Foundations for Type-Driven Probabilistic Modelling

Ohad Kammar
University of Edinburgh

Logic and Semantics Group
3 February, 2026
Department of Computer Science
University of Aarhus

THE UNIVERSITY *of* EDINBURGH informatics lfcs Laboratory for Foundations of Computer Science BayesCentre supported by: THE ROYAL SOCIETY The Alan Turing Institute Facebook Research NCSC

Computational golden era

logic-rich & type-rich computation

statistical computation

Computational golden era

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	$\hookrightarrow$	full model
familiar maths		supports discrete
introductory		and
		continuous distributions
		same language

Motivation: why foundations?

discrete probability
countably supported distributions
good type-structure
(this course)

measure theory
standard, established
poor type-structure

↘ **well-behaved probability** ↗
s-finite distributions
over standard Borel spaces

continuous probability ↗
Lebesgue measure over $\mathbb{R}^n$

↘ **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

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

- document **intent**:

  probability $(\text{Distribution}\,X)$ vs. density $(X \to [0,\infty])$ 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 (now)

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

Lecture 2: the full model

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

course page

ask questions on the Scottish PL Institute Zulip stream #qbs

Language of probability & distribution

X	type (=space) of **values**/**outcomes**
$\mathsf{D}X$	type of **distributions**/**measures** over X
$\mathsf{P}X \subseteq \mathsf{D}X$	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.:

$$\mu : \mathsf{D}X, E : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Ce}}[E] : \mathbb{W}$$

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

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

$$y_1, y_2 : Y \vdash y_1 \overset{Y}{=} y_2 : \mathsf{Prop} \qquad \mu : \mathsf{P}X, E : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Pr}}[E] = \underset{\mu}{\mathrm{Ce}}[E]$$

(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

$$\mu : \mathsf{D}X \vdash \underset{\mu}{\mathrm{Ce}}[\emptyset] = 0$$

Axioms for events and distributions

Boolean Sub-algebra of Events

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

Disjoint additivity

$$w, v : \mathbb{W} \vdash w + v : \mathbb{W} \qquad E, C : \mathcal{B}_X, \mu : \mathsf{D}X \vdash \underset{\mu}{\mathrm{Ce}}[E] = \underset{\mu}{\mathrm{Ce}}[E \cap C] + \underset{\mu}{\mathrm{Ce}}\left[E \cap C^{\complement}\right]$$

Axioms for events and distributions

Boolean Sub-algebra of Events

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

Disjoint additivity

$w, v : \mathbb{W} \vdash w + v : \mathbb{W} \qquad E, C : \mathcal{B}_X, \mu : \mathsf{D}X \vdash \underset{\mu}{\mathrm{Ce}}[E] = \underset{\mu}{\mathrm{Ce}}[E \cap C] + \underset{\mu}{\mathrm{Ce}}\left[E \cap C^{\complement}\right]$

Exercise

Derive 'axiomatically' that:

- measurement is **monotone**:

$$\mu : \mathsf{D}X, E \subseteq F \vdash \underset{\mu}{\mathrm{Ce}}[E] \leq \underset{\mu}{\mathrm{Ce}}[F]$$

- the **inclusion-exclusion** principle:

$$\mu : \mathsf{D}X, E, F : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Ce}}[E \cup F] + \underset{\mu}{\mathrm{Ce}}[E \cup F] = \underset{\mu}{\mathrm{Ce}}[E] + \underset{\mu}{\mathrm{Ce}}[F]$$

Axioms for events and distributions

Consider posets:

$$\omega \coloneqq (\mathbb{N}, \leq) \qquad (\mathcal{B}_X, \subseteq) \qquad (\mathbb{W}, \leq)$$

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

$$P^{\omega} \coloneqq \left\{ p_{\_} \in \underline{P}^{\mathbb{N}} \middle| p_0 \leq p_1 \leq \cdots \right\}$$

Chain-closure of events and weights

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

Scott-continuity of measurement

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

Axiom for probability

Probability distributions have total mass one

$$\mathsf{P}X := \{\mu \in \mathsf{D}X | \mathrm{Ce}_{\mu}[X] = 1\} \qquad \mu : \mathsf{P}X \vdash \mathsf{cast}\,\mu : \mathsf{D}X$$

i.e., if we define:

$$\square := [0,1] \qquad \mu : \mathsf{P}X, E : \mathcal{B}_X \vdash \Pr_{\mu}[E] := \mathop{\mathrm{Ce}}_{\mathsf{cast}\,\mu}[E] : \square$$

then:

$$\mu : \mathsf{P}X \vdash \Pr_{\mu}[X] = 1$$

Integration

Lebesgue integration w.r.t. a distribution

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

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

Expectation w.r.t. a probability distribution

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

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 (now)

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

Lecture 2: the full model

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

course page

ask questions on the Scottish PL Institute Zulip stream #qbs

Discrete model

X: types denote **sets**

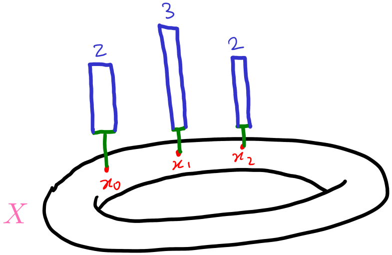
$\mathrm{D}X$: set of **histograms**:

Discrete model

X: types denote **sets**

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

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



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

Countably supported distributions

Support

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

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

The subsets supporting a weight function μ are closed under intersections.

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

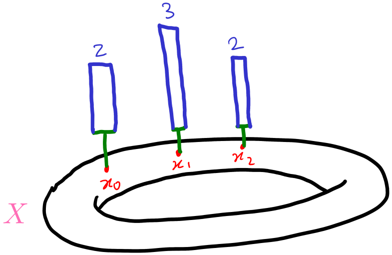
$$\mu : \mathbb{W}^X \vdash \operatorname{supp} \mu \coloneqq \{x \in X | \mu\, x > 0\}$$

Discrete model

X: types denote **sets**

DX: set of **histograms**:

$$\mathrm{D}X := \{\mu : X \to \mathbb{W} | \mu \text{ is } \textbf{countably supported} \}$$
$$:= \{\mu : X \to \mathbb{W} | \exists S \in \mathcal{P}X . S \text{ is countable}\}$$
$$:= \{\mu : X \to \mathbb{W} | \operatorname{supp} \mu \text{ is countable}\}$$



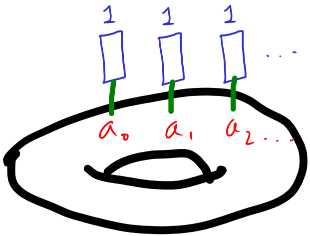
$\mu\, x_0 = 2 \quad \mu\, x_1 = 3 \quad \mu\, x_2 = 2$

Example distributions

Counting distribution

Counts the outcomes in a countable subset:

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

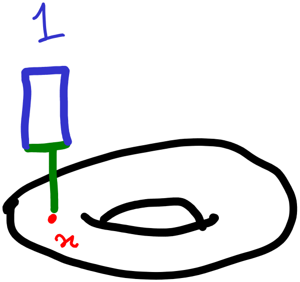


Example distributions

Dirac

A point mass:

$$x : X \vdash \boldsymbol{\delta}_x := \left(\lambda x'.\begin{cases} x' = x : & 1 \\ x' \neq x : & 0 \end{cases}\right) : \mathrm{D}X$$



(NB: $x : X \vdash \boldsymbol{\delta}_x = \#_{\{x\}}$.)

Example distributions

Zero

No mass anywhere:

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

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

Discrete model

X: types denote **sets**

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

$$\mathsf{D}X := \{\mu : X \to \mathbb{W} | \mu \text{ is } \textbf{countably supported}\}$$

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

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

Measurement: weighted sum of all (supported) outcomes:

$$\mu : \mathsf{D}X, E : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Ce}}[E] := \sum_{x \in E} \mu\, x$$

$$:= \sum_{x \in E \cap \mathrm{supp}\, \mu} \mu\, x$$

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

Example measurements

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

Counting distribution

counts supported outcomes

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

Example measurements

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

Counting distribution

counts supported outcomes

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

Dirac

detects given outcome:

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

Example measurements

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

Counting distribution

counts supported outcomes

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

Dirac

detects given outcome:

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

Zero

measures every event as zero:

$$E : \mathcal{B}_X \vdash \mathrm{Ce}_{\mathbf{0}}[E] = 0$$

The discrete model validates the axioms

Exercise

$$\mu : \mathsf{D} \qquad \vdash \underset{\mu}{\mathrm{Ce}}[\emptyset] = 0$$

$$E, C : \mathcal{B}_X, \mu : \mathsf{D} \qquad \vdash \underset{\mu}{\mathrm{Ce}}[E] = \underset{\mu}{\mathrm{Ce}}[E \cap C] + \underset{\mu}{\mathrm{Ce}}\left[E \cap C^{\complement}\right]$$

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

Parameterised distributions

Kernel

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

Kernels are open/parameterised distributions.

Examples

Dirac and the counting distribution form kernels:

$$\boldsymbol{\delta}_{\_} : X \rightsquigarrow \mathsf{D}X \qquad \#_{\_} : \mathcal{P}_{\mathrm{ctbl}} X \rightsquigarrow \mathsf{D}X$$

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

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

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

$$\oint \mathrm{d}\mu k := \lambda y. \sum_{x \in X} \mu\, x \cdot k(x; y)$$

$$:= \lambda y. \sum_{x \in \operatorname{supp} \mu} \mu\, x \cdot k(x; y)$$

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

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

Example

Weak Disintegration Problem (non-standard terminology)

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

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

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

Example

Weak Disintegration Problem (non-standard terminology)

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

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

Example disintegration

For $n \in \mathbb{N}$, write $\mathbf{Fin}\, n := \{0, \ldots, n-1\}$. For countable X, write $\# := \#_X : \mathsf{D}X$.
Here is a disintegration of $\# \in \mathsf{D}\left((\mathbf{Fin}\, 2)^{\mathbf{Fin}\,(n+1)}\right)$ w.r.t. $\# \in \mathsf{D}\,(\mathbf{Fin}\, 2)$:

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

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

Sub-type of probability distributions

Sub-types

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

$$\{x : X|\varphi\} \subseteq X \qquad y : \{x : X|\varphi\} \vdash \mathsf{cast}\, y \coloneqq y : X$$

we **lift** values in X that satisfy φ to the sub-type:

$$\frac{\Gamma \vdash M : X \qquad \Gamma \vdash \varphi\,[x \mapsto M]}{\Gamma \vdash \mathsf{lift} M : \{x : X|\varphi\}} \qquad \frac{\Gamma \vdash M : X \qquad \Gamma \vdash \{\varphi\}\, x \mapsto M}{\Gamma \vdash \mathsf{cast}(\mathsf{lift} M) = M}$$

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

$$y : \{x \in X|\varphi\} \vdash \mathsf{lift}(\mathsf{cast}\, y) = y \qquad y : \{x \in X|\varphi\} \vdash \varphi\,[x \mapsto \mathsf{cast}\, y]$$

i.e., the lifting is unique and elements in the sub-type satisfy φ.

Sub-type of probability distributions

Magnitude and probability distributions

$$\mu : \mathsf{D}X \vdash \|\mu\| \coloneqq \underset{\mu}{\mathrm{Ce}}[X] : \mathbb{W} \qquad \mathsf{P}X \coloneqq \{\mu \in \mathsf{D}X \mid \|\mu\| = 1\} \qquad \square \coloneqq [0,1] \coloneqq \{w \in \mathbb{W} \mid w \leq 1\}$$

Event probability

$$\mu : \mathsf{P}X, E : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Pr}}[E] \coloneqq \mathsf{lift}\left(\underset{\mathsf{cast}\,\mu}{\mathrm{Ce}}[E]\right) : \square$$

Stochastic kernel

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

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 $\boldsymbol{\delta}_{\_} : X \to \mathsf{D}X$ lift along cast:

$$x : X \vdash \|\boldsymbol{\delta}_x\| = \underset{\boldsymbol{\delta}_x}{\mathrm{Ce}}[X] = 1$$

so we can overload:

$$\begin{array}{ccc} & \overset{\boldsymbol{\delta}_{\_}}{\dashrightarrow} & \mathsf{P}X \\ X & =: & \downarrow \mathrm{cast} \\ & \underset{\boldsymbol{\delta}_{\_}}{\longrightarrow} & \mathsf{D}X \end{array}$$

Kock integrals of stochastic kernels by probability distributions lift along cast:

$$\mu : \mathsf{P}X, k : (\mathsf{P}Y)^X \vdash \mathrm{Ce}_{\oint(\mathsf{cast}\,\mu)(\mathrm{d}x)\,\mathsf{cast}(k\,x)}[Y] = 1$$

so we can overload:

$$\begin{array}{ccc} (\mathsf{P}X) \times (\mathsf{P}Y)^X & \overset{\oint}{\dashrightarrow} & \mathsf{P}Y \\ \mathrm{cast} \times (\mathrm{cast}\,\circ) \downarrow & =: & \downarrow \mathrm{cast} \\ (\mathsf{D}X) \times (\mathsf{D}Y)^X & \underset{\oint}{\longrightarrow} & \mathsf{D}Y \end{array}$$

Proposition

The triple $\left(\mathsf{D}, \boldsymbol{\delta}_{-}, \oint\right)$ forms a monad over $\mathbf{Set}$:

$$\begin{array}{ll} x : X, k : (\mathsf{D}Y)^X & \vdash \oint \mathrm{d}\boldsymbol{\delta}_x k = k\,x \\ \mu : \mathsf{D}X & \vdash \oint \mu(\mathrm{d}x)\boldsymbol{\delta}_x = \mu \\ \mu : \mathsf{D}X, k : (\mathsf{D}Y)^X, \ell : (\mathsf{D}Z)^Y & \vdash \oint \left(\oint \mu(\mathrm{d}x) k\,x\right)(\mathrm{d}y)\ell\,y = \oint \mu(\mathrm{d}x) \oint k(x; \mathrm{d}y)\ell\,y \end{array}$$

Corollary

The triple $\left(\mathsf{P}, \boldsymbol{\delta}_{-}, \oint\right)$ forms a monad over $\mathbf{Set}$.

Weighted average

Lebesgue integral

Integration is the raison d'être for distributions:

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

In the **discrete model**:

$$\int \mathrm{d}\mu f \coloneqq \sum_{x \in X} (\mu\, x) \cdot (f\, x) \coloneqq \sum_{x \in \operatorname{supp} \mu} (\mu\, x) \cdot (f\, x)$$

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

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

Weighted average

Expectation

To emphasise that some μ is a probability distribution, we will use the notation:

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

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:

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

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:

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

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:

$$\frac{x : X \vdash M : \mathbb{B}}{\{x \in X | M\} : \mathcal{B}_X} \qquad \{x \in X | \varphi\, x\} := \{x \in X | \varphi\, x = \textbf{True}\}$$

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:

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

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

Notation:

$$\frac{\Gamma \vdash \mu : \mathsf{D}X \qquad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \underset{x\sim\mu}{\mathrm{Ce}}[M] \coloneqq \underset{\mu}{\mathrm{Ce}}[\{x \in X | M\}] : \mathbb{W}}$$

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

Language of probability & distribution (recap)

X	type of **values**/**outcomes**
$\mathsf{D}X$	type of **distributions**/**measures** over X
$\mathsf{P}X \subseteq \mathsf{D}X$	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.:

$$\mu : \mathsf{D}X, E : \mathcal{B}_X \vdash \underset{\mu}{\mathrm{Ce}}[E] : \mathbb{W}$$

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

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

$$y_1, y_2 : Y \vdash y_1 \overset{Y}{=} y_2 : \mathsf{Prop} \qquad \mu : \mathsf{P}X, E : \mathcal{B}_X \vdash \mathsf{cast}\,\underset{\mu}{\mathrm{Pr}}[E] = \underset{\mu}{\mathrm{Ce}}[E]$$

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

Lecture plan

Lecture 1: discrete model (now)

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

Lecture 2: the full model

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

course page

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 $(\otimes) : \mathrm{D}X \times \mathrm{D}Y \to \mathrm{D}(X \times Y)$
- Random elements and their laws (push-forward measure):
  $(\lambda(\mu, \alpha).\mu_{\alpha}) : \mathrm{D}\Omega \times X^{\Omega} \to \mathrm{D}X$

NB:

- Dirac kernel $\delta_{\_} : X \to \mathrm{D}X$
- Kock integration
  $\oint : \mathrm{D}X \times (\mathrm{D}Y)^{\mathrm{D}X} \to \mathrm{D}Y$

Standard vocabulary

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

Simply-typed foundations for probabilistic modelling

Compositional building blocks for modelling

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

NB:

- Dirac kernel $\delta_{\_} : X \to \mathrm{D}X$
- Kock integration $\oint : \mathrm{D}X \times (\mathrm{D}Y)^{\mathrm{D}X} \to \mathrm{D}Y$

Standard vocabulary

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

Affine combinations of distributions: scaling

Scaling distributions

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

In the discrete model:

$$w \cdot \mu := \lambda x. w \cdot \mu\, x \qquad \operatorname{supp}(w \cdot \mu) \subseteq \operatorname{supp} \mu$$

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

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

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

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

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

$$w : \mathbb{W}, \mu : \mathsf{D}X, E : \mathcal{B}_X \vdash \underset{w \cdot \mu}{\mathrm{Ce}}[f] = w \cdot \underset{\mu}{\mathrm{Ce}}[f]$$

Affine combinations of distributions: scaling

Normalisation

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

measurement is homogeneous
↓

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:

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

In the discrete model:

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

so: $x : X \vdash \mathbf{U}_{\{x\}} = \boldsymbol{\delta}_x$.

Unit type

$$\mathbb{1} \coloneqq \{()\}$$

Proposition

The following two functions are mutually inverse:

$$\mathrm{D}\mathbb{1} \underset{(\cdot\boldsymbol{\delta}_{()})}{\overset{\|-\|}{\rightleftarrows}} \mathbb{W}$$

Proof

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

Proposition

We can recover Lebesgue integration from Kock integration:

$$\begin{array}{ccc}
\mathsf{D}X \times \mathbb{W}^{X} & \xrightarrow{\ \mathrm{id} \times (\cong \circ)\ } & \mathsf{D}X \times (\mathsf{D}\mathbb{1})^{X} \\
{\scriptstyle \int}\big\downarrow & = & \big\downarrow{\scriptstyle \fint} \\
\mathbb{W} & \xleftarrow[\ \cong\ ]{} & \mathsf{D}\mathbb{1}
\end{array}$$

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

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

In the discrete model:

$$\sum_{i \in I} \mu_i \coloneqq \lambda x. \sum_{i \in I} \mu_i\, x \qquad \operatorname{supp} \sum_{i \in I} \mu_i = \bigcup_{i \in I} \operatorname{supp} \mu_i$$

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:

$$p : \mathbb{I} \vdash \mathbf{B}p = \mathsf{lift}\,(p \cdot \boldsymbol{\delta}_{\mathbf{True}} + (1 - p) \cdot \boldsymbol{\delta}_{\mathbf{False}}) : \mathsf{P}\mathbb{B}$$

Affinity of integration and convexity of expectation

Theorem (Multi-linearity)

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

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

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

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

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

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

Weight arithmetic

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

Proposition

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

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

and $(\cdot) : \mathbb{W} \times \mathsf{D}X \to \mathsf{D}X$ makes each $\mathsf{D}X$ 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:

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

$$\sum_{i \in I} w_i \cdot \mu_i := \mathsf{lift} \sum_{i \in I} w_i \cdot (\mathsf{cast}\,\mu_i) : \mathsf{P}X$$

Proof

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

Convex combinations of probability distributions

Corollary (Multi-convexity)

Stochastic Kock integration, expectation and measurement are convex:

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

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

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

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

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

Products

Product distribution

$$\mu : \mathsf{D}X, \nu : \mathsf{D}Y \vdash \quad \mu \otimes \nu := \oint \mu(\mathrm{d}x) \oint \nu(\mathrm{d}y) \boldsymbol{\delta}_{(x,y)} : \mathsf{D}(X \times Y)$$

In the discrete model:

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

Example: counting distribution on product space

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

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

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

Products

Notation:

$$\frac{\Gamma \vdash M : \mathsf{D}(X \times Y) \qquad \Gamma, x : X, y : Y \vdash K : \mathsf{D}Z}{\Gamma \vdash \oiint M(\mathrm{d}x, \mathrm{d}y)K \coloneqq \oint \mathrm{d}M(\lambda\,(x, y)\,.K) : \mathsf{D}Z}$$

Theorem (Fubini-Tonelli)

We can integrate products in any order:

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

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

Applying Fubini-Tonelli

Theorem (Rule of Product)

We can factor out products:

$$\mu : \mathsf{D}X, f : \mathbb{W}^X, \nu : \mathsf{D}Y, g : \mathbb{W}^Y \vdash \quad \iint (\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y) f x \cdot g y = \left(\int \mathrm{d}\mu f\right) \cdot \left(\int \mathrm{d}\nu g\right)$$

$$\mu : \mathsf{D}X, E : \mathcal{B}_X, \nu : \mathsf{D}Y, F : \mathcal{B}_Y \vdash \quad \underset{\mu\otimes\nu}{\mathrm{Ce}} [E \times F] = \underset{\mu}{\mathrm{Ce}} [E] \cdot \underset{\nu}{\mathrm{Ce}} [F]$$

Theorem

The product lifts to probability distributions:

$$\mu : \mathsf{P}X, \nu : \mathsf{P}Y \vdash (\mu \otimes \nu) \coloneqq \mathsf{lift}(\mathsf{cast}\, \mu \otimes \mathsf{cast}\, \nu) : \mathsf{P}(X \times Y)$$

Products

Binomial distribution

the number of successful outcomes of n independent Bernoulli trials:

$$\mathbf{B}_n : \mathbb{I} \rightsquigarrow \mathsf{P}(\mathbf{Fin}\,(1+n)) \qquad \mathbf{B}_0 p \coloneqq \boldsymbol{\delta}_0 : \mathsf{P}(\mathbf{Fin}\,1)$$

$$\mathbf{B}_{1+n} p \coloneqq \oiint (\mathbf{B}_n p \otimes \mathbf{B} p)(\mathrm{d}c, \mathrm{d}b)\,(\mathrm{if}\; b \;\mathrm{then}\; \boldsymbol{\delta}_{1+c} \;\mathrm{else}\; \boldsymbol{\delta}_c) : \mathsf{P}(\mathbf{Fin}\,(2+n))$$

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

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

Push-forward distributions

Random element

in X any (internal) function:

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

Law

of a random element is the distribution:

$$\mu : \mathsf{D}\Omega, \alpha : X^{\Omega} \vdash \mu_{\alpha} := \oint \mu(\mathrm{d}\omega)\boldsymbol{\delta}_{\alpha\,\omega} : \mathsf{D}X$$

Example

Represent outcomes of die roll by $\mathsf{D6} := \{1, 2, \ldots, 6\}$, and two rolls by $\mathsf{D6} \times \mathsf{D6}$.
The sum of the rolls is a random element:

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

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

Push-forward distributions

Theorem (Law of the Unconcious Statistician)

Formulae for reparameterising integration and measurement:

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

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

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

Simply-typed foundations for probabilistic modelling

Compositional building blocks for modelling

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

NB:

- Dirac kernel $\delta_{\_} : X \to \mathrm{D}X$
- Kock integration
  $\oint : \mathrm{D}X \times (\mathrm{D}Y)^{\mathrm{D}X} \to \mathrm{D}Y$

Standard vocabulary

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

Standard vocabulary: concepts concerning products

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

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

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:

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

More generally, for finite I:

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

Independence

Example [Durett]

Model 3 independent coin tosses:

$$\mathsf{Toss} \coloneqq \{\mathbf{Head}, \mathbf{Tail}\} \qquad \Omega \coloneqq \mathsf{Toss}^3 \qquad \mu \coloneqq \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} : \mathsf{P}\Omega$$

The outcome of the i^{th} coin toss is the random element $\pi_i : \Omega \to \mathsf{Toss}$.
Consider the Boolean proposition in which the i^{th} and j^{th} tosses $(i \neq j)$ agree:

$$\mathsf{Same}_{ij} \coloneqq \lambda\omega.\pi_i\omega = \pi_j\omega : \Omega \to \mathbb{B}$$

Calculate:

$$\Pr_{\mu}[\mathsf{Same}_{12}] \overset{\text{LOTUS}}{=} \Pr_{(x,y)\sim\mu_{(\pi_1,\pi_2)}}[x = y] \overset{\text{marginalisation}}{=} \Pr_{(x,y)\sim\mathbf{U}\otimes\mathbf{U}}[x = y] \overset{\text{Fubini}}{=} \int \mathbf{U}(\mathrm{d}x) \Pr_{y\sim\mathbf{U}}[x = y]$$

$$= \frac{1}{2} \cdot \Pr_{y\sim\mathbf{U}}[\mathbf{Head} = y] + \frac{1}{2} \cdot \Pr_{y\sim\mathbf{U}}[\mathbf{Tail} = y] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Independence

Example [Durett]

Model 3 independent coin tosses:

$$\mathsf{Toss} \coloneqq \{\mathbf{Head}, \mathbf{Tail}\} \qquad \Omega \coloneqq \mathsf{Toss}^3 \qquad \mu \coloneqq \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} : \mathsf{P}\Omega$$

The outcome of the i^{th} coin toss is the random element $\pi_i : \Omega \to \mathsf{Toss}$.
Consider the Boolean proposition in which the i^{th} and j^{th} tosses $(i \neq j)$ agree:

$$\mathsf{Same}_{ij} \coloneqq \lambda\omega.\pi_i\omega = \pi_j\omega : \Omega \to \mathbb{B}$$

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

Independence

π_1, Same_{12}, and Same_{13} determine π_2, π_3, so:

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

Fubini-Tonelli

$\downarrow$

$$= \int \mathbf{U}_{\mathsf{Toss}}(\mathrm{d}b_1) \Pr_{(b_2,b_3)\sim(\mathbf{U}\otimes\mathbf{U})}[\mathsf{Same}_{12}(b_1,b_2,b_3) = \mathbf{True}, \mathsf{Same}_{13}(b_1,b_2,b_3) = \mathbf{True}]$$

$$= \frac{1}{2} \Pr_{(b_2,b_3)\sim(\mathbf{U}\otimes\mathbf{U})}[\mathsf{Same}_{12}(\mathbf{Head},b_2,b_3) = \mathbf{True}, \mathsf{Same}_{13}(\mathbf{Head},b_2,b_3) = \mathbf{True}]$$

$$+ \frac{1}{2} \Pr_{(b_2,b_3)\sim(\mathbf{U}\otimes\mathbf{U})}[\mathsf{Same}_{12}(\mathbf{Tail},b_2,b_3) = \mathbf{True}, \mathsf{Same}_{13}(\mathbf{Tail},b_2,b_3) = \mathbf{True}]$$

$$= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

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

Independence

Example [Durett]

Model 3 independent coin tosses:

$$\mathsf{Toss} \coloneqq \{\mathbf{Head}, \mathbf{Tail}\} \qquad \Omega \coloneqq \mathsf{Toss}^3 \qquad \mu \coloneqq \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} \otimes \mathbf{U}_{\mathsf{Toss}} : \mathsf{P}\Omega$$

The outcome of the i^{th} coin toss is the random element $\pi_i : \Omega \to \mathsf{Toss}$.
Consider the Boolean proposition in which the i^{th} and j^{th} tosses $(i \neq j)$ agree:

$$\mathsf{Same}_{ij} \coloneqq \lambda\omega.\pi_i\omega = \pi_j\omega : \Omega \to \mathbb{B}$$

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

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

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

Distribution preservation

Distribution space (Ω, μ)

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

Distribution preserving function

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

$$f : \Omega_1 \to \Omega_2 \qquad (\mu_1)_f = \mu_2$$

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

Example

Consider the swapping function: $\mathsf{swap} \coloneqq (\lambda\,(x, y)\,.\,(y, x)) : X \times Y \to Y \times X$. Then, for each $\mu : \mathrm{D}X$, $\nu : \mathrm{D}Y$, swapping is distribution preserving function:

$$\mathsf{swap} : (X \times Y, \mu \otimes \nu) \to (Y \times X, \nu \otimes \mu)$$

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

Density and scaling

Density

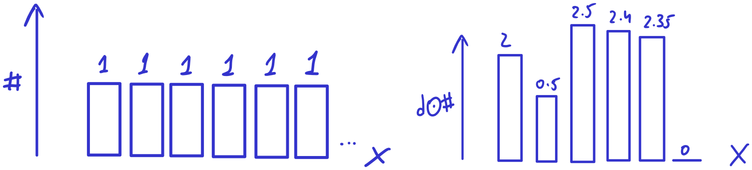
over X is any weight function $f : X \to \mathbb{W}$.

Density scaling

We can scale a distribution by a density:

$$f : \mathbb{W}^X, \mu : \mathsf{D}X \vdash f \odot \mu := \oint \mu(\mathrm{d}x)(f, x) \cdot \boldsymbol{\delta}_x : \mathsf{D}X$$

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 $f : X \to \mathbb{W}$.

Density scaling

We can scale a distribution by a density:

$$f : \mathbb{W}^X, \mu : \mathsf{D}X \vdash f \odot \mu := \oint \mu(\mathrm{d}x)(f, x) \cdot \boldsymbol{\delta}_x : \mathsf{D}X$$

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:

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

- f is the **density** of μ w.r.t. ν
- f is a **Radon-Nikodym derivative** of μ w.r.t. ν.

Absolute continuity

μ is **absolutely continuous** w.r.t. ν when μ has a density w.r.t. ν:

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

Example

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

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

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:

$$\frac{\Gamma \vdash \mu : \mathsf{D}X \qquad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \mu(\mathrm{d}x)\ \text{almost certainly}\ M \coloneqq [M] \odot \mu = \mu : \mathsf{Prop}}$$

For probabilities we define:

$$\frac{\Gamma \vdash \mu : \mathsf{P}X \qquad \Gamma, x : X \vdash M : \mathbb{B}}{\Gamma \vdash \mu(\mathrm{d}x)\ \text{almost surely}\ M \coloneqq (\mathsf{cast}\,\mu)(\mathrm{d}x)\ \text{almost certainly}\ M : \mathsf{Prop}}$$

Existence and almost-sure uniqueness of densities

Theorem (Radon-Nikodym)

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

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

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

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

Example

For all countable X, we have:

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

Indeed, apply the Radon-Nikodym theorem, since $\operatorname{supp}\# = X$.
Constructively, direct calculation shows: $(\lambda x.\mu\,x) = \frac{\mathrm{d}\mu}{\mathrm{d}\#}$.

Simply-typed foundations for probabilistic modelling

Compositional building blocks for modelling

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

NB:

- Dirac kernel $\delta_{\_} : X \to \mathrm{D}X$
- Kock integration
  $\oint : \mathrm{D}X \times (\mathrm{D}Y)^{\mathrm{D}X} \to \mathrm{D}Y$

Standard vocabulary

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

Lecture plan

Lecture 1: discrete model (now)

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

Lecture 2: the full model

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

course page

ask questions on the Scottish PL Institute Zulip stream #qbs

Type dependencies

Example: Binomial kernels

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

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

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

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

Family model

Family over an indexing set I

consists of a seqeuence $X_{\_} = (X_i)_{i \in I}$ 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:

$$\mathbf{Fin}\, n \coloneqq \{0, 1, \ldots, n - 1\}$$

Family model

Family over an indexing set I

consists of a seqeuence $X_{\_} = (X_i)_{i \in I}$ 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

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

$$\theta : I \to J \qquad (f_i : X_i \to Y_{\theta\, i})_{i \in I}$$

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:

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

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:

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

Family of distributions

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

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

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

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

Both have a **Dirac** distribution:

$$i : I \vdash \boldsymbol{\delta}\_ : X_i \to \mathsf{D}X_i \qquad i : I \vdash \boldsymbol{\delta}\_ : X_i \to \mathsf{P}X_i$$

Extension and dependent pairs

Extension

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

$$\coprod_{i \in I} X_i \coloneqq \bigcup_{i \in I} \{i\} \times X_i = \left\{ (i, x) \in I \times \bigcup_{i \in I} X_i \middle| x \in X_i \right\}$$

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

Dependent pairs

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

Functions and kernels

Dependent functions

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

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

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

are dependent elements:

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

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

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

Dependent Kock integral

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

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

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

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

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

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

Random variables

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

Signed and unsigned random variable

in a probability space (Ω, μ) are random elements $\alpha : \Omega \to \overline{\mathbb{R}}$ and $\alpha : \Omega \to \mathbb{W}$.
The **positive** and **negative parts** are unsigned random variables $-^{\pm} : \overline{\mathbb{R}}^{\Omega} \to \mathbb{W}^{\Omega}$:

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

An unsigned r.v. α is **Lebesgue integrable** when its Lebesgue integral is finite:
$\int \mathrm{d}\mu\alpha < \infty$.
For a (signed) r.v. α, when either α^{+} or α^{-} is Lebesgue integrable, we define:

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

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

Random variable spaces

Lebesgue integrability is a Boolean property:

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

Lebesgue spaces ensemble

is the family:

$$i : I, p : [1,\infty), \mu : \mathsf{P}X_i \vdash \qquad \mathcal{L}_p(X_i, \mu) := \left\{\alpha : X_i \to \overline{\mathbb{R}} \middle| \alpha^p \text{ integrable}\right\}$$

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

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

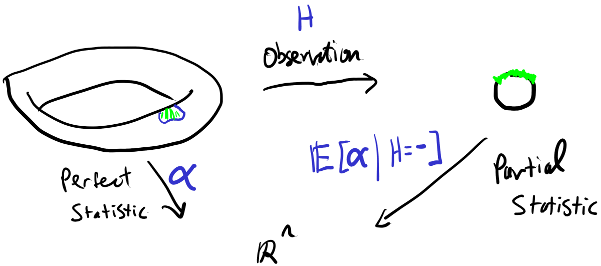
and the fibre 2 has an inner product (almost a Hilbert space!):

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

Conditioning á la Kolmogorov

Situation:

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



Conditional expectation of α along H w.r.t μ

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

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

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

Conditioning á la Kolmogorov

Theorem (Kolmogorov)

Every random variable has a conditional expectation:

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

Therefore:

Corollary (Internal conditional expectation)

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

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

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

Conditioning á la Kolmogorov

Conditional probability

of event is a conditional expectation of its characteristic function:

$$\mu : \mathsf{P}\Omega, H : \Omega \to X, E : \mathcal{B}_\Omega, \beta : \mathcal{L}_1(X, \mu_H) \vdash$$
$$\left(\beta = \Pr_{\mu}[E|H = -]\right) \coloneqq \left(\beta = \mathop{\mathbb{E}}_{\omega\sim\mu}[\omega \in E|H = -]\right) : \mathsf{Prop}$$

Regular conditional probability

a kernel that agrees with the conditional expectation of the characteristic functions:

$$\mu : \mathsf{P}\Omega, H : \Omega \to X, k : X \rightsquigarrow \Omega \vdash$$
$$\left(k = \Pr_{\mu}[-|H = -]\right) \coloneqq \left(\forall E \in \mathcal{B}_\Omega . k(-; E) = \mathop{\mathbb{E}}_{\omega\sim\mu}[\omega \in E|H = -]\right) : \mathsf{Prop}$$

Conditioning via disintegration

Kolmogorov's theorem does **not** ensure the existence of a regular conditional probability, although the constructive, discrete, definition does.

Disintegration Problem (warning: conflicting terminologies in literature)

Input: probability distribution $\mu : \mathsf{P}\Omega$, measurable map $H : \Omega \to \Theta$
induce law $\nu := \mu_H : \mathsf{P}\Theta$

Output: probability kernel $k : \Theta \rightsquigarrow \Omega$ such that: $\mu = \oint \mathrm{d}\nu k$.

We call k a **disintegration** of μ along H.

Proposition

Consider a probability kernel $k : \Theta \rightsquigarrow \Omega$. TFAE:

- k is a disintegration of μ along $H : \Omega \to \Theta$;
- k is a regular conditional probability kernel of μ conditioned on H.

Conditioning via disintegration

Fibred disintegration of $\mu : \mathsf{P}\left(\coprod_{\Theta} \Omega\right)$ (non-standard terminology and formulation)

a partial dependent kernel $k : (\theta : \Theta) \rightsquigarrow \Omega_{\bot}$, defined μ_{dep}-a.s., that disintegrates μ along the first projection $\mathsf{dep} : \left(\coprod_{\Theta} \Omega\right) \to \Theta$:

$$\mu : \mathsf{P}\left(\coprod_{\Theta} \Omega\right), k : \Theta \rightsquigarrow \Omega_{\bot} \vdash k \text{ disintegrates fibres of } \mu \coloneqq$$

$$\mu_{\mathsf{dep}}(\mathrm{Dom}\,(k)) = 1, \mu = \oint \mathrm{d}\mu_{\mathsf{dep}} k : \mathsf{Prop}$$

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

$$-^{\dagger} : \left(\mu : \mathsf{P}\left(\coprod_{\Theta} \Omega\right)\right) \to \{k : (\theta : \Theta) \rightsquigarrow \Omega_{\bot} | k \text{ disintegrates } \mu \text{ along } \mathsf{dep}\}$$

$$\mathrm{Dom}\left(\mu^{\dagger}\right) \coloneqq \{\theta | \mu_{\mathsf{dep}}\theta > 0\} \qquad \mu^{\dagger} \coloneqq \lambda\theta. \frac{1}{\mu_{\mathsf{dep}}\theta} \odot \mu|_{\mathsf{dep}^{-1}[\theta]}$$

Bayes's Theorem (adapted from Williams)

Let:

- $\lambda : \mathsf{P}(X \times \Theta)$ be a joint probability distribution.
- $\mu : \mathsf{D}X$, $\nu : \mathsf{D}\Theta$ be distributions such that $\lambda \ll \mu \otimes \nu$ $\quad X \xleftarrow{\alpha := \pi_1} X \times \Theta \xrightarrow{H := \pi_2} \Theta$
- $w_{\alpha,H} = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu\otimes\nu} : X \times \Theta \to \mathbb{W}$ a Radon-Nikodym derivative

Observation 1

- $w_\alpha := \lambda x. \int \nu(\mathrm{d}\theta) w_{\alpha,H}(x,\theta) : X \to \mathbb{W}$ then: $w_\alpha = \frac{\mathrm{d}\lambda_\alpha}{\mathrm{d}\mu}$
- $w_H := \lambda\theta. \int \mu(\mathrm{d}x) w_{\alpha,H}(x,\theta) : \Theta \to \mathbb{W}$ then: $w_H = \frac{\mathrm{d}\lambda_H}{\mathrm{d}\nu}$

Observation 2

Let: $w_\alpha(- \mid H = -) : X \times \Theta \to \mathbb{W}$ $\qquad w_\alpha(x \mid H = \theta) := \begin{cases} w_H\theta > 0 : & \frac{w_{\alpha,H}(x,\theta)}{w_H\theta} \\ \text{otherwise:} & 0 \end{cases}$

$\lambda_{\alpha|H=-} : \Theta \rightsquigarrow X \qquad \lambda_{\alpha|H=\theta} := \lambda_\alpha(- \mid H = \theta) \odot \nu$. Then:

$$\lambda_{\alpha|H=-} = \Pr_\lambda[-|H = -] \qquad \text{(Bayes's formula)}$$

Lecture plan

Lecture 1: discrete model

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

Lecture 2: the full model (now)

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

course page

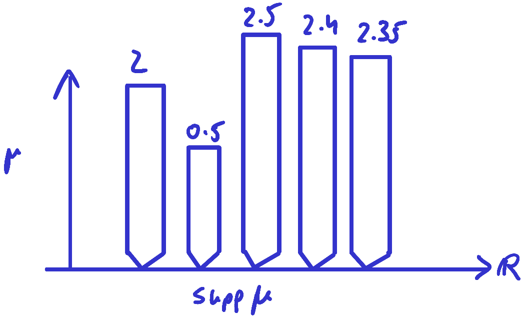
ask questions on the
Scottish PL Institute
Zulip stream #qbs

From histograms to measures

The **discrete** model expresses **histograms** only.

Also want **continuous** distributions:

- lengths
- areas
- volumes



Continuous caveat

Theorem (Vitali 1905)

There is no reasonable generalisation of 'length' that measures all subsets of the real line—there is no function $\lambda : \mathcal{P}\mathbb{R} \to \mathbb{W}$ satisfying:

$$\lambda[a,b] = (b-a) \qquad \lambda(s + [E]) = \lambda E \qquad \lambda\left(\biguplus_{i=0}^{\infty} E_n\right) = \sum_{i=0}^{\infty} \lambda E_n$$

(generalise length) (translation invariance) (σ-additivity)

Takeaway

$\mathcal{B}_{\mathbb{R}} := \mathcal{P}\mathbb{R}$ as in the **discrete** model excludes **length, area, volume** as distributions.

$\implies$ need a different model

Workaround

Only measure **well-behaved** subsets:

Borel subsets $\mathcal{B}_{\mathbb{R}} \subseteq \mathcal{P}\mathbb{R}$

smallest σ**-field** containing all **open intervals**:

$$\frac{}{\emptyset \in \mathcal{B}_{\mathbb{R}}} \qquad \frac{E \in \mathcal{B}_{\mathbb{R}}}{E^{\complement} \in \mathcal{B}_{\mathbb{R}}} \qquad \frac{E_{-} \in \mathcal{B}_{\mathbb{R}}^{\mathbb{N}}}{\bigcup_{n\in\mathbb{N}} E_n \in \mathcal{B}_{\mathbb{R}}} \qquad \frac{a, b \in \mathbb{R}}{(a,b) \in \mathcal{B}_{\mathbb{R}}}$$

(empty set) (complements) (countable unions) (intervals)

Examples

- Countable discrete subsets are Borel:

$$\{r\} = \bigcap_{\varepsilon\in\mathbb{Q}_{>0}} (r-\varepsilon, r+\varepsilon) \in \mathcal{B}_{\mathbb{R}} \quad , \quad I \text{ countable} \implies I = \bigcup_{i\in I}\{i\}$$

- Any interval is Borel, e.g.: $[a,b) = (a,b) \cup \{a\}$

Measure theory: generalise the **worst-case** scenario

Measurable space $M = (\underline{M}, \mathcal{B}_M)$

set of **points** $a \in \underline{M}$ equipped with a **σ-field** $\mathcal{B}_M \subseteq \mathcal{P}\underline{M}$:

$$\frac{}{\emptyset \in \mathcal{B}_{\mathbb{R}}} \qquad \frac{E \in \mathcal{B}_{\mathbb{R}}}{E^{\complement} \in \mathcal{B}_{\mathbb{R}}} \qquad \frac{E_{-} \in \mathcal{B}_{\mathbb{R}}^{\mathbb{N}}}{\bigcup_{n \in \mathbb{N}} E_n \in \mathcal{B}_{\mathbb{R}}}$$

(empty set) (complements) (countable unions)

Examples

- Discrete spaces: $\overline{I}^{\mathbf{Meas}} := (I, \mathcal{P}I)$
- Sub-spaces: $\dfrac{S \subseteq \underline{M}}{S_M := (S, [\mathcal{B}_M] \cap S)}$ i.e., $\mathcal{B}_{S_M} := \{E \cap S | E \in \mathcal{B}_M\}$, e.g., $[0,\infty) \hookrightarrow \mathbb{R}$
- Products: $\mathcal{B}_{\prod_{i \in I} M_i} := \sigma \bigcup_{i \in I} \pi_i^{-1}[\mathcal{B}_{M_i}] = \sigma \left\{ \bigtimes_{i \in I} E_i \middle| \begin{array}{l} E_{-} \in \prod_{i \in I} \mathcal{B}_{M_i}, \\ \exists J \subseteq_{\text{countable}} I. \\ \forall j \notin J. E_j = \underline{M}_i \end{array} \right\}$, e.g.: $\mathbb{R}^n$

Measure theory

Borel measurable function $f : M \to K$

function sending points to points and measurable subsets to measurable subsets:

$$f : \underline{M} \to \underline{K} \qquad \mathcal{B}_M \ni f^{-1}[E] \Longleftarrow E \in \mathcal{B}_K$$

Examples

- $(+), (\cdot) : \mathbb{R}^2 \to \mathbb{R}$
- $|-|, \sin : \mathbb{R} \to \mathbb{R}$
- any continuous function $\mathbb{R}^n \to \mathbb{R}$
- any function out of a discrete space: $\dfrac{f : I \to \underline{M}}{f : \overline{I} \to M}$

Measure theory

Category **Meas**

Objects M: measurable spaces

Arrows $f : M \to K$: Borel measurable functions

$$\frac{}{\mathrm{id} \coloneqq (\lambda x.x) : M \to M} \qquad \frac{f : M \to K \qquad g : K \to L}{g \circ f : (\lambda x.g(f\,x)) : M \to M}$$

Categorical structure

Products, coproducts/disjoint unions, subspaces, projective and injective limits / categorical limits and colimits are all fine.

Theorem (Aumann'61)

There are no measurable spaces of Borel subsets nor of measurable functions over $\mathbb{R}$.
In detail, there are no σ-fields $\mathcal{B}_{\mathcal{B}_\mathbb{R}}$ and $\mathcal{B}_{\mathbb{R}\to\mathbb{R}}$ such that, letting $\mathcal{B}_\mathbb{R}$ and $\mathbb{R} \to \mathbb{R}$ be the corresponding measurable spaces, the following functions are measurable:

- Membership testing:
  $$(\in) := \left(\lambda r.E.\begin{cases} r \in E: & \mathbf{True} \\ \text{otherwise:} & \mathbf{False}\end{cases}\right) : \mathbb{R} \times \mathcal{B}_\mathbb{R} \to \overline{\{\mathbf{True}, \mathbf{False}\}}$$
- Evaluation: $\text{eval} := (\lambda\,(f, r)\,.f\,r) : (\mathbb{R} \to \mathbb{R}) \times \mathbb{R} \to \mathbb{R}$.

As a consequence, $\mathbf{Meas}$ is not Cartesian closed.

Aumann's Theorem: proof preliminaries

Recall the **Borel hierarchy** over a family of subsets $\mathcal{U} \subseteq \mathcal{P}X$, defined by transfinite induction on $\omega_1 + 1$, the successor of the first uncountable ordinal:

$$\mathbf{\Sigma}^{\mathcal{U}}_{\alpha}, \mathbf{\Pi}^{\mathcal{U}}_{\alpha}, \mathbf{\Delta}^{\mathcal{U}}_{\alpha} \subseteq \mathcal{P}X \qquad (\alpha \in \omega_1)$$

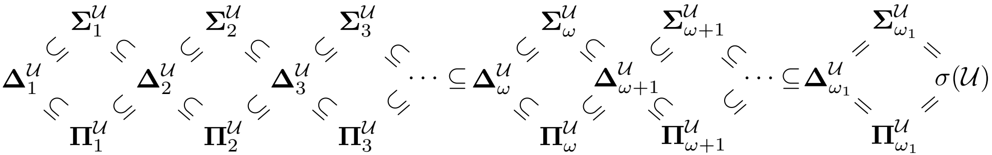
$$\mathbf{\Sigma}^{\mathcal{U}}_{1} := \mathcal{U}$$

$$\mathbf{\Sigma}^{\mathcal{U}}_{\alpha+1} := \left\{ \bigcup_{i \in I} A_i \middle| I \subseteq \mathbb{N}, A_{-} \in \mathcal{U} \cup \bigcup_{\beta \leq \alpha} \mathbf{\Pi}^{\mathcal{U}}_{\beta} \right\} \qquad (1 \leq \alpha \in \omega_1)$$

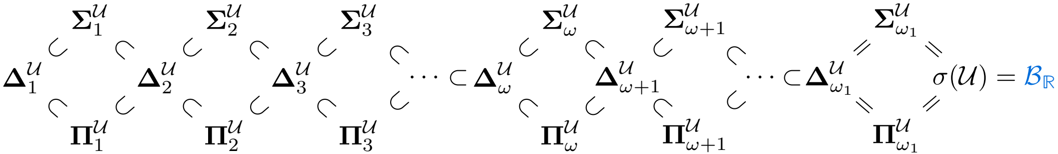
$$\mathbf{\Sigma}^{\mathcal{U}}_{\gamma} := \bigcup_{\beta < \gamma} \mathbf{\Sigma}^{\mathcal{U}}_{\beta} \qquad (1 \leq \gamma \text{ a limit ordinal in } \omega_1)$$

$$\mathbf{\Pi}^{\mathcal{U}}_{\alpha} := \left[\mathbf{\Sigma}^{\mathcal{U}}_{\alpha}\right]^{\complement} := \left\{ A^{\complement} \middle| A \in \mathbf{\Sigma}^{\mathcal{U}}_{\alpha} \right\} \qquad \mathbf{\Delta}^{\mathcal{U}}_{\alpha} := \mathbf{\Sigma}^{\mathcal{U}}_{\alpha} \cap \mathbf{\Delta}^{\mathcal{U}}_{\alpha}$$

The Borel hierarchy looks like this in general:



For $\mathcal{U} := \{(a,b) | a, b \in \mathbb{R}\}$, the hierarchy does not stabilise before ω_1:



Rank of $E \in \sigma\mathcal{U}$

first step in which it appears: $\mathrm{Rank}_E := \min\left\{\alpha < \omega_1 \middle| A \in \mathbf{\Delta}_\alpha^{\mathcal{U}}\right\}$.

Proof

Assume to the contrary there was some σ-field providing a measurable space of Borel subsets $\mathcal{B}_{\mathbb{R}}$ such that membership testing is measurable:

$$(\in) : \mathbb{R} \times \mathcal{B}_{\mathbb{R}} \to \overline{\{\mathbf{True}, \mathbf{False}\}} \qquad \text{NB: } \mathcal{B}_{\mathbb{R}\times\mathcal{B}_{\mathbb{R}}} = \sigma\left([\mathcal{B}_{\mathbb{R}}] \times [\mathcal{B}_{\mathcal{B}_{\mathbb{R}}}]\right)$$

Let $\alpha := \text{Rank}\,(\in)^{-1}[\mathbf{True}] < \omega_1$, and find $E \in \mathcal{B}_{\mathbb{R}}$ with $\text{Rank}_E > \alpha$. Then:

$$\alpha < \text{Rank}\, E = \text{Rank}\left(((\in) \circ (-, E))^{-1}[\mathbf{True}]\right) = \text{Rank}\left((-, E)^{-1}\left((\in)^{-1}[\mathbf{True}]\right)\right)$$
$$\leq \text{Rank}\left((\in)^{-1}[\mathbf{True}]\right) = \alpha$$

So $\alpha < \alpha$, a contradiction, and the postulated σ-field cannot exist. A similar proof replacing E with its characteristic function proves eval cannot be measurable. ∎

Sequences

By generalities, $(\overline{I} \to M) = \prod_{i \in I} M$. For countable I, we use $\overline{I} \to M$ for sequences.

Example

A sequence $a_- : \mathbb{N} \to \mathbb{R}$ is **Cauchy** when $\forall \varepsilon > 0. \exists N \in \mathbb{N}. \forall m, n > N. |a_n - a_n| < \varepsilon$.
We can define the Cauchy property through quantification over countable sets:

$$\mathbf{Cauchy} \in \mathcal{B}_{\mathbb{N} \to \mathbb{R}} \qquad \mathbf{Cauchy} := \bigcap_{\varepsilon \in \mathbb{Q}_{>0}} \bigcup_{N \in \mathbb{N}} \bigcap_{m,n \in \mathbb{N}} \{a_- \in \underline{\mathbb{N} \to \mathbb{R}} \mid |a_n - a_m| < \varepsilon\}$$

measurability through **type-checking**

Sequential Higher-order Structure:

I Countable: $V^{\ulcorner I \urcorner} = \prod_{i \in I} V$

$\implies$ Some higher-order structure in **Meas**:

Cauchy $\in B_{[-\infty,\infty]^{\mathbb{N}}}$

$$\mathrm{Cauchy} := \bigcap_{\varepsilon \in \mathbb{Q}^+} \bigcup_{k \in \mathbb{N}} \bigcap_{\substack{m,n \in \mathbb{N} \\ m,n \geq k}} \{ \vec{y} \in [-\infty,\infty]^{\mathbb{N}} \mid |y_m - y_n| < \varepsilon \}$$

$\lim\sup : [-\infty,\infty]^{\mathbb{N}} \to [-\infty,\infty]$ $\quad$ $\lim : \mathrm{Cauchy} \to \mathbb{R}$

handwritten-slides/popl24-tutorial-fest-slide-83.pdf

Compose higher-order building blocks:

lim is measurable!

$$\mathrm{VanishingSeq}(\mathbb{R}) := \{ \vec{r} \in \mathbb{R}^{\mathbb{N}} \mid \lim_{n\to\infty} r_n = 0 \} \in \mathcal{B}_{\mathbb{R}^{\mathbb{N}}}$$

$$\mathrm{approx}_{-} : \mathrm{VanishingSeq}(\mathbb{R}^{+}) \times \mathbb{R} \longrightarrow \mathbb{Q}^{\mathbb{N}}$$

$$\text{s.t.:} \quad \left| (\mathrm{approx}_{\vec{\Delta}}\, r)_n - r \right| < \Delta_n$$

Slogan: Measurable by Type!

Not all operations of interest fit:

Intrinsically higher-order!

$$\limsup : \left([-\infty,\infty]^{\mathbb{R}}\right)^{\mathbb{N}} \to [-\infty,\infty]^{\mathbb{R}}$$

$$\limsup := \lambda \vec{f}.\lambda x. \limsup_{n\to\infty} f_n\, x$$

handwritten-slides/popl24-tutorial-fest-slide-84.pdf

Want

Slogan: measurability by type!

But

For higher-order building blocks

defer measurability proofs until

we resume 1^{st} order fragment => non compositional

handwritten-slides/popl24-tutorial-fest-slide-85.pdf

Plan

Def: $V \in \mathbf{Meas}$ is Standard Borel when

$$V \cong A \quad \text{for some } A \in \mathcal{B}_{\mathbb{R}}$$

the "good part" of **Meas** – the subcategory

$$\mathbf{Sbs} \hookrightarrow \mathbf{Meas}$$

handwritten-slides/popl24-tutorial-fest-slide-86.pdf

Sbs includes

- Discrete ⌜I⌝, I countable
- Countable products of Sbs:

$$\mathbb{R}^n, \mathbb{R}^{\mathbb{N}}, 2^{\mathbb{N}}, \mathbb{N}^{\mathbb{N}}$$

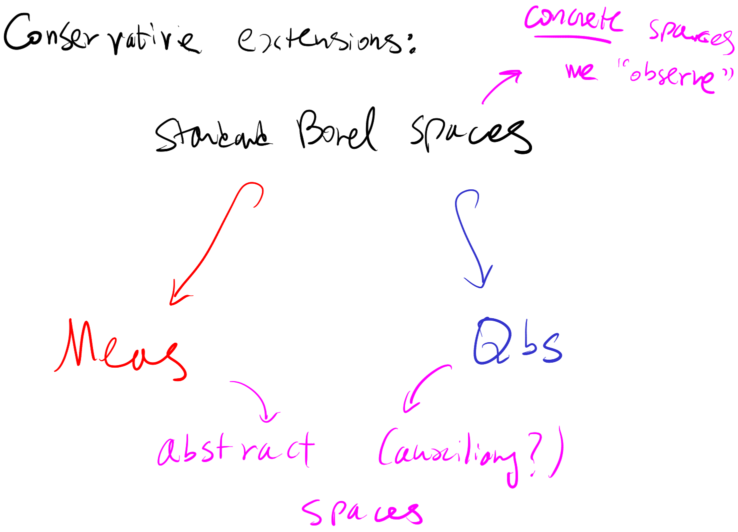
- Borel subspaces of Sbs:

$\mathbb{I} := [0,1]$ $\mathbb{R}^+ := (0,\infty)$ $\mathbb{R}_{\geq 0} := [0,\infty]$

- Countable coproducts of Sbs:

$\mathbb{W} := [0,\infty]$ $\overline{\mathbb{R}} := [-\infty,\infty]$

handwritten-slides/popl24-tutorial-fest-slide-87.pdf



handwritten-slides/popl24-tutorial-fest-slide-88.pdf