An introduction to categorical probability theory

Rob Cornish

Department of Statistics, University of Oxford

March 8, 2024

Motivation

It should be said: for someone trained in formal methods, the area of probability theory can be rather sloppy: everything is called 'P', types are hardly ever used, crucial ingredients (like distributions in expected values) are left implicit, basic notions (like conjugate prior) are introduced only via examples, calculation recipes and algorithms are regularly just given, without explanation, goal or justification, etc. This hurts, especially because there is so much beautiful mathematical structure around. (Jacobs [2019])

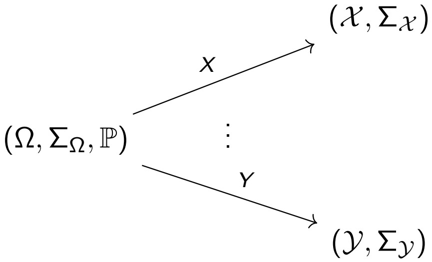
Classical probability theory

Start with an underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$

Classical probability theory

Start with an underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$

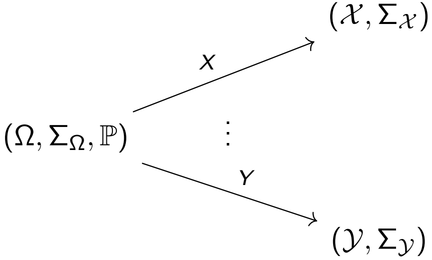
Model phenomena of interest using random variables (i.e. measurable functions) $X : \Omega \to \mathcal{X}$, i.e.



Classical probability theory

Start with an underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$

Model phenomena of interest using random variables (i.e. measurable functions) $X : \Omega \to \mathcal{X}$, i.e.

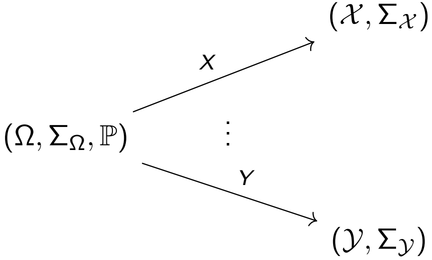


Can consider many distinct $\mathcal{X}$ and $\mathcal{Y}$, but Ω is fixed throughout

Classical probability theory

Start with an underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$

Model phenomena of interest using random variables (i.e. measurable functions) $X : \Omega \to \mathcal{X}$, i.e.



Can consider many distinct $\mathcal{X}$ and $\mathcal{Y}$, but Ω is fixed throughout

Usually study the joint or marginal behaviour of X, Y, etc.

Problems

This picture is quite complex

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$
- Random variables X, Y, etc.

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$
- Random variables X, Y, etc.
- Joint and marginal distributions of X, Y, etc.

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$
- Random variables X, Y, etc.
- Joint and marginal distributions of X, Y, etc.

Also somewhat at odds with how we think intuitively:

- Distributions are secondary objects (cf. Bayesian statistics)

Problems

This picture is quite complex

Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$
- Random variables X, Y, etc.
- Joint and marginal distributions of X, Y, etc.

Also somewhat at odds with how we think intuitively:

- Distributions are secondary objects (cf. Bayesian statistics)
- Random variables are static (can't "sample" from them)
  - OK for fixed datasets, but often ill-suited for describing computation

Problems

This picture is quite complex

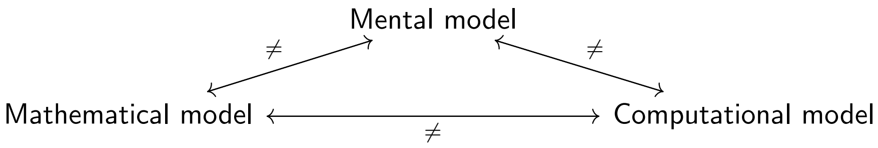
Many seemingly different components playing different roles:

- The underlying measurable space (Ω, Σ_Ω)
- The probability measure $\mathbb{P}$
- Random variables X, Y, etc.
- Joint and marginal distributions of X, Y, etc.

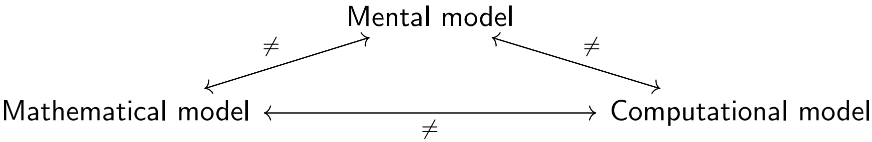
Also somewhat at odds with how we think intuitively:

- Distributions are secondary objects (cf. Bayesian statistics)
- Random variables are static (can't "sample" from them)
  - OK for fixed datasets, but often ill-suited for describing computation
- Kolmogorov-style conditioning is highly technical

Practical implications



Practical implications



Implications:

- Lack of conceptual scalability that often requires hand-waving
- Difficult to interface with other mathematical theories
- Impediment to formal verification and automation
- A challenge pedagogically

Main point

Categorical probability reorganises the existing theory in a way that makes reasoning about higher-level concepts easy and intuitive

Theory becomes much more like a (high-level, expressive) programming language

Case study

PROBABILISTIC SYMMETRIES AND INVARIANT NEURAL NETWORKS

By Benjamin Bloem-Reddy[1] and Yee Whye Teh[2]

[1]*Department of Statistics*
University of British Columbia
benbr@stat.ubc.ca

[2]*Department of Statistics*
University of Oxford
y.w.teh@stats.ox.ac.uk

Background: group invariance

Often it is desirable for a function $f : \mathcal{X} \to \mathcal{Y}$ to be invariant to the action of a group $\mathcal{G}$

Background: group invariance

Often it is desirable for a function $f : \mathcal{X} \to \mathcal{Y}$ to be invariant to the action of a group $\mathcal{G}$

Example:

- $\mathcal{X}$ consists of sequences of profiles of subjects in an i.i.d. population

Background: group invariance

Often it is desirable for a function $f : \mathcal{X} \to \mathcal{Y}$ to be invariant to the action of a group $\mathcal{G}$

Example:

- $\mathcal{X}$ consists of sequences of profiles of subjects in an i.i.d. population
- $\mathcal{G}$ consists of permutations of the indices of these sequences

Background: group invariance

Often it is desirable for a function $f : \mathcal{X} \to \mathcal{Y}$ to be invariant to the action of a group $\mathcal{G}$

Example:

- $\mathcal{X}$ consists of sequences of profiles of subjects in an i.i.d. population
- $\mathcal{G}$ consists of permutations of the indices of these sequences
- $f : \mathcal{X} \to \mathcal{Y}$ makes some prediction about the population

Background: group invariance

Often it is desirable for a function $f : \mathcal{X} \to \mathcal{Y}$ to be invariant to the action of a group $\mathcal{G}$

Example:

- $\mathcal{X}$ consists of sequences of profiles of subjects in an i.i.d. population
- $\mathcal{G}$ consists of permutations of the indices of these sequences
- $f : \mathcal{X} \to \mathcal{Y}$ makes some prediction about the population

Important question: for a given group $\mathcal{G}$, characterise the class of $f : \mathcal{X} \to \mathcal{Y}$ such that

$$f(g \cdot x) = f(x) \qquad \text{for all } g \in \mathcal{G} \text{ and } x \in \mathcal{X}$$

Probabilistic Symmetries

Bloem-Reddy and Teh [2020] consider a probabilistic version of this

Probabilistic Symmetries

Bloem-Reddy and Teh [2020] consider a probabilistic version of this

Setup: $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are random variables representing data and prediction respectively

Probabilistic Symmetries

Bloem-Reddy and Teh [2020] consider a probabilistic version of this

Setup: $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are random variables representing data and prediction respectively

Aim is to characterise when Y is conditionally $\mathcal{G}$-invariant in the sense that

$$\mathbb{P}(Y \in B \mid X \in A) = \mathbb{P}(Y \in B \mid X \in g \cdot A)$$

for all $g \in \mathcal{G}$, $A \in \Sigma_{\mathcal{X}}$ with $\mathbb{P}(X \in A) > 0$, and $B \in \Sigma_{\mathcal{Y}}$

Main result (on invariance)

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \, . \tag{14}$$

Main result (on invariance)

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

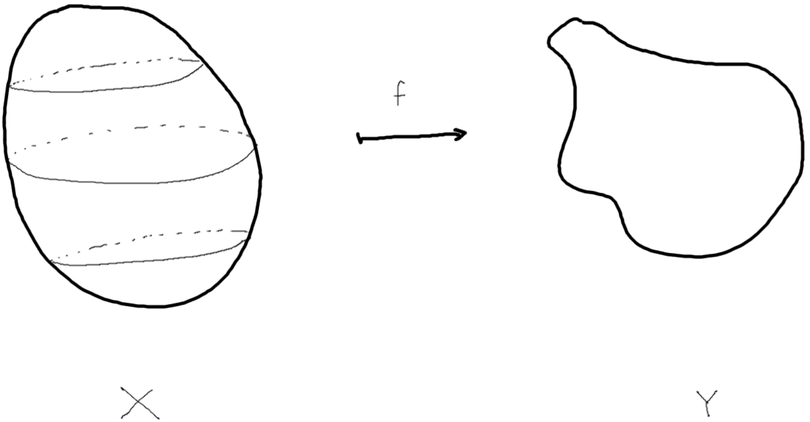
$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \,. \tag{14}$$

Here a maximal invariant is any measurable function M such that

$$M(x) = M(x') \Leftrightarrow x = g \cdot x' \text{ for some } g \in \mathcal{G}$$

(picture next slide)

Orbits



Proof

The proof of this is complex and uses highly technical ideas from advanced probability theory, e.g.

- Measurable cross section
- Normalised Haar measure
- Orbit law
- Conditional independence (of X and Y given $M(X)$)

Proof

The proof of this is complex and uses highly technical ideas from advanced probability theory, e.g.

- Measurable cross section
- Normalised Haar measure
- Orbit law
- Conditional independence (of X and Y given $M(X)$)

Also only applies when $\mathcal{G}$ is compact and X has a $\mathcal{G}$-invariant marginal

Thoughts

Why is this so hard to show? (E.g. compare deterministic case)

Thoughts

Why is this so hard to show? (E.g. compare deterministic case)

Is it optimal to model a neural network in terms of random variables (X, Y)? And why must $\mathrm{Law}[X]$ be $\mathcal{G}$-invariant?

Thoughts

Why is this so hard to show? (E.g. compare deterministic case)

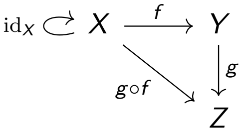
Is it optimal to model a neural network in terms of random variables (X, Y)? And why must $\mathrm{Law}[X]$ be $\mathcal{G}$-invariant?

With the tools of categorical probability, we can not only generalise this result, but we can prove it in a way that maps directly onto our intuitions

Categorical probability theory

Category theory

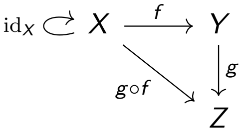
A category (often) models a collection of entities that behave like functions:



Here X, Y, Z are objects and $f : X \to Y$, $g : Y \to X$ are arrows or morphisms

Category theory

A category (often) models a collection of entities that behave like functions:



Here X, Y, Z are objects and $f : X \to Y$, $g : Y \to X$ are arrows or morphisms

Minimal structure:

- We can compose compatibly typed morphisms
- We have identity arrows

Formal definition

Definition

A category consists of a collection of objects and a collection of arrows

Each arrow f has a source and target object, denoted $f : \mathcal{X} \to \mathcal{Y}$

There is a composition operation $\circ$ on arrows such that

$$
\begin{aligned}
g \circ f : \mathcal{X} \to \mathcal{Z} \qquad & \text{whenever } f : \mathcal{X} \to \mathcal{Y} \text{ and } g : \mathcal{Y} \to \mathcal{Z} \\
h \circ (g \circ f) = (h \circ g) \circ f \qquad & \text{when } f, g, h \text{ are appropriately typed}
\end{aligned}
$$

For every object $\mathcal{X}$ there is an identity arrow $\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ satisfying

$$
\begin{aligned}
f \circ \mathrm{id}_{\mathcal{X}} = f \qquad & \text{whenever } f : \mathcal{X} \to \mathcal{Y} \\
\mathrm{id}_{\mathcal{X}} \circ g = g \qquad & \text{whenever } g : \mathcal{Z} \to \mathcal{X}
\end{aligned}
$$

Philosophy: study structural properties extrinsically in terms of arrows

Examples

Categories are everywhere:

- **Set**, the category of sets and functions
- **Top**, the category of topological spaces and continuous functions
- **Meas**, the category of measurable spaces and measurable functions
- etc...

Examples

Categories are everywhere:

- **Set**, the category of sets and functions
- **Top**, the category of topological spaces and continuous functions
- **Meas**, the category of measurable spaces and measurable functions
- etc...

Not all categories look like this, e.g.:

- **Stoch**, the category of measurable spaces and Markov kernels

Examples

Categories are everywhere:

- **Set**, the category of sets and functions
- **Top**, the category of topological spaces and continuous functions
- **Meas**, the category of measurable spaces and measurable functions
- etc...

Not all categories look like this, e.g.:

- **Stoch**, the category of measurable spaces and Markov kernels
- A group $\mathcal{G}$ can be viewed as a category (with a single object, and inverses)
- A poset can be viewed as a category (with a unique arrow $x \to y$ iff $x \leq y$)

Functors

The only other definition we will need is that of a functor

Functors

The only other definition we will need is that of a functor

Idea: a functor $F : \mathsf{C} \to \mathsf{D}$ is an arrow between categories:

$$\begin{array}{ccc} X & & FX \\ \downarrow{\scriptstyle f} & \mapsto & \downarrow{\scriptstyle Ff} \\ Y & & FY \end{array}$$

Note overloaded on both objects and arrows

Functors

The only other definition we will need is that of a functor

Idea: a functor $F : \mathsf{C} \to \mathsf{D}$ is an arrow between categories:

$$\begin{array}{ccc} X & & FX \\ \Big\downarrow f & \mapsto & \Big\downarrow Ff \\ Y & & FY \end{array}$$

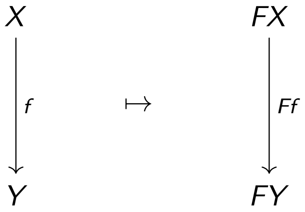
Note overloaded on both objects and arrows

Must satisfy $F(g \circ f) = Fg \circ Ff$ and $F\mathrm{id}_{\mathcal{X}} = \mathrm{id}_{F\mathcal{X}}$

Functors

The only other definition we will need is that of a functor

Idea: a functor $F : \mathsf{C} \to \mathsf{D}$ is an arrow between categories:

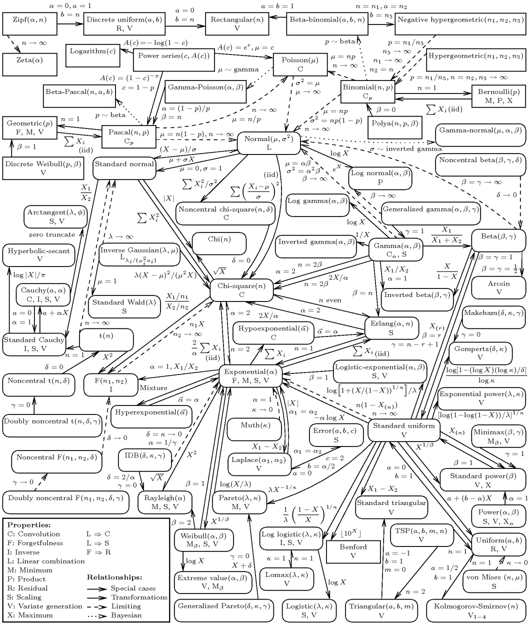


Note overloaded on both objects and arrows

Must satisfy $F(g \circ f) = Fg \circ Ff$ and $F\mathrm{id}_{\mathcal{X}} = \mathrm{id}_{F\mathcal{X}}$

Categories and functors themselves form a category...

Analogy


Zipf(α, n)
Discrete uniform(a, b) R, V
Rectangular(n) V
Beta-binomial(a, b, n)
Negative hypergeometric(n_1, n_2, n_3)
Zeta(α)
Logarithm(c)
Power series($c, A(c)$)
Poisson(μ) C
Hypergeometric(n_1, n_2, n_3)
Beta-Pascal(n, a, b)
Gamma-Poisson(α, β)
Binomial(n, p) C_p
Bernoulli(p) M, P, X
Geometric(p) F, M, V
Pascal(n, p) C_p
Polya(n, p, β)
Normal(μ, σ^2) L
Gamma-normal(μ, α, β)
Discrete Weibull(p, β) V
Standard normal
Noncentral beta(β, γ, δ)
Log normal(α, β) P
Arctangent(λ, ϕ) S, V
Noncentral chi-square(n, δ) C
Log gamma(α, β)
Generalized gamma(α, β, γ)
Hyperbolic-secant V
Chi(n)
Inverse Gaussian(λ, μ) L
Inverted gamma(α, β)
Gamma(α, β) C_α, S
Beta(β, γ)
Cauchy(a, α) C, I, S, V
Standard Wald(λ) S
Chi-square(n) C
Inverted beta(β, γ)
Arcsin V
Makeham(δ, κ, γ)
Standard Cauchy I, S, V
t(n)
Hypoexponential($\vec{\alpha}$) C
Erlang(α, n) S
Gompertz(δ, κ) V
Noncentral t(n, δ)
F(n_1, n_2) I
Exponential(α) F, M, S, V
Logistic-exponential(α, β) S, V
Exponential power(λ, κ) V
Doubly noncentral t(n, δ, γ)
Hyperexponential($\vec{\alpha}$)
Muth(κ)
Error(a, b, c) S
Standard uniform V
Minimax(β, γ) M_β, V
Noncentral F(n_1, n_2, δ)
IDB(δ, κ, γ)
Laplace(α_1, α_2) V
Standard power(β) V, X
Doubly noncentral F(n_1, n_2, δ, γ)
Rayleigh(α) M, S, V
Pareto(λ, κ) M, V
Standard triangular V
Power(α, β) S, V, X_α
Weibull(α, β) M_β, S, V
Log logistic(λ, κ) I, S, V
Benford V
TSP(a, b, m, n) V
Uniform(a, b) R, V
Extreme value(α, β) V, M_β
Lomax(λ, κ) V
von Mises (κ, μ) S
Generalized Pareto(δ, κ, γ)
Logistic(λ, κ) S, V
Triangular(a, b, m) V
Kolmogorov-Smirnov(n) V_{1-4}
Properties:
C: Convolution
F: Forgetfulness
I: Inverse
L: Linear combination
M: Minimum
P: Product
R: Residual
S: Scaling
V: Variate generation
X: Maximum
L ⇒ C
L ⇒ S
F ⇒ R
Relationships:
Special cases
Transformations
Limiting
Bayesian

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_{\mathcal{X}}$ implicit)

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_{\mathcal{X}}$ implicit)

It turns out P can be thought of as a functor **Meas** $\rightarrow$ **Meas**:

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_\mathcal{X}$ implicit)

It turns out P can be thought of as a functor **Meas** $\to$ **Meas**:

- Equip $P\mathcal{X}$ with the (initial) σ-algebra generated by the functions:

$$\begin{aligned} \mathsf{eval}_A : P\mathcal{X} &\to [0,1] \qquad \text{where } A \in \Sigma_\mathcal{X} \\ p &\mapsto p(A) \end{aligned}$$

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_{\mathcal{X}}$ implicit)

It turns out P can be thought of as a functor **Meas** $\to$ **Meas**:

- Equip $P\mathcal{X}$ with the (initial) σ-algebra generated by the functions:

$$\begin{aligned} \mathrm{eval}_A : P\mathcal{X} &\to [0,1] \qquad \text{where } A \in \Sigma_{\mathcal{X}} \\ p &\mapsto p(A) \end{aligned}$$

- For measurable $f : \mathcal{X} \to \mathcal{Y}$, define Pf by the pushforward, i.e.

$$\begin{aligned} Pf : P\mathcal{X} &\to P\mathcal{Y} \\ Pf(p) &\mapsto f\#p \end{aligned}$$

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_{\mathcal{X}}$ implicit)

It turns out P can be thought of as a functor **Meas** $\to$ **Meas**:

- Equip $P\mathcal{X}$ with the (initial) σ-algebra generated by the functions:

$$\begin{aligned} \mathrm{eval}_A : P\mathcal{X} &\to [0,1] \qquad \text{where } A \in \Sigma_{\mathcal{X}} \\ p &\mapsto p(A) \end{aligned}$$

- For measurable $f : \mathcal{X} \to \mathcal{Y}$, define Pf by the pushforward, i.e.

$$\begin{aligned} Pf : P\mathcal{X} &\to P\mathcal{Y} \\ Pf(p) &\mapsto f\#p \end{aligned}$$

- Check functor axioms hold

The Giry Functor [Giry, 1982]

Denote by $P\mathcal{X}$ the set of probability measures on $\mathcal{X}$ (where $\Sigma_{\mathcal{X}}$ implicit)

It turns out P can be thought of as a functor **Meas** → **Meas**:

- Equip $P\mathcal{X}$ with the (initial) σ-algebra generated by the functions:

$$\begin{aligned} \mathsf{eval}_A : P\mathcal{X} &\to [0,1] \qquad \text{where } A \in \Sigma_{\mathcal{X}} \\ p &\mapsto p(A) \end{aligned}$$

- For measurable $f : \mathcal{X} \to \mathcal{Y}$, define Pf by the pushforward, i.e.

$$\begin{aligned} Pf : P\mathcal{X} &\to P\mathcal{Y} \\ Pf(p) &\mapsto f\#p \end{aligned}$$

- Check functor axioms hold

This reduces already the complexity of our original picture (since $\mathbb{P} \in P\Omega$)

Markov kernels

Consider a measurable function $k : \mathcal{X} \to P\mathcal{Y}$

Markov kernels

Consider a measurable function $k : \mathcal{X} \to P\mathcal{Y}$

By definition of P:

- $k(x)(-)$ is a probability measure for all $x \in \mathcal{X}$
- $k(-)(B) = \mathrm{eval}_B \circ k$ is measurable for all $B \in \Sigma_{\mathcal{Y}}$

Markov kernels

Consider a measurable function $k : \mathcal{X} \to P\mathcal{Y}$

By definition of P:

- $k(x)(-)$ is a probability measure for all $x \in \mathcal{X}$
- $k(-)(B) = \mathrm{eval}_B \circ k$ is measurable for all $B \in \Sigma_{\mathcal{Y}}$

Hence k is a Markov kernel: can think of as $k : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0,1]$ such that

- $k(x,-)$ is a probability measure for all $x \in \mathcal{X}$
- $k(-,B)$ is measurable for all $B \in \Sigma_{\mathcal{Y}}$

Markov kernels

Consider a measurable function $k : \mathcal{X} \to P\mathcal{Y}$

By definition of P:

- $k(x)(-)$ is a probability measure for all $x \in \mathcal{X}$
- $k(-)(B) = \mathsf{eval}_B \circ k$ is measurable for all $B \in \Sigma_\mathcal{Y}$

Hence k is a Markov kernel: can think of as $k : \mathcal{X} \times \Sigma_\mathcal{Y} \to [0, 1]$ such that

- $k(x, -)$ is a probability measure for all $x \in \mathcal{X}$
- $k(-, B)$ is measurable for all $B \in \Sigma_\mathcal{Y}$

(Precisely: write $k : \mathcal{X} \to P\mathcal{Y}$ as $k : \mathcal{X} \to (\Sigma_\mathcal{Y} \mapsto [0, 1])$ and uncurry)

Giry monad

We can consider Markov kernels to be generalised measurable functions:

Giry monad

We can consider Markov kernels to be generalised measurable functions:

- Every "normal" $f : \mathcal{X} \to \mathcal{Y}$ can be canonically identified with $\delta_{\mathcal{Y}} \circ f : \mathcal{X} \to P\mathcal{Y}$, where

$$\begin{aligned} \delta_{\mathcal{Y}} : \mathcal{Y} &\to P\mathcal{Y} \\ y &\mapsto \mathrm{Dirac}(y) \end{aligned}$$

Giry monad

We can consider Markov kernels to be generalised measurable functions:

- Every "normal" $f : \mathcal{X} \to \mathcal{Y}$ can be canonically identified with $\delta_{\mathcal{Y}} \circ f : \mathcal{X} \to P\mathcal{Y}$, where

$$\begin{aligned} \delta_{\mathcal{Y}} : \mathcal{Y} &\to P\mathcal{Y} \\ y &\mapsto \mathrm{Dirac}(y) \end{aligned}$$

- Every "generalised generalised" function $k : \mathcal{X} \to PP\mathcal{Y}$ can be canonically identified with $E_{\mathcal{Y}} \circ k : \mathcal{X} \to P\mathcal{Y}$, where

$$\begin{aligned} E_{\mathcal{Y}} : PP\mathcal{Y} &\to P\mathcal{Y} \\ p &\mapsto \int_{P\mathcal{Y}} p(\mathrm{d}q)\, q(-) \end{aligned}$$

Giry monad

We can consider Markov kernels to be generalised measurable functions:

- Every "normal" $f : \mathcal{X} \to \mathcal{Y}$ can be canonically identified with $\delta_{\mathcal{Y}} \circ f : \mathcal{X} \to P\mathcal{Y}$, where

$$\begin{aligned} \delta_{\mathcal{Y}} : \mathcal{Y} &\to P\mathcal{Y} \\ y &\mapsto \mathrm{Dirac}(y) \end{aligned}$$

- Every "generalised generalised" function $k : \mathcal{X} \to PP\mathcal{Y}$ can be canonically identified with $E_{\mathcal{Y}} \circ k : \mathcal{X} \to P\mathcal{Y}$, where

$$\begin{aligned} E_{\mathcal{Y}} : PP\mathcal{Y} &\to P\mathcal{Y} \\ p &\mapsto \int_{P\mathcal{Y}} p(\mathrm{d}q)\, q(-) \end{aligned}$$

P, $\delta_{\mathcal{Y}}$, and $E_{\mathcal{Y}}$ moreover satisfy coherence conditions and so give rise to a monad structure on **Meas**

Kleisli composition

The monad structure on **Meas** yields a canonical notion of composition of generalised functions (i.e. Markov kernels)

Kleisli composition

The monad structure on **Meas** yields a canonical notion of composition of generalised functions (i.e. Markov kernels)

Given $k : \mathcal{X} \to P\mathcal{Y}$ and $\ell : \mathcal{Y} \to P\mathcal{Z}$, define $\ell \circ_{\mathsf{kl}} k : \mathcal{X} \to P\mathcal{Z}$ via the following composition:

$$\mathcal{X} \xrightarrow{k} P\mathcal{Y} \xrightarrow{P\ell} PP\mathcal{Z} \xrightarrow{E_{\mathcal{Z}}} P\mathcal{Z}$$

Kleisli composition

The monad structure on **Meas** yields a canonical notion of composition of generalised functions (i.e. Markov kernels)

Given $k : \mathcal{X} \to P\mathcal{Y}$ and $\ell : \mathcal{Y} \to P\mathcal{Z}$, define $\ell \circ_{\mathsf{kl}} k : \mathcal{X} \to P\mathcal{Z}$ via the following composition:

$$\mathcal{X} \xrightarrow{k} P\mathcal{Y} \xrightarrow{P\ell} PP\mathcal{Z} \xrightarrow{E_{\mathcal{Z}}} P\mathcal{Z}$$

Can show this is the usual Chapman-Kolmogorov equation:

$$(\ell \circ_{\mathsf{kl}} k)(x)(A) = \int_{\mathcal{Y}} k(x)(\mathrm{d}y)\, \ell(y)(A) \qquad \text{where } A \in \Sigma_{\mathcal{Z}}$$

Kleisli composition

The monad structure on **Meas** yields a canonical notion of composition of generalised functions (i.e. Markov kernels)

Given $k : \mathcal{X} \to P\mathcal{Y}$ and $\ell : \mathcal{Y} \to P\mathcal{Z}$, define $\ell \circ_{\mathsf{kl}} k : \mathcal{X} \to P\mathcal{Z}$ via the following composition:

$$\mathcal{X} \xrightarrow{k} P\mathcal{Y} \xrightarrow{P\ell} PP\mathcal{Z} \xrightarrow{E_{\mathcal{Z}}} P\mathcal{Z}$$

Can show this is the usual Chapman-Kolmogorov equation:

$$(\ell \circ_{\mathsf{kl}} k)(x)(A) = \int_{\mathcal{Y}} k(x)(\mathrm{d}y)\, \ell(y)(A) \qquad \text{where } A \in \Sigma_{\mathcal{Z}}$$

Dirac maps $\delta_{\mathcal{X}} : \mathcal{X} \to P\mathcal{X}, x \mapsto \mathsf{Dirac}(x)$ behave like identities

Kleisli category

This gives rise to the Kleisli category of **Meas**, known as **Stoch**:

	Meas	**Stoch**
Objects	Measurable spaces	Measurable spaces
Arrows	Measurable functions	Markov kernels
Composition	Composition of functions	Chapman-Kolmogorov
Identities	Identity functions	Dirac maps

Kleisli adjunction

We have a bijective correspondence (in fact an adjunction):

Markov kernels $\mathcal{X} \to \mathcal{Y}$ $\quad\leftrightsquigarrow\quad$ Measurable functions $\mathcal{X} \to P\mathcal{Y}$

Kleisli adjunction

We have a bijective correspondence (in fact an adjunction):

$$\text{Markov kernels } \mathcal{X} \to \mathcal{Y} \qquad \leftrightsquigarrow \qquad \text{Measurable functions } \mathcal{X} \to P\mathcal{Y}$$

We saw that identity kernels correspond to Dirac maps, i.e.

$$\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X} \qquad \leftrightsquigarrow \qquad \delta_{\mathcal{X}} : \mathcal{X} \to P\mathcal{X}$$

Kleisli adjunction

We have a bijective correspondence (in fact an adjunction):

Markov kernels $\mathcal{X} \to \mathcal{Y}$ $\quad\leftrightsquigarrow\quad$ Measurable functions $\mathcal{X} \to P\mathcal{Y}$

We saw that identity kernels correspond to Dirac maps, i.e.

$$\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X} \qquad \leftrightsquigarrow \qquad \delta_{\mathcal{X}} : \mathcal{X} \to P\mathcal{X}$$

Interesting question: what Markov kernel corresponds to the measurable function $\mathrm{id}_{P\mathcal{Y}} : P\mathcal{Y} \to P\mathcal{Y}$?

$$\leftrightsquigarrow \qquad \mathrm{id}_{P\mathcal{Y}} : P\mathcal{Y} \to P\mathcal{Y}$$

Kleisli adjunction

We have a bijective correspondence (in fact an adjunction):

$$\text{Markov kernels } \mathcal{X} \to \mathcal{Y} \qquad \leftrightsquigarrow \qquad \text{Measurable functions } \mathcal{X} \to P\mathcal{Y}$$

We saw that identity kernels correspond to Dirac maps, i.e.

$$\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X} \qquad \leftrightsquigarrow \qquad \delta_{\mathcal{X}} : \mathcal{X} \to P\mathcal{X}$$

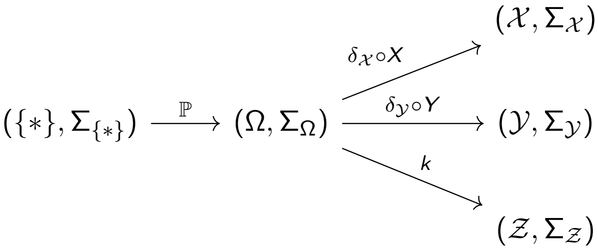
Interesting question: what Markov kernel corresponds to the measurable function $\mathrm{id}_{P\mathcal{Y}} : P\mathcal{Y} \to P\mathcal{Y}$?

$$\mathsf{samp}_{\mathcal{Y}} : P\mathcal{Y} \to \mathcal{Y} \qquad \leftrightsquigarrow \qquad \mathrm{id}_{P\mathcal{Y}} : P\mathcal{Y} \to P\mathcal{Y}$$

Here $\mathsf{samp}_{\mathcal{Y}}(p)(B) = p(B)$, i.e. $\mathsf{samp}_{\mathcal{Y}}$ draws a sample from its input

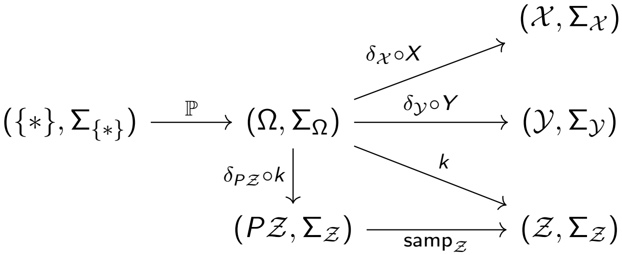
New picture

Stoch unifies and generalises the elements in our original picture:



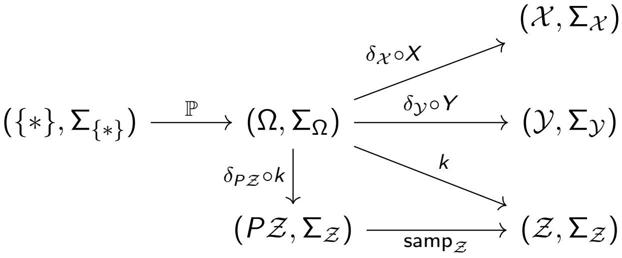
New picture

Stoch unifies and generalises the elements in our original picture:



New picture

Stoch unifies and generalises the elements in our original picture:



Although to some extent (Ω, Σ_Ω) is redundant now ...

Return to case study

Recap

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \, . \tag{14}$$

Conditional distributions/disintegrations

Proposition

If $\mathcal{Y}$ is standard Borel, then for any distribution p on $\mathcal{X} \otimes \mathcal{Y}$, there exists a Markov kernel $k : \mathcal{X} \to \mathcal{Y}$ such that

$$p(A \times B) = \int_A \mathrm{proj}_{\mathcal{X}}(p)(\mathrm{d}x)\, k(x)(B) \qquad \text{for all } A \in \Sigma_{\mathcal{X}} \text{ and } B \in \Sigma_{\mathcal{Y}}.$$

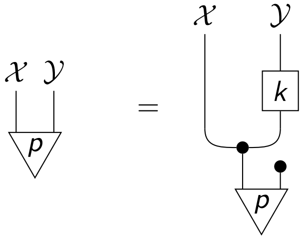
Conditional distributions/disintegrations

Proposition

If $\mathcal{Y}$ is standard Borel, then for any distribution p on $\mathcal{X} \otimes \mathcal{Y}$, there exists a Markov kernel $k : \mathcal{X} \to \mathcal{Y}$ such that

$$p(A \times B) = \int_A \mathrm{proj}_{\mathcal{X}}(p)(\mathrm{d}x)\, k(x)(B) \qquad \textit{for all } A \in \Sigma_{\mathcal{X}} \textit{ and } B \in \Sigma_{\mathcal{Y}}.$$

It is convenient to have a graphical way to denote this. Standard commutative diagrams get complex, but string diagrams work:



(Read from bottom to top)

Informal usage

We use these informally all the time already, e.g. [Vaswani et al., 2017]:

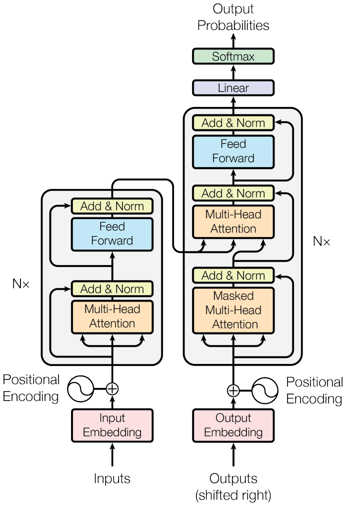


Figure 1: The Transformer - model architecture.

Invariance under an equivalence relation

Suppose $\sim$ is an arbitrary equivalence relation on $\mathcal{X}$

Invariance under an equivalence relation

Suppose $\sim$ is an arbitrary equivalence relation on $\mathcal{X}$

Definition

A distribution p on $\mathcal{X} \otimes \mathcal{Y}$ is conditionally $\sim$-invariant if p admits a disintegration $k : \mathcal{X} \to \mathcal{Y}$ that is $\sim$-invariant, i.e. $k(x) = k(x')$ if $x \sim x'$.

Invariance under an equivalence relation

Suppose $\sim$ is an arbitrary equivalence relation on $\mathcal{X}$

Definition

A distribution p on $\mathcal{X} \otimes \mathcal{Y}$ is conditionally $\sim$-invariant if p admits a disintegration $k : \mathcal{X} \to \mathcal{Y}$ that is $\sim$-invariant, i.e. $k(x) = k(x')$ if $x \sim x'$.

For $p = \mathrm{Law}[X, Y]$, equivalent to conditional invariance in sense of Bloem-Reddy and Teh [2020] under their setup, i.e. $\mathcal{G}$ is compact, $\mathrm{Law}[X]$ is $\mathcal{G}$-invariant, $\mathcal{Y}$ standard Borel, and

$$x \sim x' \Leftrightarrow x = g \cdot x' \text{ for some } g \in \mathcal{G},$$

Invariance under an equivalence relation

Suppose $\sim$ is an arbitrary equivalence relation on $\mathcal{X}$

Definition

A distribution p on $\mathcal{X} \otimes \mathcal{Y}$ is conditionally $\sim$-invariant if p admits a disintegration $k : \mathcal{X} \to \mathcal{Y}$ that is $\sim$-invariant, i.e. $k(x) = k(x')$ if $x \sim x'$.

For $p = \mathrm{Law}[X, Y]$, equivalent to conditional invariance in sense of Bloem-Reddy and Teh [2020] under their setup, i.e. $\mathcal{G}$ is compact, $\mathrm{Law}[X]$ is $\mathcal{G}$-invariant, $\mathcal{Y}$ standard Borel, and

$$x \sim x' \Leftrightarrow x = g \cdot x' \text{ for some } g \in \mathcal{G},$$

Makes sense more generally – could even start with k as the definition of a (probabilistic) neural network

Quotient spaces

Given any measurable space $\mathcal{X}$ and an equivalence relation $\sim$ on $\mathcal{X}$, we can form the quotient space $\mathcal{X}/\sim$ of equivalence classes under $\sim$

Quotient spaces

Given any measurable space $\mathcal{X}$ and an equivalence relation $\sim$ on $\mathcal{X}$, we can form the quotient space $\mathcal{X}/\sim$ of equivalence classes under $\sim$

The σ-algebra is final with respect to the quotient map $q : \mathcal{X} \to \mathcal{X}/\sim$

Quotient spaces

Given any measurable space $\mathcal{X}$ and an equivalence relation $\sim$ on $\mathcal{X}$, we can form the quotient space $\mathcal{X}/\!\sim$ of equivalence classes under $\sim$

The σ-algebra is final with respect to the quotient map $q : \mathcal{X} \to \mathcal{X}/\!\sim$

Explicitly, $\Sigma_{\mathcal{X}/\sim} \coloneqq \{B \subseteq \mathcal{X}/\!\sim \mid q^{-1}(B) \in \Sigma_{\mathcal{X}}\}$.

Universal property of the quotient

Proposition

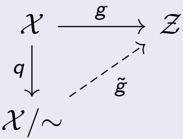
A measurable function $g : \mathcal{X} \to \mathcal{Z}$ is $\sim$-invariant iff there exists a (necessarily unique) measurable function $\tilde{g} : \mathcal{X}/\sim \to \mathcal{Z}$ such that $\tilde{g} \circ q = g$, i.e. the following diagram commutes:

$$
\begin{array}{ccc}
\mathcal{X} & \xrightarrow{\;g\;} & \mathcal{Z} \\
\Big\downarrow q & \nearrow_{\tilde{g}} & \\
\mathcal{X}/\sim & &
\end{array}
$$

Universal property of the quotient

Proposition

A measurable function $g : \mathcal{X} \to \mathcal{Z}$ is $\sim$-invariant iff there exists a (necessarily unique) measurable function $\tilde{g} : \mathcal{X}/{\sim} \to \mathcal{Z}$ such that $\tilde{g} \circ q = g$, i.e. the following diagram commutes:

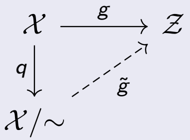


Requires proof, but can do so via only elementary definitions

Universal property of the quotient

Proposition

A measurable function $g : \mathcal{X} \to \mathcal{Z}$ is $\sim$-invariant iff there exists a (necessarily unique) measurable function $\tilde{g} : \mathcal{X}/\sim \to \mathcal{Z}$ such that $\tilde{g} \circ q = g$, i.e. the following diagram commutes:



Requires proof, but can do so via only elementary definitions

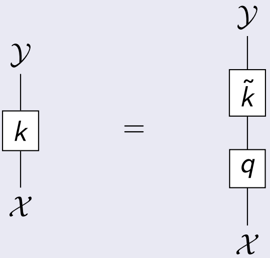
A very natural result in the context of category theory

Invariant kernels via the quotient

Now take $\mathcal{Z} = P\mathcal{Y}$ and interpret within **Stoch**

Corollary

A Markov kernel $k : \mathcal{X} \to \mathcal{Y}$ *is* $\sim$*-invariant iff there exists a Markov kernel* $\tilde{k} : \mathcal{X}/{\sim} \to \mathcal{Y}$ *with*

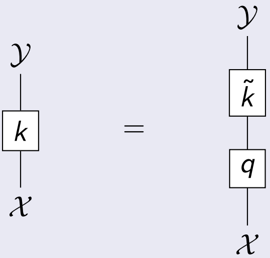


Invariant kernels via the quotient

Now take $\mathcal{Z} = P\mathcal{Y}$ and interpret within **Stoch**

Corollary

A Markov kernel $k : \mathcal{X} \to \mathcal{Y}$ is $\sim$-invariant iff there exists a Markov kernel $\tilde{k} : \mathcal{X}/\sim \to \mathcal{Y}$ with

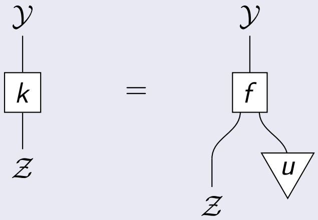


(Note that we are identifying q with its lifted version $\delta_{\mathcal{X}/\sim} \circ q$)

Noise outsourcing

Proposition

For any Markov kernel $k : \mathcal{Z} \to \mathcal{Y}$ *with* $\mathcal{Y}$ *standard Borel, there exists a measurable function* $f : \mathcal{Z} \otimes [0, 1] \to \mathcal{Y}$ *such that*

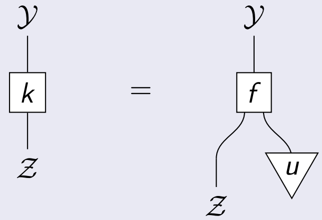


where $u = \mathrm{Uniform}(0, 1)$.

Noise outsourcing

Proposition

For any Markov kernel $k : \mathcal{Z} \to \mathcal{Y}$ with $\mathcal{Y}$ standard Borel, there exists a measurable function $f : \mathcal{Z} \otimes [0,1] \to \mathcal{Y}$ such that



where $u = \mathrm{Uniform}(0,1)$.

Standard result (e.g. Lemma 3.22 of Kallenberg [2002])

Combining these results

Proposition

If $\mathcal{Y}$ is standard Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/\!\sim \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \stackrel{\mathrm{d}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1),\ \eta \perp\!\!\!\perp X$$

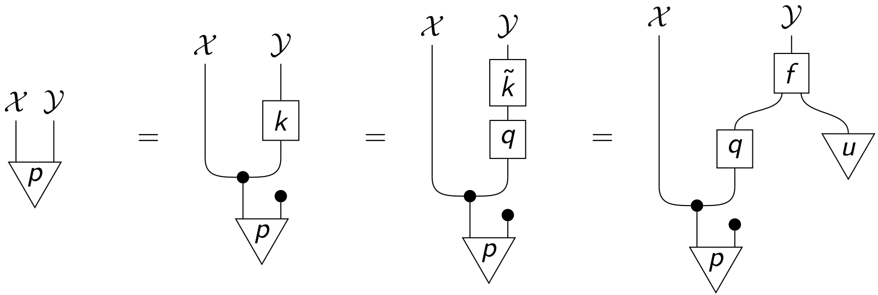
Combining these results

Proposition

If $\mathcal{Y}$ is standard Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/{\sim} \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \stackrel{\mathrm{d}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1),\ \eta \perp\!\!\!\perp X$$

Proof: writing $p := \mathrm{Law}[X, Y]$, conditional $\sim$-invariance implies



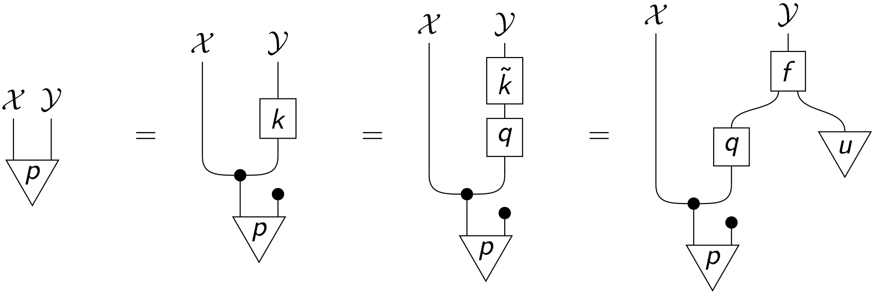
Combining these results

Proposition

If $\mathcal{Y}$ is standard Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/\sim \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \stackrel{\mathrm{d}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1),\ \eta \perp\!\!\!\perp X$$

Proof: writing $p := \mathrm{Law}[X, Y]$, conditional $\sim$-invariance implies



Conversely, right-hand side is conditionally $\sim$-invariant since q is.

Comparison with original result

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \text{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \,. \tag{14}$$

Comparison with original result

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

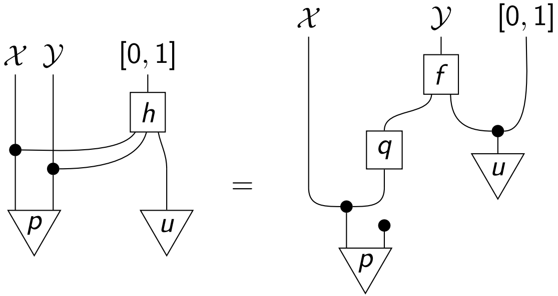
$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \,. \tag{14}$$

Not quite done:

$$(X, Y) \overset{\mathrm{d}}{=} (X, f(q(X), \eta)) \qquad \not\Rightarrow \qquad Y \overset{\text{a.s.}}{=} f(q(X), \eta)$$

Completing the proof

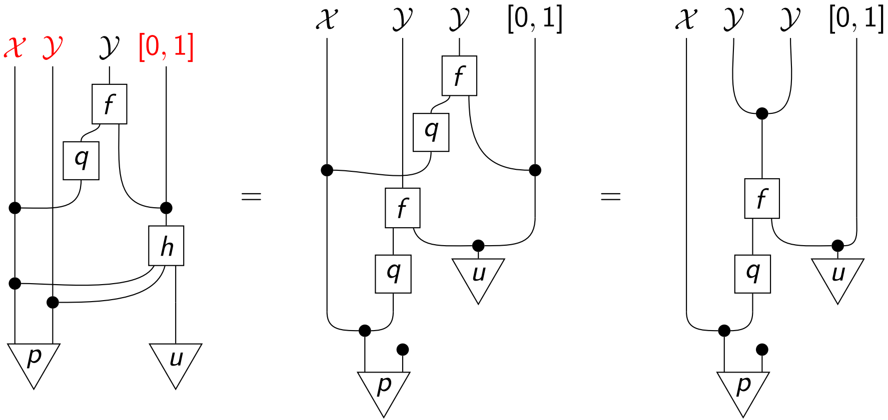
Choose $h : \mathcal{X} \otimes \mathcal{Y} \otimes [0,1] \to [0,1]$ such that



Existence of h follows by disintegrating right-hand side along $\mathcal{X} \times \mathcal{Y}$ and then applying noise outsourcing result

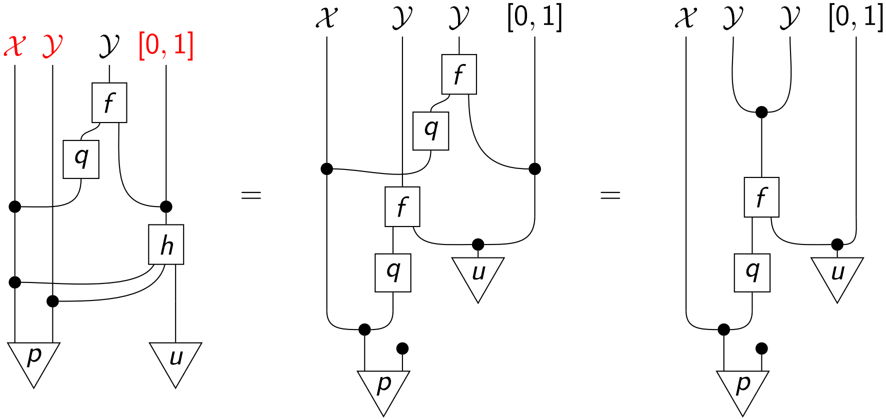
Completing the proof

Now affix the same (q, f) construction to both sides:



Completing the proof

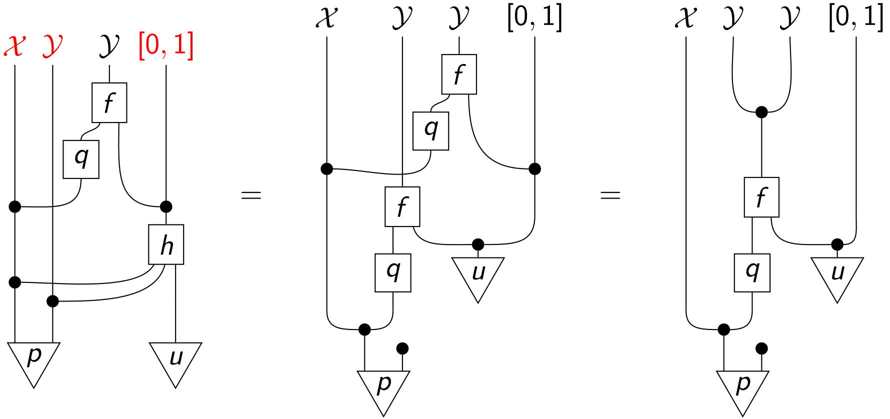
Now affix the same (q, f) construction to both sides:



$\Rightarrow$ If $\xi \sim \mathrm{Uniform}(0,1)$ with $\xi \perp\!\!\!\perp (X, Y)$, then letting $\eta := h(X, Y, \xi)$, have $(X, Y, f(q(X), \eta), \eta) \overset{\mathrm{d}}{=} (X, f(q(X), \xi), f(q(X), \xi), \xi)$

Completing the proof

Now affix the same (q, f) construction to both sides:



$\Rightarrow$ If $\xi \sim \mathrm{Uniform}(0,1)$ with $\xi \perp\!\!\!\perp (X, Y)$, then letting $\eta := h(X, Y, \xi)$, have $(X, Y, f(q(X), \eta), \eta) \overset{\mathrm{d}}{=} (X, f(q(X), \xi), f(q(X), \xi), \xi)$

$\Rightarrow Y \overset{\mathrm{a.s.}}{=} f(q(X), \eta)$ and $\eta \overset{\mathrm{d}}{=} \xi \sim \mathrm{Uniform}(0,1)$ with $\eta \perp\!\!\!\perp X$

Combining these results

Theorem 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \text{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X\ . \tag{14}$$

Combining these results

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \text{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \, . \tag{14}$$

Theorem (Our version)

If $\mathcal{Y}$ is Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/\!\sim \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1) \textit{ with } \eta \perp\!\!\!\perp X$$

Combining these results

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \ . \tag{14}$$

Theorem (Our version)

If $\mathcal{Y}$ is Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/{\sim} \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1) \textit{ with } \eta \perp\!\!\!\perp X$$

(More precisely, both statements should refer to the existence of an extension of the underlying probability space that admits suitable choices of η and f)

Combining these results

THEOREM 7. *Let X and Y be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

(14) $$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X .$$

Theorem (Our version)

If $\mathcal{Y}$ is Borel, then $\mathrm{Law}[X, Y]$ is conditionally $\sim$-invariant iff there exists a measurable function $f : \mathcal{X}/\!\sim \otimes [0,1] \to \mathcal{Y}$ such that

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(q(X), \eta)) \qquad \textit{where } \eta \sim \mathrm{Uniform}(0,1) \textit{ with } \eta \perp\!\!\!\perp X$$

(More precisely, both statements should refer to the existence of an extension of the underlying probability space that admits suitable choices of η and f)

Possibly better to express entirely via Markov kernels

Conclusion

Summary and Outlook

Categorical probability offers a high-level perspective on the classical theory that makes abstraction easier and helps theory follow intuition

Summary and Outlook

Categorical probability offers a high-level perspective on the classical theory that makes abstraction easier and helps theory follow intuition

The outlook is very positive:

- Lots of activity in categorical probability, e.g. Perrone [2018], Cho and Jacobs [2019], Jacobs [2019], Fritz [2020], Moss and Perrone [2023], Perrone [2023]

Summary and Outlook

Categorical probability offers a high-level perspective on the classical theory that makes abstraction easier and helps theory follow intuition

The outlook is very positive:

- Lots of activity in categorical probability, e.g. Perrone [2018], Cho and Jacobs [2019], Jacobs [2019], Fritz [2020], Moss and Perrone [2023], Perrone [2023]
- Category theory has been hugely successful elsewhere, e.g. pure maths, computer science, quantum mechanics

Summary and Outlook

Categorical probability offers a high-level perspective on the classical theory that makes abstraction easier and helps theory follow intuition

The outlook is very positive:

- Lots of activity in categorical probability, e.g. Perrone [2018], Cho and Jacobs [2019], Jacobs [2019], Fritz [2020], Moss and Perrone [2023], Perrone [2023]
- Category theory has been hugely successful elsewhere, e.g. pure maths, computer science, quantum mechanics

The programming language has been (increasingly) written – now is time for practitioners to write new software

Upcoming work

Equivariant stochastic neural networks in Markov categories

Rob Cornish

Bart Jacobs. Structured probabilistic reasoning. 2019.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *J. Mach. Learn. Res.*, 21:90–1, 2020.

Michele Giry. A categorical approach to probability theory. In *Categorical aspects of topology and analysis*, pages 68–85. Springer, 1982.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Olav Kallenberg. *Foundations of modern probability*. Springer, 2 edition, 2002.

Paolo Perrone. Categorical probability and stochastic dominance in metric spaces, 2018.

References II

Kenta Cho and Bart Jacobs. Disintegration and bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, mar 2019. doi: 10.1017/s0960129518000488. URL https://doi.org/10.1017%2Fs0960129518000488.

Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, aug 2020. doi: 10.1016/j.aim.2020.107239. URL https://doi.org/10.1016%2Fj.aim.2020.107239.

Sean Moss and Paolo Perrone. A category-theoretic proof of the ergodic decomposition theorem. *Ergodic Theory and Dynamical Systems*, 43(12): 4166–4192, February 2023. ISSN 1469-4417. doi: 10.1017/etds.2023.6. URL http://dx.doi.org/10.1017/etds.2023.6.

Paolo Perrone. Markov categories and entropy, 2023.