.3.4,4.3,4.4,4.5,4.6.
[2] It was also widely based on Athur Parzygnat's work found **here** and **there** for the general understanding of what we are trying to do, as Culbertson and Sturtz's paper is less preoccupied with vulgarization.
[3] Parts 3.2 4.1, 4.2 were found on Wikipedia **here**, **there** and **there**.