Adjoint School 2024 Subgroup 1B Blogpost

Drew, Nico, and Utku

Introduction (this probably doesn't need a section title)

We assume that the reader is familiar with symmetric monoidal categories and string diagrams. All of our string diagrams are to be read bottom to top.

- As said by Tomas, spell out the motivation here, and make sure readers know that there's a big reward at the end.
- Maybe an illustration? My motivator to learn this was the difference between the cumbersome algebra from my professor's notes (Section 9.4 of the attached estimation theory pdf) vs the simplicity of Proposition 3.7 in this paper

Basics of Probability Theory (UTKU)

Probability distributions

We want to proceed with our discussion through an example, and so before we introduce everything, consider the following:

> You've just installed a sprinkler system to your lawn! It is a very advanced piece of technology, measuring a myriad of different things to determine when to turn on the sprinklers... and you have no idea how it does this. In your effort to have an idea of when the system turns on (you pay the water bill, after all) you decided to keep track of how the weather feels and whether your sprinkler is on or not.

Here's what you have: You make the following distinctions: Weather = {sunny, cloudy, rainy}, Humidity = {dry, humid}, Temperature = {hot, mild, cold}, Sprinkler = {on, off}

Weather	Humidity	Temperature	Sprinkler
sunny	humid	mild	off
sunny	dry	hot	on
cloudy	dry	hot	on
cloudy	humid	mild	on
rainy	humid	cold	off
cloudy	dry	cold	on
sunny	humid	cold	off

> You make an assumption that the frequency with which each weather event occurred would be an accurate estimate for how it will be in

the future, and so you assemble the previous 3 months' weather data into probability distributions.

A probability distribution on a finite set X is a function $p : 2^X \to [0, 1]$ assigning to each subset $A \subset X$ a number $p(A)$ such that

- $p(\emptyset) = 0$,
- $p(X) = 1$,
- and for disjoint subsets $A_1, \ldots, A_k \subset X$, $\sum_i p(A_i) = p(\bigcup_i A_i)$.

For our purposes, a simpler characterization exists from the fact that we can consider a set to disjointly consist of its individual points; namely we can think of a probability distribution on X to be a function $p : X \to [0, 1]$ such that

$$\sum_{x \in X} p(x) = 1$$

We will also make use of the bra-ket notation to denote a distribution/state on X; for $X := \{x_1, \ldots, x_k\}$ with the values $\lambda_i := p(x_i)$, the following notation also describes a distribution on X:

$$\sum_{i=1}^{k} \lambda_i = 1 \leftrightsquigarrow \lambda_1 \mid x_1\rangle + \lambda_2 \mid x_2\rangle + \cdots + \lambda_k \mid x_k\rangle$$

Given this notion, we can model the transition between "state spaces" X to Y by means of a *stochastic matrix*, which is a matrix $f : X \times Y \to [0, 1]$ such that each column sums to 1, which we denote

$$\sum_{y \in Y} f(y \mid x) = 1$$

Following our established bra-ket notation, we can equivalently describe the action of the channel $f : X \to Y$ by

$$f_x : \gamma_1 \mid y_1\rangle + \gamma_2 \mid y_2\rangle + \cdots + \gamma_n \mid y_n\rangle$$

with $\gamma_i := f(y_i \mid x)$ and f_x forming a probability distribution on Y.

Furthermore, given two channels $f : X \to Y$ and $g : Y \to Z$, we also have a way of obtaining a composite channel $g \circ f : X \to Z$, by the Chapman-Kolmogorov formula, defining the channel

$$(g \circ f)(z \mid x) := \sum_{y \in Y} g(z \mid y) f(y \mid x)$$

You can interpret these distributions to be channels from the singleton set to their respective sets: $p : * \to W$, $q : * \to H$, $r : * \to T$. Then, composing any such distribution with a channel will again yield a distribution

$$* \xrightarrow{p} X \xrightarrow{f} Y$$

Consider the example scenario we described above. Suppose that you compiled the historical weather data into the following probability distribution $p : * \to W \otimes H \otimes T$ (more to come about $\otimes$ in just a second):

$$p_* : 0.2 \mid s, d, h\rangle + 0.3 \mid r, h, c\rangle + 0.3 \mid c, h, m\rangle + 0.2 \mid c, d, h\rangle$$

From the table in the example, we can obtain the following channel $f : W \otimes H \otimes T \to S$ if we assume the principle of indifference, i.e., that the entries in the table all occur with equal probability (which would be the case if these were a list of observations), we get a channel

$$f_{(w,h,t)} = \delta^{\text{on}}_{wht} \mid \text{on}\rangle + \delta^{\text{off}}_{wht} \mid \text{off}\rangle$$

Then, by everything we've established so far, we can reason about the likelihood that the sprinkler will turn on the next day by composing the state p of the climate with the channel f to obtain a state $f \circ p$ of the sprinkler, computed

$$f \circ p : 0.7 \mid \text{on}\rangle + 0.3 \mid \text{off}\rangle$$

All in all, along with the identity matrices, all this data assembles into the category FinStoch with

- objects: finite sets
- morphisms: stochastic matrices
- where the composition is determined through the Chapman-Kolmogorov formula

This is one of the first examples of a Markov category that we will be looking at, and it will be a good baseline to observe why a Markov category is defined the way it is.

Possibility distribution

Markov categories need not only house probabilistic models of uncertainty; we'll see that the following also forms a Markov category:

Consider a channel between two finite sets X, Y to be an assignment $f : X \to Y$ such that each $f(x) \subset Y$ is a non-empty subset. Defining the composition to be

$$g \circ f(x) := \bigcup_{y \in f(x)} g(y)$$

and the identities as $x \mapsto \{x\}$ gives us the Markov category FinSetMulti of possibilities!

The same data from the example can be used in a possibilistic way as well; a channel $S \to W \otimes H \otimes T$ can map the sprinkler to all the possible states of weather/climate where the sprinkler has turned on etc.

Channels are Kleisli maps

Something you may have noticed from the two examples of morphisms of Markov categories is that fixing an element $x \in X$ yields some structure attached to Y with "desirable properties": in the case of FinStoch, we have that each f_x is a probability distribution on Y – in fact, the Chapman-Kolmogorov formula further provides a way to obtain a probability distribution from a probability distribution of probability distributions. In the case of FinSetMulti, each f_x is a non-empty subset of Y, and the composition is provided through the union of a set of sets.

This is not a coincidence: we will see that for certain monads, the Kleisli category they yield turn out to be Markov categories! The monads in question will provide us descriptions of what the channels are, as well as the rule for composition.

Kleisli Categories (Should this be a subsection of above?) (NICO)

- Example probability monads (Construct `flatten`, `dirac`, and `zipper` for each)
  - Finite distribution monad
  - Powerset monad
  - Briefly mention giry monad
- Kleisli categories
  - What structures do the Kleisli categories lose (and what do they keep) from their base (Cartesian) counterparts?
    - They do keep comonoid structures
    - But they're no longer Cartesian
    - Copy map is no longer natural
    - Products are no longer categorical products, ie. projections are no longer universal. What does this mean in terms of probability? (Answer: unlike Cartesian projection, you cannot in general reconstruct a joint probability distribution from its marginals)
    - Delete is still natural though, ie. unit object is still final
    - This all plays into "equivalent characterizations of deterministic Markov categories"

Markov Categories

Formal definition

Let's start with the terse definition that category theorists love so much: A Markov category is a semiCartesian category where every object is a comonoid compatible with the monoidal structure.

(Now give a more explicit definition. Should we give both string diagram equations and commutative diagrams? Or just stick to one?)

Each Axiom Explained

Let's go a little bit more in-depth into why each of these axioms are required. (Bring in our established example setting into each of the subsections below.)

Composition and Identity (Utku)

The necessity for composition and identities in a categorical setting requires no explanation, though we note that the mental image of "information flow" is essentially channels/Markov kernels taking states to states. The flow of information is essentially a pushforward.

Monoidal Products (Nico)

We want to describe distributions over joint variables.

Swap Map (Drew)

Copy Map (Drew)

We want this because it makes sense to process the same data in multiple different ways and then compare them. Show for instance the "graph" of a morphism

Why should this be compatible with the monoidal structure?

Delete Map (Nico)

In probability theory: marginalization. In information processing:deleting information seems desirable (even though impossible in quantum information theory)

Why should it be natural? Equivalently, why should the tensor unit be terminal? In this sense, why should del be compatible with the monoidal structure?

- This corresponds to normalization
- Deleting an output of a process deletes the whole process
- Omitting this leads to CD-categories

Important Markov categories

- The most important construction: Kleisli categories of symmetric monoidal monads
- FinSupStoch := Kl(D)
- Finstoch
- Gauss

Additional Axioms and definitions (Drew)

Independence

Conditionals, Bayesian Inversion

Determinism

Almost-sure equality

Representability?

Conclusion: Cool things you can do with Markov categories

- De Finetti
- HMMs and Bayesian Inversion
- Causal Inferencing