

# Foundations for Type-Driven Probabilistic Modelling

Ohad Kammar  
University of Edinburgh

Laboratory for Software Science  
9 December, 2025  
Chair of Programming Languages  
Institute of Computer Science  
University of Tartu



THE UNIVERSITY OF EDINBURGH  
**informatics** **lfcs**

Laboratory for Foundations  
of Computer Science



supported by:



THE ROYAL  
SOCIETY

The  
Alan Turing  
Institute

Facebook Research  
NCSC

logic-rich & type-rich computation

statistical computation

## logic-rich & type-rich computation

- ▶ Expressive type systems: Haskell, OCaml, Rust, Agda, Idris
- ▶ Mechanised mathematics: Agda, Rocq, Isabelle/HOL, Lean
- ▶ Verification: SMT-powered real-world systems

## statistical computation

Generative modelling with efficient inference: Monte-Carlo simulation or gradient-based optimisation

# This course

## Typed interface to probability/statistics

Every concept has:

- ▶ a type
- ▶ associated operations
- ▶ properties in terms of these operations.



course page

## Two implementations/models

**discrete model**

familiar maths  
introductory



**full model**

supports discrete  
and  
continuous distributions  
same language

# Motivation: why foundations?

## discrete probability

countably supported distributions

good type-structure

(this course)

## continuous probability

Lebesgue measure over  $\mathbb{R}^n$

## measure theory

standard, established

poor type-structure

## well-behaved probability

s-finite distributions

over standard Borel spaces

## quasi-Borel spaces

new, experimental

rich type-structure

(this course)

## Takeaway

Use types to abstract away from the model

# Motivation: why types?

- ▶ **spotlights** meaningful operations

$$\int : (\text{Distribution } X) \times (\text{RandomVariable } X) \rightarrow [0, \infty]$$

- ▶ document **intent**:

probability ( $\text{Distribution } X$ ) vs. density ( $X \rightarrow [0, \infty]$ ) vs. random variable

- ▶ succinctness: omit and elaborate details
- ▶ especially **formal** types, allow using theory correctly without fully understanding it

# Lecture plan

## Lecture 1: discrete model (today)

- ▶ Motivation
- ▶ Language of probability and distribution
- ▶ Discrete model
- ▶ Simply-typed probability
- ▶ Dependently-typed probability



course page

## Lecture 2: the full model

- ▶ Borel sets and measurable spaces
- ▶ Quasi-Borel spaces
- ▶ Type structure & standard Borel spaces
- ▶ Integration & random variables



ask questions on the  
Scottish PL Institute  
Zulip stream #qbs

# Language of **probability** & **distribution**

$X$  type (=space) of **values/outcomes**

$DX$  type of **distributions/measures** over  $X$

$PX \subseteq DX$  sub-type of **probability distributions** over  $X$

$\mathcal{B}_X \subseteq \mathcal{P}X$  type of **events**: subsets we wish to measure

$\mathbb{W}$  type of **weights**: values in  $[0, \infty]$

$\int, \mathbb{E}$  Lebesgue integration and the expectation operation

Type judgements describe well-formed values/outcomes of a given type, e.g.:

$$\mu : DX, E : \mathcal{B}_X \vdash \text{Ce}_{\mu}[E] : \mathbb{W}$$

(measures weight  $\text{Ce}_{\mu}[E]$  of event  $E$  according to distribution  $\mu$ )

Propositions describe properties of well-formed values/outcomes of a given type, e.g.:

$$y_1, y_2 : Y \vdash y_1 \stackrel{Y}{=} y_2 : \text{Prop} \quad \mu : PX, E : \mathcal{B}_X \vdash \text{Pr}_{\mu}[E] = \text{Ce}_{\mu}[E]$$

(probability of event according to probability distribution is its measure)



# Axioms for events and distributions

Empty event

$$\emptyset : \mathcal{B}_X$$

Empty events weight zero

$$\mu : \mathcal{D}X \vdash \text{Ce}_{\mu}[\emptyset] = 0$$

# Axioms for events and distributions

## Boolean Sub-algebra of Events

$E : \mathcal{B}_X \vdash E^c : \mathcal{B}_X$        $E, F : \mathcal{B}_X \vdash E \cap F : \mathcal{B}_X$     so also:  $E, F : \mathcal{B}_X \vdash X, E \cup F : \mathcal{B}_X$

## Disjoint additivity

$w, v : \mathbb{W} \vdash w + v : \mathbb{W}$        $E, C : \mathcal{B}_X, \mu : \mathbf{D}X \vdash \text{Ce}_{\mu}[E] = \text{Ce}_{\mu}[E \cap C] + \text{Ce}_{\mu}[E \cap C^c]$

# Axioms for events and distributions

## Boolean Sub-algebra of Events

$E : \mathcal{B}_X \vdash E^c : \mathcal{B}_X$      $E, F : \mathcal{B}_X \vdash E \cap F : \mathcal{B}_X$     so also:  $E, F : \mathcal{B}_X \vdash X, E \cup F : \mathcal{B}_X$

## Disjoint additivity

$w, v : \mathbb{W} \vdash w + v : \mathbb{W}$      $E, C : \mathcal{B}_X, \mu : \mathcal{D}X \vdash \text{Ce}_\mu[E] = \text{Ce}_\mu[E \cap C] + \text{Ce}_\mu[E \cap C^c]$

## Exercise

Derive 'axiomatically' that:

- ▶ measurement is **monotone**:

$$\mu : \mathcal{D}X, E \subseteq F \vdash \text{Ce}_\mu[E] \leq \text{Ce}_\mu[F]$$

- ▶ the **inclusion-exclusion** principle:

$$\mu : \mathcal{D}X, E, F : \mathcal{B}_X \vdash \text{Ce}_\mu[E \cup F] + \text{Ce}_\mu[E \cap F] = \text{Ce}_\mu[E] + \text{Ce}_\mu[F]$$

# Axioms for events and distributions

Consider posets:

$$\omega := (\mathbb{N}, \leq) \quad (\mathcal{B}_X, \subseteq) \quad (\mathbb{W}, \leq)$$

$\omega$ -chains in a poset  $P = (\underline{P}, \leq)$ :

$$P^\omega := \{p_- \in \underline{P}^{\mathbb{N}} \mid p_0 \leq p_1 \leq \dots\}$$

Chain-closure of events and weights

$$E_- : (\mathcal{B}_X, \subseteq)^\omega \vdash \bigcup_n E_n : \mathcal{B}_X \quad w_- : (\mathbb{W}, \leq)^\omega \vdash \sup_n w_n : \mathbb{W}$$

Scott-continuity of measurement

$$E_- : (\mathcal{B}_X, \subseteq)^\omega, \mu : \mathbf{D}X \vdash \mathbf{Ce}_\mu [\bigcup_n E_n] = \sup_n \mathbf{Ce}_\mu [E_n]$$

# Axiom for probability

Probability distributions have total mass one

$$\mathbf{PX} := \{\mu \in \mathbf{DX} \mid \mathbf{Ce}_\mu[X] = 1\} \quad \mu : \mathbf{PX} \vdash \mathbf{cast} \mu : \mathbf{DX}$$

i.e., if we define:

$$\mathbb{I} := [0,1] \quad \mu : \mathbf{PX}, E : \mathcal{B}_X \vdash \mathbf{Pr}_\mu[E] := \mathbf{Ce}_{\mathbf{cast} \mu}[E] : \mathbb{I}$$

then:

$$\mu : \mathbf{PX} \vdash \mathbf{Pr}_\mu[X] = 1$$

# Integration

## Lebesgue integration w.r.t. a distribution

$$\mu : \mathsf{DX}, f : \mathbb{W}^X \vdash \int \mu(\mathrm{d}x) f(x) : \mathbb{W}$$

(NB: We succinctly write  $\mathbb{W}^X$  for the type of functions  $X \rightarrow \mathbb{W}$ .)

## Expectation w.r.t. a probability distribution

$$\mu : \mathsf{PX}, f : \mathbb{W}^X \vdash \mathbb{E}_{x \sim \mu} [f(x)] := \int (\mathsf{cast} \mu)(\mathrm{d}x) f(x) : \mathbb{W}$$

We'll use variations on this notation, e.g.:

$$\int \mathrm{d}\mu f, \int f \mathrm{d}\mu, \int f(x) \mu(\mathrm{d}x), \mathbb{E}_\mu [f]$$

# Summary

Have: Language and (some) axioms

Want: Model

Today: **discrete** model

Next week: **full** model

# Lecture plan

## Lecture 1: discrete model (today)

- ▶ Motivation
- ▶ Language of probability and distribution
- ▶ Discrete model
- ▶ Simply-typed probability
- ▶ Dependently-typed probability



course page

## Lecture 2: the full model

- ▶ Borel sets and measurable spaces
- ▶ Quasi-Borel spaces
- ▶ Type structure & standard Borel spaces
- ▶ Integration & random variables



ask questions on the  
Scottish PL Institute  
Zulip stream #qbs



$X$ : types denote **sets**

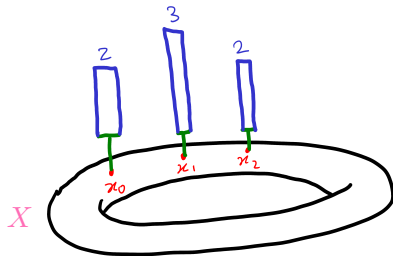
$\mathcal{D}X$ : set of **histograms**:

# Discrete model

$X$ : types denote **sets**

$\mathcal{D}X$ : set of **histograms**:

$$\mathcal{D}X := \{\mu : X \rightarrow \mathbb{W} \mid \mu \text{ is countably supported (next slide)}\}$$



$$\mu x_0 = 2 \quad \mu x_1 = 3 \quad \mu x_2 = 2$$

# Countably supported distributions

## Support

A subset  $S$  **supports** a weight function  $\mu : X \rightarrow \mathbb{W}$  when  $\mu$  is 0 outside  $S$ :

$$\mu : \mathbb{W}^X, S : \mathcal{P}X \vdash S \text{ supports } \mu := (\forall x : X. (\mu x > 0) \implies x \in S) : \text{Prop}$$

The subsets supporting a weight function  $\mu$  are closed under intersections.

$\implies$  There is a smallest supporting subset, called the **support** of  $\mu$ :

$$\mu : \mathbb{W}^X \vdash \text{supp } \mu := \{x \in X \mid \mu x > 0\}$$

# Discrete model

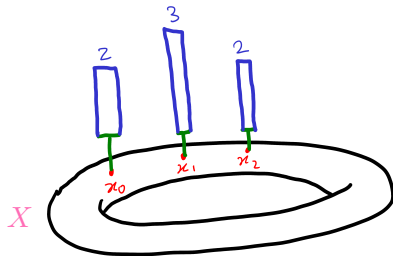
$X$ : types denote **sets**

$\mathcal{D}X$ : set of **histograms**:

$$\mathcal{D}X := \{ \mu : X \rightarrow \mathbb{W} \mid \mu \text{ is countably supported} \}$$

$$:= \{ \mu : X \rightarrow \mathbb{W} \mid \exists S \in \mathcal{P}X. S \text{ is countable} \}$$

$$:= \{ \mu : X \rightarrow \mathbb{W} \mid \text{supp } \mu \text{ is countable} \}$$



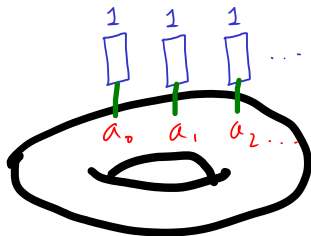
$$\mu x_0 = 2 \quad \mu x_1 = 3 \quad \mu x_2 = 2$$

# Example distributions

## Counting distribution

Counts the outcomes in a countable subset:

$$S : \mathcal{P}_{\text{ctbl}} X \vdash \#_S := \left( \lambda x. \begin{cases} x \in S : 1 \\ x \notin S : 0 \end{cases} \right) : \mathsf{D}X$$

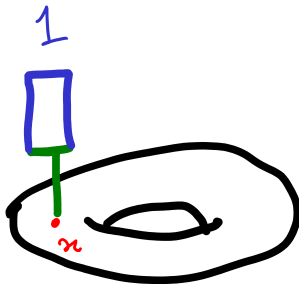


# Example distributions

## Dirac

A point mass:

$$x : X \vdash \delta_x := \left( \lambda x'. \begin{cases} x' = x : 1 \\ x' \neq x : 0 \end{cases} \right) : \mathsf{D}X$$



(NB:  $x : X \vdash \delta_x = \#_{\{x\}}.$ )

# Example distributions

## Zero

No mass anywhere:

$$\vdash \mathbf{0} := \underline{0} := (\lambda x.0) : \mathbf{D}X$$

$$(\text{NB: } \vdash \mathbf{0} = \#_{\emptyset}.)$$

# Discrete model

$X$ : types denote **sets**

$\mathsf{D}X$ : set of **histograms**:

$$\mathsf{D}X := \{ \mu : X \rightarrow \mathbb{W} \mid \mu \text{ is countably supported} \}$$

$\mathcal{B}_X$ : **every subset** can be measured:

$$\mathcal{B}_X := \mathcal{P}X$$

**Measurement**: weighted sum of all (supported) outcomes:

$$\begin{aligned} \mu : \mathsf{D}X, E : \mathcal{B}_X \vdash \mathsf{Ce}_\mu[E] &:= \sum_{x \in E} \mu x \\ &:= \sum_{x \in E \cap \text{supp } \mu} \mu x \end{aligned}$$

$$\text{NB: } \mu : \mathsf{D}X, E : \mathcal{B}_X, S : \mathcal{P}_{\text{ctbl}}X, S \text{ supports } \mu \vdash \mathsf{Ce}_\mu[E] = \sum_{x \in E \cap S} \mu x.$$



# Example measurements

(NB:  $\mu : \mathcal{D}X$ ,  $E : \mathcal{B}_X$ ,  $S : \mathcal{P}_{\text{ctbl}}X$ ,  $S$  supports  $\mu \vdash \text{Ce}_\mu[E] = \sum_{x \in E \cap S} \mu x$ .)

## Counting distribution

counts supported outcomes

$$S : \mathcal{P}_{\text{ctbl}}X, E : \mathcal{B}_X \vdash \underset{\#_S}{\text{Ce}}[E] = |E \cap S| := \begin{cases} E \cap S \text{ has } n \in \mathbb{N} \text{ elements:} & n \\ E \cap S \text{ is infinite:} & \infty \end{cases}$$

# Example measurements

(NB:  $\mu : \mathcal{D}X, E : \mathcal{B}_X, S : \mathcal{P}_{\text{ctbl}}X, S \text{ supports } \mu \vdash \text{Ce}_\mu[E] = \sum_{x \in E \cap S} \mu x.$ )

## Counting distribution

counts supported outcomes

$$S : \mathcal{P}_{\text{ctbl}}X, E : \mathcal{B}_X \vdash \underset{\#_S}{\text{Ce}}[E] = |E \cap S| := \begin{cases} E \cap S \text{ has } n \in \mathbb{N} \text{ elements:} & n \\ E \cap S \text{ is infinite:} & \infty \end{cases}$$

## Dirac

detects given outcome:

$$x : X, E : \mathcal{B}_X \vdash \text{Ce}_{\delta_x}[E] = \begin{cases} x \in E : & 1 \\ x \notin E : & 0 \end{cases}$$

# Example measurements

(NB:  $\mu : \mathcal{D}X$ ,  $E : \mathcal{B}_X$ ,  $S : \mathcal{P}_{\text{ctbl}}X$ ,  $S$  supports  $\mu \vdash \text{Ce}_\mu[E] = \sum_{x \in E \cap S} \mu x$ .)

## Counting distribution

counts supported outcomes

$$S : \mathcal{P}_{\text{ctbl}}X, E : \mathcal{B}_X \vdash \underset{\#_S}{\text{Ce}}[E] = |E \cap S| := \begin{cases} E \cap S \text{ has } n \in \mathbb{N} \text{ elements:} & n \\ E \cap S \text{ is infinite:} & \infty \end{cases}$$

## Dirac

detects given outcome:

$$x : X, E : \mathcal{B}_X \vdash \text{Ce}_{\delta_x}[E] = \begin{cases} x \in E : & 1 \\ x \notin E : & 0 \end{cases}$$

## Zero

measures every event as zero:

$$E : \mathcal{B}_X \vdash \text{Ce}_0[E] = 0$$

# The discrete model validates the axioms

## Exercise

$$\mu : \mathbf{D} \quad \vdash \text{Ce}_{\mu}[\emptyset] = 0$$

$$E, C : \mathcal{B}_X, \mu : \mathbf{D} \quad \vdash \text{Ce}_{\mu}[E] = \text{Ce}_{\mu}[E \cap C] + \text{Ce}_{\mu}[E \cap C^c]$$

$$E_- : (\mathcal{B}_X, \subseteq)^\omega, \mu : \mathbf{D} x \vdash \text{Ce}_{\mu}\left[\bigcup_n E_n\right] = \sup_n \text{Ce}_{\mu}[E_n]$$

# Parameterised distributions

## Kernel

$k : X \rightsquigarrow Y$  from  $X$  to  $Y$ : function  $k : X \rightarrow \mathsf{D}Y$ .

Kernels are open/parameterised distributions.

## Examples

Dirac and the counting distribution form kernels:

$$\delta_- : X \rightsquigarrow \mathsf{D}X \qquad \#_- : \mathcal{P}_{\text{ctbl}} X \rightsquigarrow \mathsf{D}X$$

NB: This definition is **internal**: when we consider the full model, we will define kernels as those functions internal to the model rather than the set-theoretic functions.

# Action of kernels on distributions

## Kock integral

$$\mu : \mathbf{D}X, k : (\mathbf{D}Y)^X \vdash \oint d\mu k : \mathbf{D}Y$$

This **distribution-valued** integral is implicit in many probability texts. It corresponds to integrating against an arbitrary weight function or random variable.

## Discrete model interpretation

$$\begin{aligned}\oint d\mu k &:= \lambda y. \sum_{x \in X} \mu x \cdot k(x; y) \\ &:= \lambda y. \sum_{x \in \text{supp } \mu} \mu x \cdot k(x; y)\end{aligned}$$

NB1: we write  $k(x; y) := k(x)(y)$  for the uncurried function.

NB2:  $\mu : \mathbf{D}X, k : (\mathbf{D}Y)^X, S : \mathcal{P}_{\text{ctbl}} X, S \text{ supports } \mu \vdash \oint d\mu k = \lambda y. \sum_{x \in S} \mu x \cdot k(x; y)$

# Example

## Weak Disintegration Problem (non-standard terminology)

Input: distributions  $\mu : \mathcal{D}\Theta$ ,  $\nu : \mathcal{D}X$

Output: kernel  $k : \Theta \rightsquigarrow \mathcal{D}X$  such that:  $\nu = \oint d\mu k$ .

Such a **weak disintegration** of  $\nu$  w.r.t.  $\mu$  provides an ‘explanation’ of an observed distribution  $\nu \in \mathcal{D}X$  in terms of a given distribution on parameters  $\mu \in \mathcal{D}\Theta$ . I use the term ‘explanation’ because it explains how the parameters transform into observations.

# Example

## Weak Disintegration Problem (non-standard terminology)

Input: distributions  $\mu : \mathbf{D}\Theta$ ,  $\nu : \mathbf{D}X$

Output: kernel  $k : \Theta \rightsquigarrow \mathbf{D}X$  such that:  $\nu = \oint d\mu k$ .

## Example disintegration

For  $n \in \mathbb{N}$ , write  $\mathbf{Fin} \, n := \{0, \dots, n-1\}$ . For countable  $X$ , write  $\# := \#_X : \mathbf{D}X$ .

Here is a disintegration of  $\# \in \mathbf{D}((\mathbf{Fin} \, 2)^{\mathbf{Fin} \, (n+1)})$  w.r.t.  $\# \in \mathbf{D}(\mathbf{Fin} \, 2)$ :

$$k(x; f) := \begin{cases} f n = x : & 1 \\ \text{otherwise:} & 0 \end{cases} \quad \text{Indeed: } \left( \oint d\# k \right) f = \sum_{b \in \mathbf{Fin} \, 2} \overbrace{\# b}^1 \cdot k(b; f) = k(0; f) + k(1; f)$$

$f : \mathbf{Fin} \, (n+1) \rightarrow \mathbf{Fin} \, 2$  function  
so can take only one value: 0 or 1

$$\downarrow \\ = 1 = \# f$$



# Sub-type of probability distributions

## Sub-types

Given type  $X$  and  $x : X \vdash \varphi : \text{Prop}$ , take the **sub-type** and the **coercion** as follows:

$$\{x : X \mid \varphi\} \subseteq X \quad y : \{x : X \mid \varphi\} \vdash \text{cast } y := y : X$$

we **lift** values in  $X$  that satisfy  $\varphi$  to the sub-type:

$$\frac{\Gamma \vdash M : X \quad \Gamma \vdash \varphi [x \mapsto M]}{\Gamma \vdash \text{lift } M : \{x : X \mid \varphi\}} \quad \frac{\Gamma \vdash M : X \quad \Gamma \vdash \{\varphi\} x \mapsto M}{\Gamma \vdash \text{cast}(\text{lift } M) = M}$$

The axiom implies that  $\text{lift } M$  lifts  $M$  along  $\text{cast}$ . Moreover:

$$y : \{x \in X \mid \varphi\} \vdash \text{lift}(\text{cast } y) = y \quad y : \{x \in X \mid \varphi\} \vdash \varphi [x \mapsto \text{cast } y]$$

i.e., the lifting is unique and elements in the sub-type satisfy  $\varphi$ .

# Sub-type of probability distributions

## Magnitude and probability distributions

$$\mu : DX \vdash \|\mu\| := \text{Ce}_{\mu}[X] : \mathbb{W} \quad PX := \{\mu \in DX \mid \|\mu\| = 1\} \quad \mathbb{I} := [0,1] := \{w \in \mathbb{W} \mid w \leq 1\}$$

## Event probability

$$\mu : PX, E : \mathcal{B}_X \vdash \text{Pr}_{\mu}[E] := \text{lift} \left( \text{Ce}_{\text{cast}_{\mu}}[E] \right) : \mathbb{I}$$

## Stochastic kernel

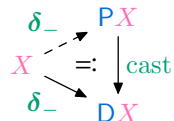
$k : X \rightsquigarrow Y$  from  $X$  to  $Y$ : function  $X \rightarrow PY$ .

NB: in the **discrete model** these distinctions and rules amount to pure pedantry. This pedantry will pay off in the **full model**.

# Lifting Dirac and Kock

## Lemma

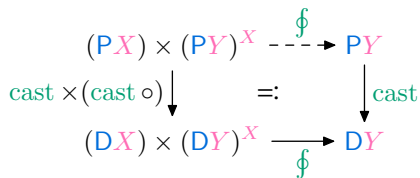
Dirac kernels  $\delta_- : X \rightarrow DX$  lift along  $\text{cast}$ :

$$x : X \vdash \|\delta_x\| = \text{Ce}_{\delta_x}[X] = 1 \quad \text{so we can overload:}$$


Kock integrals of stochastic kernels by probability distributions lift along  $\text{cast}$ :

$$\mu : PX, k : (PY)^X \vdash \text{Ce}_{\oint(\text{cast } \mu)(dx) \text{ cast}(kx)}[Y] = 1$$

so we can overload:



## Proposition

The triple  $(D, \delta_-, \oint)$  forms a monad over **Set**:

$$\begin{array}{ll}
 x : X, k : (DY)^X & \vdash \oint d\delta_x k = k x \\
 \mu : DX & \vdash \oint \mu(dx) \delta_x = \mu \\
 \mu : DX, k : (DY)^X, \ell : (DZ)^Y & \vdash \oint (\oint \mu(dx) k x) (dy) \ell y = \oint \mu(dx) \oint k(x; dy) \ell y
 \end{array}$$

## Corollary

The triple  $(P, \delta_-, \oint)$  forms a monad over **Set**.

# Weighted average

## Lebesgue integral

Integration is the raison d'être for distributions:

$$\mu : \mathbb{D}X, f : \mathbb{W}^X \vdash \int d\mu f : \mathbb{W}$$

In the **discrete model**:

$$\int d\mu f := \sum_{x \in X} (\mu x) \cdot (f x) := \sum_{x \in \text{supp } \mu} (\mu x) \cdot (f x)$$

As usual, replace  $\text{supp } \mu$  by any countable supporting set:

$$\mu : \mathbb{D}X, f : \mathbb{W}^X, S : \mathcal{P}X, S \text{ supports } \mu \vdash \int d\mu f = \sum_{x \in S} (\mu x) \cdot (f x)$$

# Weighted average

## Expectation

To emphasise that some  $\mu$  is a probability distribution, we will use the notation:

$$\mu : \mathbf{P}X, f : \mathbb{W}^X \vdash \quad \mathbb{E}_\mu[f] := \int d(\text{cast } \mu) f : \mathbb{W}$$

When calculating, however, we will usually use  $\int$  and implicitly `cast` any probability distribution to its corresponding distribution.

# Booleans

## Boolean type

The simplest kind of distinguishing outcomes:

$$\mathbb{B} := \{\mathbf{True}, \mathbf{False}\} \quad \frac{\Gamma \vdash M : \mathbb{B} \quad \Gamma \vdash N_1 : X \quad \Gamma \vdash N_2 : X}{\Gamma \vdash \text{if } M \text{ then } N_1 \text{ else } N_2 : X}$$

## Iverson bracket

Lets us replace Boolean propositions with arithmetic expressions:

$$b : \mathbb{B} \vdash [b] := (\text{if } b \text{ then } 1 \text{ else } 0) : \mathbb{W}$$

For example:

$$b : \mathbb{B}, w, v : \mathbb{W} \vdash \text{if } b \text{ then } w \text{ else } v = [b] \cdot w + (1 - [b]) \cdot w$$

# Simplest probabilistic model

## Bernoulli kernel

Single trial succeeding with the given probability:

$$\mathbf{B} : \mathbb{I} \rightsquigarrow \mathbb{B} \quad \mathbf{B}p := \lambda b. \begin{cases} b = \mathbf{True} : & p \\ b = \mathbf{False} : & 1 - p \end{cases}$$

For example, for a payoff of 10 units if the trial succeeds then the expected payoff is:

$$\mathbb{E}_{b \sim \mathbf{B} \frac{1}{4}} [[b] \cdot 10] = \frac{1}{4} \cdot 10 + (1 - \frac{1}{4}) \cdot 0 = \frac{10}{4} + 0 = \frac{5}{2}$$



# Events as functions

## Proposition

Membership testing induces an isomorphism between events and Boolean propositions:

$$(\in) : \mathcal{B}_X \xrightarrow{\cong} \mathbb{B}^X$$

Its inverse sends each Boolean property to the set of outcomes satisfying it:

$$\frac{x : X \vdash M : \mathbb{B}}{\{x \in X \mid M\} : \mathcal{B}_X} \quad \{x \in X \mid \varphi x\} := \{x \in X \mid \varphi x = \mathbf{True}\}$$

## Characteristic function

represents an event as weight functions:  $E : \mathcal{B}_X \vdash [- \in E] : \mathbb{W}^X$

By the above proposition, every (internal)  $\{0, 1\}$ -valued weight function is the characteristic function of some event, namely, the inverse image of **1**.

# Measurement through integration

## Lemma

We can replace event measurement by integration of characteristic functions:

$$\mu : \mathbf{D}X, E : \mathcal{B}_X \vdash \mathbf{Ce}_{\mu}[E] = \int \mu(\mathrm{d}x) [x \in E]$$

We can deduce properties for  $\mathbf{Ce}[-]$  and  $\mathbf{Pr}[-]$  from those of the Lebesgue integral.

Notation:

$$\frac{\Gamma \vdash \mu : \mathbf{D}X \quad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \mathbf{Ce}_{x \sim \mu}[M] := \mathbf{Ce}_{\mu}[\{x \in X \mid M\}] : \mathbb{W}}$$

and similarly for  $\mathbf{Pr}_{x \sim \mu}[M]$ .

# Language of **probability** & **distribution** (recap)

$X$  type of **values/outcomes**

$\mathcal{D}X$  type of **distributions/measures** over  $X$

$\mathcal{P}X \subseteq \mathcal{D}X$  sub-type of **probability distributions** over  $X$

$\mathcal{B}_X \subseteq \mathcal{P}X$  type of **events**: subsets we wish to measure

$\mathbb{W}$  type of **weights**: values in  $[0, \infty]$

$\int, \mathbb{E}$  Lebesgue integration and the expectation operation

Type judgements describe well-formed values/outcomes of a given type, e.g.:

$$\mu : \mathcal{D}X, E : \mathcal{B}_X \vdash \text{Ce}_{\mu}[E] : \mathbb{W}$$

(measures weight  $\text{Ce}_{\mu}[E]$  of event  $E$  according to distribution  $\mu$ )

Propositions describe properties of well-formed values/outcomes of a given type, e.g.:

$$y_1, y_2 : Y \vdash y_1 \stackrel{Y}{=} y_2 : \text{Prop} \quad \mu : \mathcal{P}X, E : \mathcal{B}_X \vdash \text{cast}_{\mu} \text{Pr}[E] = \text{Ce}_{\mu}[E]$$

(probability of event according to probability distribution is its measure)

# Lecture plan

## Lecture 1: discrete model (today)

- ▶ Motivation
- ▶ Language of probability and distribution
- ▶ Discrete model
- ▶ **Simply-typed probability**
- ▶ Dependently-typed probability



course page

## Lecture 2: the full model

- ▶ Borel sets and measurable spaces
- ▶ Quasi-Borel spaces
- ▶ Type structure & standard Borel spaces
- ▶ Integration & random variables



ask questions on the  
Scottish PL Institute  
Zulip stream #qbs

# Simply-typed foundations for probabilistic modelling

## Compositional building blocks for modelling

- ▶ Affine combinations of distributions
- ▶ Product measures  $(\otimes) : \mathbf{D}X \times \mathbf{D}Y \rightarrow \mathbf{D}(X \times Y)$
- ▶ Random elements and their laws (push-forward measure):  
 $(\lambda(\mu, \alpha) \cdot \mu_\alpha) : \mathbf{D}\Omega \times X^\Omega \rightarrow \mathbf{D}X$

NB:

## Standard vocabulary

- ▶ Joint and marginal distributions
- ▶ Independence
- ▶ Distribution/probability preservation and invariance
- ▶ Density and absolute continuity
- ▶ Almost certain/sure properties

- ▶ Dirac kernel  $\delta_- : X \rightarrow \mathbf{D}X$

- ▶ Kock integration  
 $\oint : \mathbf{D}X \times (\mathbf{D}Y)^{\mathbf{D}X} \rightarrow \mathbf{D}Y$

# Simply-typed foundations for probabilistic modelling

## Compositional building blocks for modelling

- ▶ Affine combinations of distributions
- ▶ Product measures  $(\otimes) : \mathsf{D}X \times \mathsf{D}Y \rightarrow \mathsf{D}(X \times Y)$
- ▶ Random elements and their laws (push-forward measure):  
 $(\lambda(\mu, \alpha) . \mu_\alpha) : \mathsf{D}\Omega \times X^\Omega \rightarrow \mathsf{D}X$

NB:

## Standard vocabulary

- ▶ Joint and marginal distributions
- ▶ Independence
- ▶ Distribution/probability preservation and invariance
- ▶ Density and absolute continuity
- ▶ Almost certain/sure properties

- ▶ Dirac kernel  $\delta_- : X \rightarrow \mathsf{D}X$

- ▶ Kock integration  
 $\oint : \mathsf{D}X \times (\mathsf{D}Y)^{\mathsf{D}X} \rightarrow \mathsf{D}Y$

# Affine combinations of distributions: scaling

## Scaling distributions

$$w : \mathbb{W}, \mu : \mathbf{D}X \vdash w \cdot \mu : \mathbf{D}X$$

In the discrete model:

$$w \cdot \mu := \lambda x. w \cdot \mu x \quad \text{supp}(w \cdot \mu) \subseteq \text{supp } \mu$$

The function  $(\cdot) : \mathbb{W} \times \mathbf{D}X \rightarrow \mathbf{D}X$  is a **monoid action** for the monoid  $(\mathbb{W}, (\cdot), \mathbf{1})$ :

$$\mu : \mathbf{D}X \vdash \mathbf{1} \cdot \mu = \mu \quad w, v : \mathbb{W}, \mu : \mathbf{D}X \vdash w \cdot (v \cdot \mu) = (w \cdot v) \cdot \mu$$

Integration and measurement are homogeneous w.r.t. scaling:

$$w : \mathbb{W}, \mu : \mathbf{D}X, k : (\mathbf{D}Y)^X \vdash \oint d(w \cdot \mu)k = w \cdot \oint d\mu k$$

$$w : \mathbb{W}, \mu : \mathbf{D}X, f : \mathbb{W}^X \vdash \int d(w \cdot \mu)f = w \cdot \int d\mu f$$

$$w : \mathbb{W}, \mu : \mathbf{D}X, E : \mathcal{B}_X \vdash \text{Ce}_{w \cdot \mu}[f] = w \cdot \text{Ce}_{\mu}[f]$$

# Affine combinations of distributions: scaling

## Normalisation

$$\mu : \mathbf{D}X, \|\mu\| \neq 0, \infty \vdash \frac{\mu}{\|\mu\|} := \text{lift} \left( \frac{1}{\|\mu\|} \cdot \mu \right) : \mathbf{P}X$$

measurement is homogeneous

↓

$$\text{Indeed: } \left\| \frac{\mu}{\|\mu\|} \right\| = \left\| \frac{1}{\|\mu\|} \cdot \mu \right\| = \frac{1}{\|\mu\|} \cdot \|\mu\| = 1$$

## Discrete uniform / categorical distribution

Random unbiased choice between finitely many options/categories:

$$S : \mathcal{P}_{\text{fin}}(X), S \neq \emptyset \vdash \mathbf{U}_S := \frac{\text{lift} \#_S}{\|\text{lift} \#_S\|} : \mathbf{P}X$$

In the discrete model:

$$\mathbf{U}_S = \lambda x. \begin{cases} x \in S : \frac{1}{|S|} \\ x \notin S : 0 \end{cases}$$

so:  $x : X \vdash \mathbf{U}_{\{x\}} = \delta_x$ .



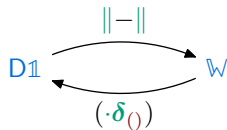
# Weights as distributions

## Unit type

$$\mathbb{1} := \{()\}$$

## Proposition

The following two functions are mutually inverse:



## Proof

Calculate:  $\mu : D\mathbb{1} \vdash \mu \mapsto \mu () \mapsto \lambda().\mu () = \mu$  and  $w : \mathbb{W} \vdash w \mapsto \lambda().w \mapsto w$ . ■

# Internalising Lebesgue integration

## Proposition

We can recover Lebesgue integration from Kock integration:

$$\begin{array}{ccc} DX \times W^X & \xrightarrow{\text{id} \times (\cong \circ)} & DX \times (D1)^X \\ \int \downarrow & = & \downarrow \oint \\ W & \xleftarrow{\cong} & D1 \end{array}$$

Since measurement also reduced to Lebesgue integration, it usually suffices to prove properties of Kock integration and derive them for Lebesgue integration and for measurement.

# Affine combinations of distributions: addition

## Summation

$$\mu_- : (\mathbf{D}X)^I, I \text{ countable} \vdash \sum_{i \in I} \mu_i : \mathbf{D}X$$

In the discrete model:

$$\sum_{i \in I} \mu_i := \lambda x. \sum_{i \in I} \mu_i x \quad \text{supp } \sum_{i \in I} \mu_i = \bigcup_{i \in I} \text{supp } \mu_i$$

## Affine and convex combinations

An **affine** combination is a countable sequence of weights  $w_- : \mathbb{W}^I$ .

It is **convex** when  $\sum_{i \in I} w_i = 1$ .

## Bernoulli revisited

We can express the Bernoulli distribution as follows:

$$p : \mathbb{I} \vdash \mathbf{B}p = \text{lift } (p \cdot \delta_{\mathbf{True}} + (1 - p) \cdot \delta_{\mathbf{False}}) : \mathbf{PB}$$

# Affinity of integration and convexity of expectation

## Theorem (Multi-linearity)

The Kock and Lebesgue integrals and measurement are affine in each argument:

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, k : X \rightsquigarrow Y \vdash \oint d(\sum_{i \in I} w_i \cdot \mu_i) k = \sum_{i \in I} w_i \cdot \oint d\mu_i k$$

$$\mu : \mathbf{D}X, w_- : \mathbb{W}^I, k_- : (X \rightsquigarrow B)^I \vdash \oint d\mu(\sum_{i \in I} w_i \cdot k_i) = \sum_{i \in I} w_i \cdot \oint d\mu k_i$$

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, \varphi : \mathbb{W}^X \vdash \int d(\sum_{i \in I} w_i \cdot \mu_i) \varphi = \sum_{i \in I} w_i \cdot \int d\mu_i \varphi$$

$$\mu : \mathbf{D}X, w_- : \mathbb{W}^I, \varphi_- : (\mathbb{W}^X)^I \vdash \int d\mu(\sum_{i \in I} w_i \cdot \varphi_i) = \sum_{i \in I} w_i \cdot \int d\mu \varphi_i$$

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, E : \mathcal{B}_X \vdash \sum_{i \in I} \text{Ce}_{w_i \cdot \mu_i} [E] = \sum_{i \in I} w_i \cdot \text{Ce}_{\mu_i} [E]$$

This theorem, a working horse in probability, has several important consequences:

## Proposition

The isomorphism  $\mathbb{D}\mathbb{1} \cong \mathbb{W}$  is a  $\sigma$ -semiring isomorphism:

$$(\mathbb{D}\mathbb{1}, \sum, (\cdot)) \cong (\mathbb{W}, \sum, (\cdot))$$

and  $(\cdot) : \mathbb{W} \times \mathbb{D}X \rightarrow \mathbb{D}X$  makes each  $\mathbb{D}X$  into a  $\mathbb{W}$ -module:

$$\left( \sum_{i \in I} w_i \right) \cdot \mu = \sum_{i \in I} (w_i \cdot \mu) \qquad w \cdot \sum_{i \in I} \mu_i = \sum_{i \in I} w \cdot \mu_i$$

# Convex combinations of probability distributions

## Lemma

**Convex** combination lifts to probability distributions:

$$w_- : \mathbb{W}^I, \mu_- : (\mathsf{PX})^I, I \text{ countable}, \sum_{i \in I} w_i = 1 \vdash$$

$$\sum_{i \in I} w_i \cdot \mu_i := \text{lift} \sum_{i \in I} w_i \cdot (\text{cast } \mu_i) : \mathsf{PX}$$

## Proof

Calculate:  $\left\| \sum_{i \in I} w_i \cdot (\text{cast } \mu_i) \right\| = \sum_{i \in I} w_i \cdot \|\text{cast } \mu_i\| = \sum_{i \in I} w_i \cdot 1 = 1$  ■

# Convex combinations of probability distributions

## Corollary (Multi-convexity)

Stochastic Kock integration, expectation and measurement are convex:

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, k : X \rightsquigarrow Y, \sum_{i \in I} w_i = 1 \vdash \oint d(\sum_{i \in I} w_i \cdot \mu_i) k = \sum_{i \in I} w_i \cdot \oint d\mu_i k$$

$$\mu : \mathbf{D}X, w_- : \mathbb{W}^I, k_- : (X \rightsquigarrow B)^I, \sum_{i \in I} w_i = 1 \vdash \oint d\mu(\sum_{i \in I} w_i \cdot k_i) = \sum_{i \in I} w_i \cdot \oint d\mu k_i$$

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, \varphi : \mathbb{W}^X, \sum_{i \in I} w_i = 1 \vdash \mathbb{E}_{\sum_{i \in I} w_i \cdot \mu_i} [\varphi] = \sum_{i \in I} w_i \cdot \mathbb{E}_{\mu_i} [\varphi]$$

$$\mu : \mathbf{D}X, w_- : \mathbb{W}^I, \varphi_- : (\mathbb{W}^X)^I, \sum_{i \in I} w_i = 1 \vdash \mathbb{E}_\mu \left[ \sum_{i \in I} w_i \cdot \varphi_i \right] = \sum_{i \in I} w_i \cdot \mathbb{E}_\mu [\varphi_i]$$

$$\mu_- : (\mathbf{D}X)^I, w_- : \mathbb{W}^I, E : \mathcal{B}_X, \sum_{i \in I} w_i = 1 \vdash \sum_{i \in I} \text{Pr}_{w_i \cdot \mu_i} [E] = \sum_{i \in I} w_i \cdot \text{Pr}_{\mu_i} [E]$$

# Products

## Product distribution

$$\mu : \mathsf{D}X, \nu : \mathsf{D}Y \vdash \mu \otimes \nu := \int \mu(\mathrm{d}x) \int \nu(\mathrm{d}y) \delta_{(x,y)} : \mathsf{D}(X \times Y)$$

In the discrete model:

$$\mu \otimes \nu = \lambda(x, y) . (\mu x) \cdot (\nu y) \quad \text{supp}(\mu \otimes \nu) = (\text{supp } \mu) \times (\text{supp } \nu)$$

Example: counting distribution on product space

$$S : \mathcal{P}_{\text{fin}}(X), T : \mathcal{P}_{\text{fin}}(Y) \vdash \#_{S \times T} \stackrel{\mathsf{D}(X \times Y)}{=} \#_S \otimes \#_T$$

Indeed:  $\text{supp}(\#_S \otimes \#_T) = S \times T = \text{supp } \#_{S \times T}$  and for  $(x, y) \in S \times T$ :

$$(\#_S \otimes \#_T)(x, y) = 1 \cdot 1 = 1 = \#_{S \times T}(x, y)$$



# Products

Notation: 
$$\frac{\Gamma \vdash M : \mathbf{D}(X \times Y) \quad \Gamma, x : X, y : Y \vdash K : \mathbf{D}Z}{\Gamma \vdash \oint M(\mathrm{d}x, \mathrm{d}y) K := \oint \mathrm{d}M(\lambda(x, y).K) : \mathbf{D}Z}$$

## Theorem (Fubini-Tonelli)

We can integrate products in any order:

$$\mu : \mathbf{D}X, \nu : \mathbf{D}Y, k : (\mathbf{D}Z)^{X \times Y} \vdash$$
$$\oint \mu(\mathrm{d}x) \oint \nu(\mathrm{d}y) k(x, y) = \oint (\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y) k(x, y) = \oint \nu(\mathrm{d}y) \oint \mu(\mathrm{d}x) k(x, y)$$

$$\mu : \mathbf{D}X, \nu : \mathbf{D}Y, \varphi : \mathbf{W}^{X \times Y} \vdash$$
$$\int \mu(\mathrm{d}x) \int \nu(\mathrm{d}y) \varphi(x, y) = \iint (\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y) \varphi(x, y) = \int \nu(\mathrm{d}y) \int \mu(\mathrm{d}x) \varphi(x, y)$$

# Applying Fubini-Tonelli

## Theorem (Rule of Product)

We can factor out products:

$$\begin{aligned} \mu : \mathbf{D}X, f : \mathbb{W}^X, \nu : \mathbf{D}Y, g : \mathbb{W}^Y &\vdash \iint (\mu \otimes \nu)(dx, dy) f x \cdot g y = \left( \int d\mu f \right) \cdot \left( \int d\nu g \right) \\ \mu : \mathbf{D}X, E : \mathcal{B}_X, \nu : \mathbf{D}Y, F : \mathcal{B}_Y &\vdash \text{Ce}_{\mu \otimes \nu} [E \times F] = \text{Ce}_{\mu} [E] \cdot \text{Ce}_{\nu} [F] \end{aligned}$$

## Theorem

The product lifts to probability distributions:

$$\mu : \mathbf{P}X, \nu : \mathbf{P}Y \vdash (\mu \otimes \nu) := \text{lift}(\text{cast } \mu \otimes \text{cast } \nu) : \mathbf{P}(X \times Y)$$

## Binomial distribution

the number of successful outcomes of  $n$  independent Bernoulli trials:

$$\begin{aligned} \mathbf{B}_n : \mathbb{I} &\rightsquigarrow \mathbf{P}(\mathbf{Fin}(1+n)) & \mathbf{B}_0 p &:= \delta_0 : \mathbf{P}(\mathbf{Fin} 1) \\ \mathbf{B}_{1+n} p &:= \iint (\mathbf{B}_n p \otimes \mathbf{B} p)(d\mathbf{c}, d\mathbf{b}) \text{ (if } \mathbf{b} \text{ then } \delta_{1+\mathbf{c}} \text{ else } \delta_{\mathbf{c}}) : \mathbf{P}(\mathbf{Fin}(2+n)) \end{aligned}$$

We can prove by induction on  $n$ , using Fubini-Tonelli and the Iverson bracket that:

$$p : \mathbb{I}, k : \mathbf{Fin}(1+n) \vdash \Pr_{\mathbf{c} \sim \mathbf{B}_n p} [\mathbf{c} = k] = \binom{n}{k}$$

# Push-forward distributions

## Random element

in  $X$  any (internal) function:

$$\mu : \mathsf{D}\Omega \vdash \alpha : \Omega \rightarrow X$$

## Law

of a random element is the distribution:

$$\mu : \mathsf{D}\Omega, \alpha : X^\Omega \vdash \mu_\alpha := \int \mu(d\omega) \delta_{\alpha\omega} : \mathsf{D}X$$

## Example

Represent outcomes of die roll by  $\mathsf{D6} := \{1, 2, \dots, 6\}$ , and two rolls by  $\mathsf{D6} \times \mathsf{D6}$ .

The sum of the rolls is a random element:

$$(+): \mathsf{D6} \times \mathsf{D6} \rightarrow \mathbb{N}$$

The law of the distribution  $\# \otimes \#$  counts the number of configurations in which the two rolls sum to a given number, e.g.:  $(\# \otimes \#)_{(+)} : 1 \mapsto 0, 2 \mapsto 1$ .

## Theorem (Law of the Unconscious Statistician)

Formulae for reparameterising integration and measurement:

$$\mu : \Omega, \alpha : X^\Omega, k : X \rightsquigarrow Y \vdash \oint d\mu_\alpha k = \oint d\mu (k \circ \alpha)$$

$$\mu : \Omega, \alpha : X^\Omega, f : \mathbb{W}^X \vdash \int d\mu_\alpha f = \int d\mu (f \circ \alpha)$$

$$\mu : \Omega, \alpha : X^\Omega, E : \mathcal{B}_X \vdash \text{Ce}_{\mu_\alpha}[E] = \text{Ce}_\mu[\alpha^{-1}[E]] = \text{Ce}_{\omega \sim \mu}[\alpha \omega \in E]$$

# Simply-typed foundations for probabilistic modelling

## Compositional building blocks for modelling

- ▶ Affine combinations of distributions
- ▶ Product measures  $(\otimes) : \mathsf{D}X \times \mathsf{D}Y \rightarrow \mathsf{D}(X \times Y)$
- ▶ Random elements and their laws (push-forward measure):  
 $(\lambda(\mu, \alpha) . \mu_\alpha) : \mathsf{D}\Omega \times X^\Omega \rightarrow \mathsf{D}X$

NB:

## Standard vocabulary

- ▶ Joint and marginal distributions
- ▶ Independence
- ▶ Distribution/probability preservation and invariance
- ▶ Density and absolute continuity
- ▶ Almost certain/sure properties

▶ Dirac kernel  $\delta_- : X \rightarrow \mathsf{D}X$

▶ Kock integration  
 $\oint : \mathsf{D}X \times (\mathsf{D}Y)^{\mathsf{D}X} \rightarrow \mathsf{D}Y$

# Standard vocabulary: concepts concerning products

Let  $\pi_i : \prod_{i \in I} X_i \rightarrow X_i$  be the  $i$ -th projection.

**Joint distribution:**  $\mu : \mathsf{D}(X \times Y)$ ,  $\mu : \mathsf{D}(\prod_{i \in I} X_i)$

**Marginal distribution:** the law of a projection:

$$\mu : \mathsf{D}\left(\prod_{i \in I} X_i\right) \vdash \mu_{\pi_i} : \mathsf{D}X_i$$

Sometimes refers to any law of a r.e..

**Marginalisation:** the action of calculating a marginal distribution by integrating all other components.

Exercise

$$\mu : \mathsf{P}X, \nu : \mathsf{D}X \vdash (\mu \otimes \nu)_{\pi_2} = \nu$$

# Independence

## Pairing random elements

$$\alpha : X^\Omega, \beta : Y^\Omega \vdash \lambda \omega. (\alpha \omega, \beta \omega) : (X \times Y)^\Omega$$

## Independent random elements

The joint law is the product of the marginals:

$$\mu : \mathbb{D}\Omega, \alpha : X^\Omega, \beta : Y^\Omega \vdash \alpha \underset{\mu}{\perp} \beta := \left( \mu_{(\alpha, \beta)} \stackrel{\mathbb{D}(X \times Y)}{=} \mu_\alpha \otimes \mu_\beta \right)$$

More generally, for finite  $I$ :

$$\mu : \mathbb{D}\Omega, \alpha_- : (X^\Omega)^I \vdash \underset{\mu}{\perp}_i \alpha_i := \left( \mu_{(\alpha_i)_i} \stackrel{\mathbb{D}(\prod_i X_i)}{=} \bigotimes_{i \in I} \mu_{\alpha_i} \right)$$



# Independence

## Example [Durett]

Model 3 independent coin tosses:

$$\text{Toss} := \{\text{Head}, \text{Tail}\} \quad \Omega := \text{Toss}^3 \quad \mu := \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} : \mathbf{P}\Omega$$

The outcome of the  $i^{\text{th}}$  coin toss is the random element  $\pi_i : \Omega \rightarrow \text{Toss}$ .

Consider the Boolean proposition in which the  $i^{\text{th}}$  and  $j^{\text{th}}$  tosses ( $i \neq j$ ) agree:

$$\text{Same}_{ij} := \lambda\omega. \pi_i\omega = \pi_j\omega : \Omega \rightarrow \mathbb{B}$$

$$\begin{array}{ccccc} \text{Calculate:} & \text{LOTUS} & & \text{marginalisation} & & \text{Fubini} \\ & \downarrow & & \downarrow & & \downarrow \\ \Pr_{\mu} [\text{Same}_{12}] & = & \Pr_{(x,y) \sim \mu(\pi_1, \pi_2)} [x = y] & = & \Pr_{(x,y) \sim \mathbf{U} \otimes \mathbf{U}} [x = y] & = & \int \mathbf{U}(dx) \Pr_{y \sim \mathbf{U}} [x = y] \\ & & & & & & \\ & = & \frac{1}{2} \cdot \Pr_{y \sim \mathbf{U}} [\text{Head} = y] & + & \frac{1}{2} \cdot \Pr_{y \sim \mathbf{U}} [\text{Tail} = y] & = & \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{array}$$

# Independence

## Example [Durett]

Model 3 independent coin tosses:

$$\text{Toss} := \{\text{Head}, \text{Tail}\} \quad \Omega := \text{Toss}^3 \quad \mu := \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} : \mathbf{P}\Omega$$

The outcome of the  $i^{\text{th}}$  coin toss is the random element  $\pi_i : \Omega \rightarrow \text{Toss}$ .

Consider the Boolean proposition in which the  $i^{\text{th}}$  and  $j^{\text{th}}$  tosses ( $i \neq j$ ) agree:

$$\text{Same}_{ij} := \lambda \omega. \pi_i \omega = \pi_j \omega : \Omega \rightarrow \mathbb{B}$$

Therefore  $\mu_{\text{Same}_{12}} = \mathbf{U}_{\mathbb{B}}$  and similarly  $\mu_{\text{Same}_{ij}} = \mathbf{U}_{\mathbb{B}}$  for  $i \neq j$ .

# Independence

$\pi_1$ ,  $\text{Same}_{12}$ , and  $\text{Same}_{13}$  determine  $\pi_2, \pi_3$ , so:

$$\Pr_{\omega \sim \mu} [\text{Same}_{12}\omega = \text{True}, \text{Same}_{13}\omega = \text{True}]$$

Fubini-Tonelli

$$\begin{aligned} & \downarrow \\ &= \int \mathbf{U}_{\text{Toss}}(d\mathbf{b}_1) \Pr_{(b_2, b_3) \sim (\mathbf{U} \otimes \mathbf{U})} [\text{Same}_{12}(b_1, b_2, b_3) = \text{True}, \text{Same}_{13}(b_1, b_2, b_3) = \text{True}] \\ &= \frac{1}{2} \Pr_{(b_2, b_3) \sim (\mathbf{U} \otimes \mathbf{U})} [\text{Same}_{12}(\text{Head}, b_2, b_3) = \text{True}, \text{Same}_{13}(\text{Head}, b_2, b_3) = \text{True}] \\ &+ \frac{1}{2} \Pr_{(b_2, b_3) \sim (\mathbf{U} \otimes \mathbf{U})} [\text{Same}_{12}(\text{Tail}, b_2, b_3) = \text{True}, \text{Same}_{13}(\text{Tail}, b_2, b_3) = \text{True}] \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \end{aligned}$$

and similarly we get  $\frac{1}{4}$  in all other cases.

# Independence

## Example [Durett]

Model 3 independent coin tosses:

$$\text{Toss} := \{\text{Head}, \text{Tail}\} \quad \Omega := \text{Toss}^3 \quad \mu := \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} \otimes \mathbf{U}_{\text{Toss}} : \mathbf{P}\Omega$$

The outcome of the  $i^{\text{th}}$  coin toss is the random element  $\pi_i : \Omega \rightarrow \text{Toss}$ .

Consider the Boolean proposition in which the  $i^{\text{th}}$  and  $j^{\text{th}}$  tosses ( $i \neq j$ ) agree:

$$\text{Same}_{ij} := \lambda \omega. \pi_i \omega = \pi_j \omega : \Omega \rightarrow \mathbb{B}$$

Therefore  $\mu_{\text{Same}_{12}} = \mathbf{U}_{\mathbb{B}}$  and similarly  $\mu_{\text{Same}_{ij}} = \mathbf{U}_{\mathbb{B}}$  for  $i \neq j$ . So:

$$\mu_{(\text{Same}_{12}, \text{Same}_{13})} = \mathbf{U}_{\mathbb{B} \times \mathbb{B}} = \mathbf{U}_{\mathbb{B}} \otimes \mathbf{U}_{\mathbb{B}} = \mu_{\text{Same}_{12}} \otimes \mu_{\text{Same}_{13}}$$

So  $\text{Same}_{12} \perp_{\mu} \text{Same}_{13}$  even though their values depend on the outcome of the first toss.

# Distribution preservation

## Distribution space $(\Omega, \mu)$

A type  $\Omega$  equipped with a distribution  $\mu : \mathbf{D}\Omega$ . Define **probability space** analogously.

## Distribution preserving function

$f : (\Omega_1, \mu_1) \rightarrow (\Omega_2, \mu_2)$  is a function whose is the co domain distribution:

$$f : \Omega_1 \rightarrow \Omega_2 \quad (\mu_1)_f = \mu_2$$

$\mu : \mathbf{D}X$  is **invariant** under  $f : X \rightarrow X$  when  $f : (X, \mu) \rightarrow (X, \mu)$  is dist. preserving.

## Example

Consider the swapping function:  $\text{swap} := (\lambda (x, y). (y, x)) : X \times Y \rightarrow Y \times X$ . Then, for each  $\mu : \mathbf{D}X$ ,  $\nu : \mathbf{D}Y$ , swapping is distribution preserving function:

$$\text{swap} : (X \times Y, \mu \otimes \nu) \rightarrow (Y \times X, \nu \otimes \mu)$$

$\text{swap}$  is invariant in the case  $X = Y$  and  $\mu = \nu$ .

# Density and scaling

## Density

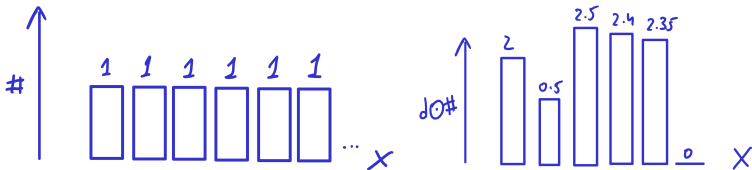
over  $X$  is any weight function  $f : X \rightarrow \mathbb{W}$ .

## Density scaling

We can scale a distribution by a density:

$$f : \mathbb{W}^X, \mu : \mathbb{D}X \vdash f \odot \mu := \int \mu(dx)(f, x) \cdot \delta_x : \mathbb{D}X$$

Scaling does not lift to probability distributions:  $\|f \odot \mu\| \neq 1$  even if  $\|\mu\| = 1$ .



# Density and scaling

## Density

over  $X$  is any weight function  $f : X \rightarrow \mathbb{W}$ .

## Density scaling

We can scale a distribution by a density:

$$f : \mathbb{W}^X, \mu : \mathsf{DX} \vdash f \odot \mu := \int \mu(\mathrm{d}x)(f, x) \cdot \delta_x : \mathsf{DX}$$

Scaling does not lift to probability distributions:  $\|f \odot \mu\| \neq 1$  even if  $\|\mu\| = 1$ .

## Warning!

The types of distributions and densities over  $X$  in the **discrete** model are close, but still **different**. They coincide on **countable** types, so people often confused them.

Types help us keep them separate.

# Density and absolute continuity

## Having density

This concept has several names in the literature:

$$\mu, \nu : \mathbf{DX}, f : \mathbb{W}^X \vdash \left( f = \frac{d\mu}{d\nu} \right) := (\mu = f \odot \nu) : \mathbf{Prop}$$

- ▶  $f$  is the **density** of  $\mu$  w.r.t.  $\nu$
- ▶  $f$  is a **Radon-Nikodym derivative** of  $\mu$  w.r.t.  $\nu$ .

## Absolute continuity

$\mu$  is **absolutely continuous** w.r.t.  $\nu$  when  $\mu$  has a density w.r.t.  $\nu$ :

$$\mu, \nu : \mathbf{DX} \vdash (\mu \ll \nu) := \exists f : \mathbb{W}^X. f = \frac{d\mu}{d\nu} : \mathbf{Prop}$$



# Density and absolute continuity

## Example

The **uniform distribution** is absolutely continuous w.r.t. the **counting measure** over the same support. Indeed, it has these two densities:

$$S : \mathcal{P}_{\text{fin}}(X) \vdash \left( \lambda x. \frac{1}{|S|} \right), \left( \lambda x. \begin{cases} x \in S : \frac{1}{|S|} \\ x \notin S : 0 \end{cases} \right) = \frac{d\mathbf{U}_S}{d\#_S}$$

These two densities are different, but they agree on the support, motivating the following concept.

# Almost certain/sure properties

## Almost certain event

is one we can assert without changing the distribution:

$$\frac{\Gamma \vdash \mu : \mathbf{D}X \quad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \mu(dx) \text{ almost certainly } M := [M] \odot \mu = \mu : \mathbf{Prop}}$$

For probabilities we define:

$$\frac{\Gamma \vdash \mu : \mathbf{P}X \quad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \mu(dx) \text{ almost surely } M := (\text{cast } \mu)(dx) \text{ almost certainly } M : \mathbf{Prop}}$$

# Existence and almost-sure uniqueness of densities

## Theorem (Radon-Nikodym)

For **probability** distributions, we characterise absolute continuity as follows:

$$\mu, \nu : \mathbf{P}X \vdash (\mu \ll \nu) \iff \forall E : \mathcal{B}_X. \Pr_{\nu}[E] = 0 \implies \Pr_{\mu}[E] = 0$$

In that case, if  $f, g = \frac{d\mu}{d\nu}$  then  $\nu(dx)$  **almost surely**  $f x = g x$ .

In the **discrete model**, this characterisation amounts to  $\text{supp } \mu \subseteq \text{supp } \nu$ .

## Example

For all countable  $X$ , we have:

$$\forall \mu : \mathbf{D}X. \mu \ll \#_X$$

Indeed, apply the Radon-Nikodym theorem, since  $\text{supp } \# = X$ .

Constructively, direct calculation shows:  $(\lambda x. \mu x) = \frac{d\mu}{d\#}$ .

# Simply-typed foundations for probabilistic modelling

## Compositional building blocks for modelling

- ▶ Affine combinations of distributions
- ▶ Product measures  $(\otimes) : \mathbf{D}X \times \mathbf{D}Y \rightarrow \mathbf{D}(X \times Y)$
- ▶ Random elements and their laws (push-forward measure):  
 $(\lambda(\mu, \alpha) \cdot \mu_\alpha) : \mathbf{D}\Omega \times X^\Omega \rightarrow \mathbf{D}X$

NB:

## Standard vocabulary

- ▶ Joint and marginal distributions
- ▶ Independence
- ▶ Distribution/probability preservation and invariance
- ▶ Density and absolute continuity
- ▶ Almost certain/sure properties

- ▶ Dirac kernel  $\delta_- : X \rightarrow \mathbf{D}X$

- ▶ Kock integration  
 $\oint : \mathbf{D}X \times (\mathbf{D}Y)^{\mathbf{D}X} \rightarrow \mathbf{D}Y$

# Lecture plan

## Lecture 1: discrete model (today)

- ▶ Motivation
- ▶ Language of probability and distribution
- ▶ Discrete model
- ▶ Simply-typed probability
- ▶ Dependently-typed probability



course page

## Lecture 2: the full model

- ▶ Borel sets and measurable spaces
- ▶ Quasi-Borel spaces
- ▶ Type structure & standard Borel spaces
- ▶ Integration & random variables



ask questions on the  
Scottish PL Institute  
Zulip stream #qbs

## Example: Binomial kernels

We've defined, for every  $n \in \mathbb{N}$ , the binomial kernel:

$$\vdash \mathbf{B}_n : \mathbb{I} \rightsquigarrow \mathbf{Fin}(1 + n)$$

We will now look at **dependent-type** structure which allows us to view these as one kernel internally:

$$n : \mathbb{N} \vdash \mathbf{B}_n : \mathbb{I} \rightsquigarrow \mathbf{Fin}(1 + n)$$

# Family model

## Family over an indexing set $I$

consists of a sequence  $X_- = (X_i)_{i \in I}$  of sets.

We call each set  $X_i$  the **fibre over  $i$** .

## Family $F$

a pair  $F = (I, X_-)$  consisting of (indexing) set  $I$  and a family  $X_-$  over it.

Notation:  $F = I \vdash X_-$

$$= i : I \vdash X_i.$$

## Example

The family  $n : \mathbb{N} \vdash \mathbf{Fin} \, n$  has  $\mathbb{N}$  as the indexing set. The fibre over  $n \in \mathbb{N}$  is:

$$\mathbf{Fin} \, n := \{0, 1, \dots, n - 1\}$$

# Family model

## Family over an indexing set $I$

consists of a sequence  $X_- = (X_i)_{i \in I}$  of sets.

We call each set  $X_i$  the **fibre over  $i$** .

## Family $F$

a pair  $F = (I, X_-)$  consisting of (indexing) set  $I$  and a family  $X_-$  over it.

Notation:  $F = I \vdash X_-$

$$= i : I \vdash X_i.$$

## Family map

$(\theta, f_-) : (I \vdash X_-) \rightarrow (J \vdash Y_-)$  is a pair of a function between the indexing sets and a sequence of functions between the corresponding fibres:

$$\theta : I \rightarrow J \quad (f_i : X_i \rightarrow Y_{\theta i})_{i \in I}$$

Notation:  $\theta \vdash f_-$ . We won't use these maps explicitly, but they are the foundation.



# Terms in context

Dependent elements  $i : I \vdash M : X_i$

in family  $i : I \vdash X_i$  are  $I$ -indexed sequences of elements from the corresponding fibres:

$$(M \in X_i)_{i \in I}$$

## Example

We have the elements:

$$n : \mathbb{N} \vdash 0, \dots, n - 1 : \mathbf{Fin} \, n$$

## Subsumption

Every simple type becomes a family by ignoring the dependency through the constant family, e.g.,  $i : I \vdash \mathbb{N}$  and  $i : I \vdash 42 : \mathbb{N}$ .

# Simple functions

## Fibred exponential

of two families over the same indexing set  $i : I \vdash X_i, Y_i$  is the family:

## Family of distributions

$$i : I \vdash X_i \rightarrow Y_i$$

over a family  $i : I \vdash X_i$  is the family:

$$i : I \vdash \mathbf{D}X_i$$

Its sub-family of fibred **probability** distributions:

$$i : I \vdash \mathbf{P}X_i$$

Both have a **Dirac** distribution:

$$i : I \vdash \delta_- : X_i \rightarrow \mathbf{D}X_i \quad i : I \vdash \delta_- : X_i \rightarrow \mathbf{P}X_i$$

# Extension and dependent pairs

## Extension

of indexing set  $I$  by a **variable** of the family  $i : I \vdash X_i$  is the (indexing) set:

$$\coprod_{i \in I} X_i := \bigcup_{i \in I} \{i\} \times X_i = \left\{ (i, x) \in I \times \bigcup_{i \in I} X_i \mid x \in X_i \right\}$$

Notation:  $(i : I, x : X_i) := \coprod_{i \in I} X_i$  and we'll often write  $i, x$  instead of  $(i, x)$ .

## Dependent pairs

$$\frac{i : I \vdash X_i \quad i : I, x : X_i \vdash Y_{i,x}}{i : I \vdash (x : X_i) \times (Y_{i,x}) := \coprod_{x \in X_i} Y_{i,x}}$$

# Functions and kernels

## Dependent functions

we identify a function  $f$  with a tuple  $(f\ x)_x$  as usual:

$$\frac{i : I \vdash X_i \quad i : I, x : X_i \vdash Y_{i,x}}{i : I \vdash ((x : X) \rightarrow Y_{i,x}) := \prod_{x \in X} Y_{i,x}}$$

Dependent kernels  $i : I \vdash k : (x : X_i) \rightsquigarrow Y_{i,x}$

are dependent elements:

$$i : I \vdash k : (x : X_i) \rightarrow \mathbf{D}Y_{i,x}$$

Dependent **stochastic** kernels  $i : I \vdash k : (x : X_i) \rightsquigarrow Y_{i,x}$  are similarly:

$$i : I \vdash k : (x : X_i) \rightarrow \mathbf{P}Y_{i,x}$$

## Dependent Kock integral

$$i : I, \mu : \mathbf{D}X_i, k : (x : X_i) \rightsquigarrow Y_{i,x} \vdash \oint d\mu k : \mathbf{D}Y_{i,x}$$

and in the **discrete model** we define it for  $i, \mu, k$  as in the simply-typed case:

$$(\oint d\mu k)y := \sum_{x \in X_i} \mu x \cdot k(x; y) : \mathbb{W}$$

Through the identification  $\mathbb{W} \cong \mathbf{D}\mathbf{1}$  and characteristic functions, we reduce dependent Lebesgue integration and measurement to dependent Kock integration:

$$i : I, \mu : \mathbf{D}X_i, f : (x : X_i) \rightarrow \mathbb{W} \vdash \int d\mu f : \mathbb{W} \qquad i : I, \mu : \mathbf{D}X_i, E : \mathcal{B}_{X_i} \vdash \mathbf{Ce}_\mu[E] : \mathbb{W}$$

$$\int d\mu f = \sum_{x \in X} \mu x \cdot f x \qquad \mathbf{Ce}_\mu[E] = \sum_{x \in E} \mu x$$

# Random variables

Let  $\overline{\mathbb{R}} := [-\infty, \infty]$  be the extended real line.

## Signed and unsigned random variable

in a probability space  $(\Omega, \mu)$  are random elements  $\alpha : \Omega \rightarrow \overline{\mathbb{R}}$  and  $\alpha : \Omega \rightarrow \mathbb{W}$ .

The **positive** and **negative parts** are unsigned random variables  $\alpha^\pm : \Omega \rightarrow \mathbb{W}$ :

$$\alpha^+ := \lambda\omega. \max(\alpha\omega, 0) = [\alpha \geq 0] \cdot |\alpha| \quad \alpha^- := \lambda\omega. -\min(\alpha\omega, 0) = [\alpha \leq 0] \cdot |\alpha|$$

An unsigned r.v.  $\alpha$  is **Lebesgue integrable** when its Lebesgue integral is finite:

$$\int d\mu\alpha < \infty.$$

For a (signed) r.v.  $\alpha$ , when either  $\alpha^+$  or  $\alpha^-$  is Lebesgue integrable, we define:

$$\mu : \mathbb{D}X, \alpha : \overline{\mathbb{R}}^X, \int d\mu\alpha^+, \int d\mu\alpha^- < \infty \vdash \quad \int d\mu\alpha := \int d\mu\alpha^+ - \int d\mu\alpha^-$$

A signed variable is **Lebesgue integrable** when both its parts are Lebesgue integrable.

# Random variable spaces

Lebesgue integrability is a Boolean property:

$$\mu : \mathbf{D}X, \alpha : X \rightarrow \overline{\mathbb{R}} \vdash \alpha \text{ integrable} := \int d\mu \alpha^+ < \infty \wedge \int d\mu \alpha^- < \infty : \mathbb{B}$$

Lebesgue spaces ensemble

is the family:

$$i : I, p : [1, \infty), \mu : \mathbf{P}X_i \vdash \mathcal{L}_p(X_i, \mu) := \{ \alpha : X_i \rightarrow \overline{\mathbb{R}} \mid \alpha^p \text{ integrable} \}$$

Every fibre has a vector space structure and a norm (almost a Banach space!):

$$i : I, p : [1, \infty), \mu : \mathbf{P}X_i, \alpha : \mathcal{L}_p(X_i, \mu) \vdash \|\alpha\|_p := \sqrt[p]{\mathbb{E}_\mu[|\alpha|^p]} : \mathbb{W}$$

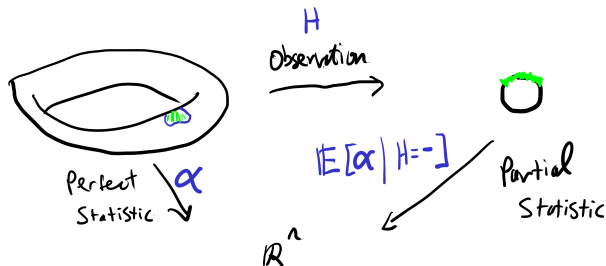
and the fibre **2** has an inner product (almost a Hilbert space!):

$$i : I, \mu : \mathbf{P}X_i, \alpha, \beta : \mathcal{L}_2(X_i, \mu) \vdash (\alpha, \beta) := \sqrt{\mathbb{E}_\mu[\alpha \cdot \beta]} : \mathbb{W}$$

# Conditioning

Situation:

- ▶ Statistical model  $\mu : \mathbf{D}\Omega$   
(voters in the next election)
- ▶ Perfect statistic  $\alpha : \Omega \rightarrow \mathbb{R}$   
(expected winning candidate)
- ▶ Observation  $H : \mathbf{D} \rightarrow X$   
(poll voting intention)



Conditional expectation of  $\alpha$  along  $H$  w.r.t  $\mu$

Statistic  $\beta : X \rightarrow \mathbb{R}$  that 'best' approximates  $H \circ \alpha$  statistically. Halmos and Doob's definition: any measurement we make of  $\beta$  agrees with measurement of  $\alpha$ :

$$\mu : \mathbf{D}\Omega, H : \Omega \rightarrow X, \alpha : \mathcal{L}_1(\Omega, \mu), \beta : \mathcal{L}_1(X, \mu_H) \vdash$$

$$\left( \beta = \mathbb{E}_{\mu} [\alpha | H = -] \right) := \left( \forall \varphi : \mathcal{L}_1 X, \mu_H. \int d\mu_H \beta \cdot \varphi = \int d\mu \alpha (\varphi \circ H) \right) : \text{Prop}$$



# Conditioning

## Theorem (Kolmogorov)

Every random variable has a conditional expectation:

$$\mu : \mathsf{D}\Omega, H : \Omega \rightarrow X, \alpha : \mathcal{L}_1(\Omega, \mu) \vdash \quad \exists \beta : \mathcal{L}_1(X, \mu_H). \beta = \mathbb{E}_{\mu}[\alpha | H = -]$$

Therefore:

## Corollary (Internal conditional expectation)

In the **discrete model** we have a dependent function:

$$\mathbb{E}_{-}[- | - = -] :$$

$$(\mu : \mathsf{D}\Omega) \rightarrow (H : \Omega \rightarrow X) \rightarrow (\alpha : \mathcal{L}_1(\Omega, \mu)) \rightarrow \left\{ \beta : \mathcal{L}_1(X, \mu_H) \left| \beta = \mathbb{E}_{\mu}[\alpha | H = -] \right. \right\}$$

# Conditioning

Disintegration of  $H : \Omega \rightarrow X$  w.r.t.  $\mu : \mathsf{D}\Omega$

is a kernel  $k : X \rightsquigarrow \Omega$  that weakly disintegrates the law of  $H$  and, moreover,  $\mathrm{d}x$ -almost surely each probability distribution  $k(x)$  is in the fibre:

$$\mu : \mathsf{D}\Omega, H : \Omega \rightarrow X, k : X \rightsquigarrow \Omega \vdash k \text{ disintegrates } H \text{ w.r.t. } \mu$$

$$:= \mu = \oint \mathrm{d}\mu_H k,$$

$$\mu_H(\mathrm{d}x) \text{ almost surely } k(x; \mathrm{d}\omega) \text{ almost surely } H\omega = x \quad : \mathsf{Prop}$$

In the **discrete model** we have an internal disintegration:

$$-\dagger- : (H : \Omega \rightarrow X) \rightarrow (\mu : \mathsf{P}\Omega) \rightarrow \{k : X \rightsquigarrow \Omega \mid k \text{ disintegrates } H \text{ w.r.t. } \mu\}$$

$$H^\dagger \mu := \lambda x. \begin{cases} \mu x > 0 : & \frac{[H = x]}{\mu_H x} \odot \mu \\ \mu x = 0 : & \mu \end{cases}$$

# Lecture plan

## Lecture 1: discrete model (today)

- ▶ Motivation
- ▶ Language of probability and distribution
- ▶ Discrete model
- ▶ Simply-typed probability
- ▶ Dependently-typed probability



course page

## Lecture 2: the full model

- ▶ Borel sets and measurable spaces
- ▶ Quasi-Borel spaces
- ▶ Type structure & standard Borel spaces
- ▶ Integration & random variables



ask questions on the  
Scottish PL Institute  
Zulip stream #qbs