

# Probability as Logic

Prakash Panangaden<sup>1</sup>

<sup>1</sup>School of Computer Science  
McGill University

Estonia Winter School March 2015

# Outline

- 1 Introduction
- 2 Conditional probability
- 3 Measures and measurable functions
- 4 Probabilistic relations

# What am I trying to do?

- 1 Probability as logic: the central role of conditional probability. [Today]
- 2 Describe the key mathematical concepts behind modern probability: [Today] measure and integration.
- 3 Probabilistic systems and bisimulation [Lecture 2]
- 4 Metrics for probabilistic behaviour [Lecture 3]
- 5 Semantics of probabilistic programming languages [Lecture 4]

# What I am not trying to do

- Drown you in category theory.
- Discuss applications to *e.g.* Bayes nets.
- Discuss approximation theory.
- Deal with continuous time.

## A puzzle

- Imagine a town where every birth is equally likely to give a boy or a girl.  $\Pr(\text{boy}) = \Pr(\text{girl}) = \frac{1}{2}$ .
- Each birth is an *independent* random event.
- There is a family with two children.
- One of them is a boy (not specified which one), what is the probability that the other one is a boy?
- Since the births are independent, the probability that the other child is a boy should be  $\frac{1}{2}$ . Right?
- Wrong! Before you are given the additional information that one child is a boy, there are 4 *equally likely* situations: bb, bg, gb, gg.
- The possibility gg is ruled out. So of the three equally likely scenarios: bb, bg, gb, only one has the other child being a boy. The correct answer is  $\frac{1}{3}$ .
- If I had said, “The *elder* child is a boy”, then the probability that the other child is a boy is indeed  $\frac{1}{2}$ .

# The point of the puzzle

- Conditional probability is tricky!
- Conditional probability/expectation is *the* heart of probabilistic reasoning.
- Conditioning = revising probability (expectation) values in the presence of new information.
- Analogous to *inference* in ordinary logic.

# Basic Terminology

- Sample space: set of possible outcomes;  $X$ .
- Event: subset of the sample space;  $A, B \subset X$ .
- Probability:  $\Pr : X \rightarrow [0, 1]$ ,  $\sum_{x \in X} \Pr(x) = 1$ .
- Probability of an event  $A$ :  $\Pr(A) = \sum_{x \in A} \Pr(x)$ .
- $A, B$  are independent:  $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$ .
- Subprobability:  $\sum_{x \in X} \Pr(x) \leq 1$ .

# Conditional probability

## Definition

If  $A$  and  $B$  are events, the *conditional probability of  $A$  given  $B$* , written  $\Pr(A \mid B)$ , is defined by:

$$\Pr(A \mid B) = \Pr(A \cap B) / \Pr(B).$$

What happens if  $\Pr(B) = 0$ ?



# Revising probabilities

## Bayes' Rule

$$\Pr(A | B) = \frac{\Pr(B | A) \cdot \Pr(A)}{\Pr(B)}.$$

- Trivial proof: calculate from the definition.
- Example: Two coins, one fake (two heads) one OK. One coin chosen with equal probability and then tossed to yield a H. What is the probability the coin was fake?
- Answer:  $\frac{2}{3}$ .
- Bayes' rule shows how to update the *prior* probability of  $A$  with the new information that the outcome was  $B$ : this gives the *posterior* probability of  $A$  given  $B$ .

# Expectation values

- A *random variable*  $r$  is a real-valued function on  $X$ .
- The *expectation value* of  $r$  is

$$\mathbb{E}[r] = \sum_{x \in X} \Pr(x)r(x).$$

- The *conditional expectation value* of  $r$  given  $A$  is:

$$\mathbb{E}[r \mid A] = \sum_{x \in X} r(x)\Pr(\{x\} \mid A).$$

- Conditional probability is a special case of conditional expectation.

# Expectation value puzzle

- Game: 2 players, each rolls a fair 6-sided die repeatedly.
- Player 1 wins if she rolls 1 followed by 2.
- Player 2 wins if he rolls 1 followed by 1.
- Which one is expected to win first?
- More precisely: what is the expected number of rolls for each one to win?
- Hint: use *conditional* expectation.

# Logic and probability

## Kozen's correspondence

Classical logic	Generalization
Truth values $\{0, 1\}$	Probabilities $[0, 1]$
Predicate	Random variable
State	Distribution
The satisfaction relation $\models$	Integration $\int$

# Motivation

Model and reason about systems with *continuous* state spaces.

- Hybrid control systems; e.g. flight management systems.
- Telecommunication systems with spatial variation; e.g. mobile (cell) phones.
- Performance modelling.
- Continuous time systems.
- Probabilistic programming languages with recursion.

# The Need for Measure Theory

- Basic fact: There are subsets of  $\mathbf{R}$  for which no sensible notion of size can be defined.
- More precisely, there is no translation-invariant measure defined on all the subsets of the reals.

# Measurable spaces

- Countability is the key: basic analysis works well with countable summations.
- A  $\sigma$ -algebra  $\Omega$  on a set  $X$  is a family of subsets with the following conditions:
  - 1  $\emptyset, X \in \Omega$
  - 2  $A \in \Omega \Rightarrow A^c \in \Omega$
  - 3  $\{A_i \in \Omega\}_{i \in \mathbb{N}} \Rightarrow \bigcup_i A_i \in \Omega$
- Closure under countable intersections is automatic.
- $A \in \Omega$  and  $A \subset B$  or  $B \subset A$  does **not** imply  $B \in \Omega$ .
- A set with a  $\sigma$ -algebra  $(X, \Omega)$  is called a *measurable space*.

# Properties of $\sigma$ -algebras

- The collection of all subsets of  $X$  is always a  $\sigma$ -algebra.
- The intersection of *any* collection of  $\sigma$ -algebras is a  $\sigma$ -algebra.
- Thus, given *any* family  $\mathcal{F}$  of subsets of  $X$  there is a *least*  $\sigma$ -algebra containing them:  $\sigma(\mathcal{F})$ ; the  $\sigma$ -algebra *generated* by  $\mathcal{F}$ .
- For most  $\sigma$ -algebras of interest a “generic” member is hard to describe. We try to work with simpler generating families.
- Because measurable sets are closed under complementation, the character of the subject is very different from topology; *e.g.* closure under limits.



## Two Examples

- **R**: the real line. The open intervals do not form a  $\sigma$ -algebra. However, they generate one: the Borel algebra.
- Let  $\mathcal{A}$  be an “alphabet” of symbols (say finite) and consider  $\mathcal{A}^*$ : words over  $\mathcal{A}$ . Let  $\mathcal{A}^\omega$  be finite and infinite words.
- Let  $u \in \mathcal{A}^*$  and let  $u \uparrow \stackrel{\text{def}}{=} \{v \in \mathcal{A}^\omega \mid u \leq v\}$ .
- A “natural”  $\sigma$ -algebra on  $\mathcal{A}^\omega$  is the  $\sigma$ -algebra generated by  $\{u \uparrow \mid u \in \mathcal{A}^*\}$ .

# Measurable functions

- $f : (X, \Sigma) \rightarrow (Y, \Omega)$  is *measurable* if for every  $B \in \Omega$ ,  $f^{-1}(B) \in \Sigma$ .
- Just like the definition of continuous in topology.
- Why is this the definition? Why backwards?
- $x \in f^{-1}(B)$  if and only if  $f(x) \in B$ .
- No such statement for the forward image.
- Exactly the same reason why we give the Hoare triple for the assignment statement in terms of preconditions.
- Older books (Halmos) give a more general definition that is not compositional.

# Examples

- If  $A \subset X$  is a measurable set,  $\mathbf{1}_A(x) = 1$  if  $x \in A$  and 0 otherwise is called the *indicator* or *characteristic* function of  $A$  and is measurable.
- The sum and product of real-valued measurable functions is measurable.
- If we take *finite* linear combinations of indicators we get *simple* functions: measurable functions with finite range.

# Convergence properties

- If  $\{f_i : \mathbf{R} \rightarrow \mathbf{R}\}_{i \in \mathbf{N}}$  converges pointwise to  $f$  and all the  $f_i$  are measurable then so is  $f$ .
- Stark difference with continuity.
- If  $f : (X, \Sigma) \rightarrow (\mathbf{R}, \mathcal{B})$  is non-negative and measurable then there is a sequence of non-negative *simple* functions  $s_i$  such that  $s_i \leq s_{i+1} \leq f$  and the  $s_i$  converge pointwise to  $f$ .
- The secret of integration.

# Measures

- Want to define a “size” for measurable sets.
- A **measure** on  $(X, \Sigma)$  is a function  $\mu : \Sigma \rightarrow [0, \infty]$  or  $\mu : \Sigma \rightarrow [0, 1]$  (probability) such that
  - 1  $\mu(\emptyset) = 0$
  - 2  $A \cap B = \emptyset$  implies  $\mu(A \cup B) = \mu(A) + \mu(B)$ .
  - 3  $A \subset B$  implies  $\mu(A) \leq \mu(B)$ , follows.
  - 4  $\{A_i\}_{i \in \mathbb{N}} \subset \Sigma$  pairwise disjoint implies  $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$ ;  
subsumes (2).
  - 5 Actually, (4) is the only axiom needed.

# Up and down continuity

## Up continuity

Suppose  $A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots$  are all measurable and that  $A = \bigcup_{i=1}^{\infty} A_i$ . Then  $\mu(A) = \lim_{1 \rightarrow \infty} \mu(A_i)$ .

## Down continuity

Suppose  $A_1 \supseteq A_2 \supseteq \dots \supseteq A_n \supseteq \dots$  are all measurable and that  $A = \bigcap_{i=1}^{\infty} A_i$  and  $\mu(A_1) < \infty$ . Then  $\mu(A) = \lim_{1 \rightarrow \infty} \mu(A_i)$ .

Both follow from  $\sigma$ -additivity but they are not strong enough to imply it. A Choquet capacity is finitely sub-additive (or super-additive) and satisfies both continuity properties.

# Examples of measures

- $X$  countable,  $\sigma$ -algebra all subsets of  $X$ ;  $c(A)$  = number of elements in  $A$ . Counting measure; not very useful.
- $X$  any set,  $\sigma$ -algebra  $\mathcal{P}(X)$ , fix  $x_0 \in X$   $\delta_{x_0}(A) = 1$  if  $x_0 \in A$ , 0 otherwise. Dirac delta “function.”
- $X = \mathbf{R}$ ,  $\sigma$ -algebra generated by the open (or closed) intervals, the Borel sets  $\mathcal{B}$ .  $\lambda : \mathcal{B} \rightarrow \mathbf{R}^{\geq 0}$  defined as *the* measure which assigns to intervals their lengths.
- How do we know that such a measure is defined or that it is unique?
- Similarly, we can define measures on  $\mathbf{R}^n$ .

# Extension theorems

- We look for simple “well-structured” families of sets, *e.g.* intervals in  $\mathbf{R}$  and define “suitable” functions on them.
- Then we rely on extension theorems to obtain a unique measure on the generated  $\sigma$ -algebra.



# Well structured families of sets

## Definition

A **Semi-ring** **A semi-ring** of subsets of  $X$  is a family  $\mathcal{F}$  of subsets of  $X$  such that: (i)  $\emptyset \in \mathcal{F}$ , (ii)  $A, B \in \mathcal{F}$  implies  $A \cap B \in \mathcal{F}$  (iii) if  $A, B \in \mathcal{F}$  and  $A \subset B$  then there are *disjoint* sets  $C_1, \dots, C_k$  in  $\mathcal{F}$  such that

$$B \setminus A = \bigcup_{i=1}^k C_i.$$

Think of rectangles in the plane.

# The extension theorem

## Extension theorem

If  $\mathcal{F}$  is a semi-ring and  $\mu$  is a set function on  $\mathcal{F}$  with values in  $[0, \infty]$  such that  $\mu(\emptyset) = 0$ ,  $\mu$  is finitely additive and countably *subadditive*, then  $\mu$  has an extension to a measure on  $\sigma(\mathcal{F})$ .

## $\Pi$ systems

- A  $\pi$ -system is a family of sets closed under finite intersection.
- If two measures agree on a  $\pi$ -system then they agree on the generated  $\sigma$ -algebra.
- Fantastically useful, because one can work with the *much* simpler sets of a  $\pi$ -system instead of the horribly complicated sets of the generated  $\sigma$ -algebra.

# The Lebesgue integral

- Want to define  $\int f d\mu$ , where  $f$  is measurable and  $\mu$  is a measure.
- Assume that  $f$  is everywhere non-negative and bounded and  $\mu$  is a probability measure.
- If  $f$  is  $\mathbf{1}_A$  then we *define*  $\int \mathbf{1}_A d\mu = \mu(A)$ .
- If  $f$  is  $r \cdot \mathbf{1}_A$  then we *define*  $\int f d\mu = r \cdot \mu(A)$ .
- If  $f = \sum_{i=1}^k r_i \mathbf{1}_{A_i}$  (simple function) then we define

$$\int f d\mu = \sum_{i=1}^k r_i \cdot \mu(A_i).$$

- Need to check that it does not matter how we write such an  $f$  as a simple function.
- There are some subtleties if sets can have infinite measure but these do not arise if we are dealing with probability measures and bounded measurable functions.

# The Lebesgue integral II

## The Lebesgue integral

If  $f$  is non-negative and measurable and  $\mu$  a probability measure we define

$$\int f d\mu = \sup \int s d\mu$$

where the *sup* is over all *simple* non-negative functions below  $f$ .

- One can define integrals of general functions by splitting them into positive and negative pieces.
- One can prove that the integral is linear and monotone.

# Monotone convergence

## The monotone convergence theorem

Let  $\{f_n\}$  be a sequence of measurable functions on  $X$  such that (1)  
 $\forall x \in X, 0 \leq f_1(x) \leq f_2(x) \leq \dots \leq f_n(x) \leq \dots \leq f(x)$  and (2)  
 $\forall x \in X, \sup_n f_n(x) = f(x)$  then

$$\sup_n \int f_n d\mu = \int f d\mu.$$

- Should remind you of things in domain theory.
- The integral is continuous in an order-theoretic sense.

# The monotone convergence mantra

- Want to prove  $\int \mathcal{E}(f) d\mu = \int \mathcal{E}'(f) d\nu$ .
- Prove it for the special case  $f = \mathbf{1}_A$ , usually easy.
- Then automatic for simple functions by linearity.
- Then automatic for non-negative bounded measurable functions by the monotone convergence theorem.
- Then clear for general bounded measurable functions.

# The mantra in action

- Suppose  $T : (X, \Sigma, \mu) \rightarrow (Y, \Omega, \nu)$  measurable and measure preserving:  $\forall B \in \Omega \nu(B) = \mu(T^{-1}(B))$ .
- $f : Y \rightarrow \mathbf{R}$  is measurable.
- Want to show  $\forall B \in \Omega, \int_B f d\nu = \int_{T^{-1}(B)} T \circ f d\mu$ .
- Assume that  $f$  is  $\chi_A$  for some  $A \in \Omega$ .
- Left-hand Side is  $\nu(A \cap B)$ .
- Right-hand side is  $\mu(T^{-1}(A) \cap T^{-1}(B)) = \mu(T^{-1}(A \cap B)) = \nu(A \cap B)$ .
- And that's all we have to do!!



# Ordinary binary relations

- $R : A \rightarrow B$  is just  $R \subseteq A \times B$
- Natural converse relation  $R^\circ : B \rightarrow A$ .
- Composition:  $R_1 : A \rightarrow B, R_2 : B \rightarrow C$  then  $R_1 \circ R_2 = \{(x, z) \mid \exists y \in B, xR_1y \text{ and } yR_2z\}$ .
- Close relation with the powerset construction:
- $\hat{R} : A \rightarrow \mathcal{P}(B)$  is an equivalent description of  $R$ .

# Markov kernels

- A *Markov kernel* on a measurable space  $(S, \Sigma)$  is a function  $h : S \times \Sigma \rightarrow [0, 1]$  with (a)  $h(s, \cdot) : \Sigma \rightarrow [0, 1]$  a (sub)probability measure and (b)  $h(\cdot, A) : S \rightarrow [0, 1]$  a measurable function.
- Though apparently asymmetric, these are the probabilistic analogues of binary relations
- and the uncountable generalization of a matrix.
- They describe transition probabilities in situations where a “point-to-point” approach does not make sense.
- Composition:  $k$  “after”  $h$ ,  $(k \circ h)(x, A) = \int k(x', A) dh(x, \cdot)$ , where we are integrating the variable  $x'$  using the measure  $h(x, \cdot)$ .
- We construct these things using a major theorem (the Radon-Nikodym theorem).

# Probabilistic relations

- Want to define  $R : (X, \Sigma) \rightarrow (Y, \Omega)$ .
- Define a probabilistic relation  $R$  from  $X$  to  $Y$  to be a Markov kernel of type  $R : X \times \Omega \rightarrow [0, 1]$  with the same measurability conditions.
- Given relations  $R_1 : (X, \Sigma) \rightarrow (Y, \Omega)$  and  $R_2 : (Y, \Omega) \rightarrow (Z, \Lambda)$  we define  $R_2 \circ R_1$  ( $R_1; R_2$ ) as
- $(R_2 \circ R_1)(x, C \in \Lambda) = \int R_2(y, C) dR_1(x, \cdot)$ .
- Just like the formula for composing ordinary relations with integration for  $\exists$ .
- Converse is tricky and requires more machinery and more structure.

# The category **SRel**

- Objects: measurable spaces  $(X, \Sigma_X)$
- Morphisms:  $h : (X, \Sigma_X) \rightarrow (Y, \Sigma_Y)$  are Markov kernels  $h : X \times \Sigma_Y \rightarrow [0, 1]$ .
- Composition:  $h : X \rightarrow Y, k : Y \rightarrow Z$  then  $\forall x \in X, C \in \Sigma_Z$ ,  
 $(k \circ h)(x, C) = \int_Y k(y, C)h(x, dy)$ .
- The identity morphisms:  $id : X \rightarrow X$  is  $\delta(x, A)$ .
- Prove associativity of composition by using the monotone convergence mantra.
- It has countable coproducts; very useful for semantics.
- Unlike **Rel** this category is not self dual.

# The Gíry Monad

- Define  $\Pi : \mathbf{Mes} \rightarrow \mathbf{Mes}$  by  $\Pi((X, \Sigma_X)) = \{\nu \mid \nu : \Sigma_X \rightarrow [0, 1]\}$  where  $\nu$  is a *subprobability* measure on  $X$ .
- Actually, Gíry used probability measures; I made the small change to subprobability measures in order to adapt it to programming language semantics.
- But  $\Pi(X)$  has to be a measurable space not just a set.
- For every  $A \in \Sigma_X$  we define  $\text{ev}_A : \Pi(X) \rightarrow [0, 1]$  by  $\text{ev}_A(\nu) = \nu(A)$ .
- We define the  $\sigma$ -algebra on  $\Pi(X)$  to be the *least*  $\sigma$ -algebra making all the  $\text{ev}_A$  measurable.
- Given  $f : X \rightarrow Y$  define  $(\Pi(f)(\nu))(B \in \Sigma_Y) = \nu(f^{-1}(B))$ .
- Need natural transformations:  $\eta : I \rightarrow \Pi$  and  $\mu : \Pi^2 \rightarrow \Pi$ .
- $\eta_X(x) = \delta(x, \cdot)$
- $\mu_X(\Omega \in \Pi^2(X)) = \lambda B \in \Sigma_X. \int \text{ev}_B d\Omega_{\Pi(X)}$ .

# The Kleisli category of $\Pi$

- If  $T : \mathcal{C} \rightarrow \mathcal{C}$  is a monad, then  $\mathcal{C}_T$  has the same objects as  $\mathcal{C}$  and the morphisms in  $\mathcal{C}_T$  from  $X$  to  $Y$  are morphisms in  $\mathcal{C}$  from  $X$  to  $TY$ .
- For the powerset monad we get morphisms  $X \rightarrow \mathcal{P}(Y)$  which we recognize as just binary relations.
- Here we get  $h : X \rightarrow \Pi(Y)$  or  $h : X \rightarrow (\Sigma_Y \rightarrow [0, 1])$  or  $h : X \times \Sigma_Y \rightarrow [0, 1]$ .
- These are exactly the Markov kernels.
- How do we prove associativity of composition of Markov kernels?
- Use the monotone convergence mantra Luke!