Probability as Logic

Prakash Panangaden[1]

[1]School of Computer Science
McGill University

Estonia Winter School March 2015

Outline

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 isa 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: $\mathsf{Pr} : X \longrightarrow [0, 1]$, $\sum_{x \in X} \mathsf{Pr}(x) = 1$.
- Probability of an event A: $\mathsf{Pr}(A) = \sum_{x \in A} \mathsf{Pr}(x)$.
- A, B are independent: $\mathsf{Pr}(A \cap B) = \mathsf{Pr}(A) \cdot \mathsf{Pr}(B)$.
- Subprobability: $\sum_{x \in X} \mathsf{Pr}(x) \leq 1$.

Conditional probability

Definition

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

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

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

Revising probabilities

Bayes' Rule

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

- 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

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

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

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

- 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 σ-algebra Ω 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. $\{A_i \in \Omega\}_{i \in \mathbf{N}} \Rightarrow \bigcup_i A_i \in \Omega$
- 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 σ-algebra (X, Ω) is called a *measurable space*.

Properties of σ-algebras

- The collection of all subsets of X is always a σ-algebra.
- The intersection of *any* collection of σ-algebras is a σ-algebra.
- Thus, given *any* family $\mathcal{F}$ of subsets of X there is a *least* σ-algebra containing them: $\sigma(\mathcal{F})$; the σ-algebra *generated* by $\mathcal{F}$.
- For most σ-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

- $\mathbf{R}$: the real line. The open intervals do not form a σ-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" σ-algebra on $\mathcal{A}^\omega$ is the σ-algebra generated by $\{u \uparrow \mid u \in \mathcal{A}^*\}$.

Measurable functions

- $f : (X, \Sigma) \to (Y, \Omega)$ 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 $\{f_i : \mathbf{R} \longrightarrow \mathbf{R}\}_{i \in \mathbf{N}}$ converges pointwise to f and all the f_i are measurable then so is f.
- Stark difference with continuity.
- If $f : (X, \Sigma) \longrightarrow (\mathbb{R}, \mathcal{B})$ 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, Σ) is a function $\mu : \Sigma \longrightarrow [0, \infty]$ or $\mu : \Sigma \longrightarrow [0, 1]$ (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. $\{A_i\}_{i \in \mathbf{N}} \subset \Sigma$ 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 $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq \ldots$ are all measurable and that $A = \bigcup_{i=1}^{\infty} A_i$. Then $\mu(A) = \lim_{1 \longrightarrow \infty} \mu(A_i)$.

Down continuity

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

Both follow from σ-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, σ-algebra all subsets of X; $c(A) =$ number of elements in A. Counting measure; not very useful.
- X any set, σ-algebra $\mathcal{P}(X)$, fix $x_0 \in X$ $\delta_{x_0}(A) = 1$ if $x_0 \in A$, 0 otherwise. Dirac delta "function."
- $X = \mathbf{R}$, σ-algebra generated by the open (or closed) intervals, the Borel sets $\mathcal{B}$. $\lambda : \mathcal{B} \longrightarrow \mathbf{R}^{\geq 0}$ 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 σ-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 $C_1, \ldots, C_k$ 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 μ is a set function on $\mathcal{F}$ with values in $[0, \infty]$ such that $\mu(\emptyset) = 0$, μ is finitely additive and countably *subadditive*, then μ has an extension to a measure on $\sigma(\mathcal{F})$.

Π systems

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

The Lebesgue integral

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

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

- 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 μ a probability measure we define

$$\int f \mathrm{d}\mu = \sup \int s \mathrm{d}\mu$$

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) $\forall x \in X,\ 0 \leq f_1(x) \leq f_2(x) \leq \ldots \leq f_n(x) \leq \ldots \leq f(x)$ and (2) $\forall x \in X,\ \sup_n f_n(x) = f(x)$ then

$$\sup_n \int f_n \mathrm{d}\mu = \int f \mathrm{d}\mu.$$

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

The monotone convergence mantra

- Want to prove $\int \mathcal{E}(f)\mathrm{d}\mu = \int \mathcal{E}'(f)\mathrm{d}\nu$.
- 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 $T : (X, \Sigma, \mu) \longrightarrow (Y, \Omega, \nu)$ measurable and measure preserving: $\forall B \in \Omega \; \nu(B) = \mu(T^{-1}(B))$.
- $f : Y \longrightarrow \mathbf{R}$ is measurable.
- Want to show $\forall B \in \Omega, \int_B f \mathrm{d}\nu = \int_{T^{-1}(B)} T \circ f \mathrm{d}\mu$.
- Assume that f is χ_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

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

Markov kernels

- A *Markov kernel* on a measurable space (S, Σ) is a function $h : S \times \Sigma \longrightarrow [0, 1]$ with (a) $h(s, \cdot) : \Sigma \longrightarrow [0, 1]$ a (sub)probability measure and (b) $h(\cdot, A) : S \longrightarrow [0, 1]$ a measurable function.
- Though apparantly 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, $(k \circ h)(x, A) = \int k(x', A)\mathrm{d}h(x, \cdot)$, 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 $R : (X, \Sigma) \longrightarrow (Y, \Omega)$.
- Define a probabilistic relation R from X to Y to be a Markov kernel of type $R : X \times \Omega \longrightarrow [0, 1]$ with the same measurability conditions.
- Given relations $R_1 : (X, \Sigma) \longrightarrow (Y, \Omega)$ and $R_2 : (Y, \Omega) \longrightarrow (Z, \Lambda)$ we define $R_2 \circ R_1$ $(R_1; R_2)$ as
- $(R_2 \circ R_1)(x, C \in \Lambda) = \int R_2(y, C) \mathrm{d}R_1(x, \cdot)$.
- Just like the formula for composing ordinary relations with integration for $\exists$.
- Converse is tricky and requires more machinery and more structure.

The category $\mathbf{SRel}$

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

The Gíry Monad

- Define $\Pi : \mathbf{Mes} \longrightarrow \mathbf{Mes}$ by $\Pi((X, \Sigma_X)) = \{\nu \mid \nu : \Sigma_X \longrightarrow [0, 1]\}$ where ν 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 $\mathrm{ev}_A : \Pi(X) \longrightarrow [0, 1]$ by $\mathrm{ev}_A(\nu) = \nu(A)$.
- We define the σ-algebra on $\Pi(X)$ to be the *least* σ-algebra making all the ev_A measurable.
- Given $f : X \longrightarrow Y$ define $(\Pi(f)(\nu))(B \in \Sigma_Y) = \nu(f^{-1}(B))$.
- Need natural transformations: $\eta : I \longrightarrow \Pi$ and $\mu : \Pi^2 \longrightarrow \Pi$.
- $\eta_X(x) = \delta(x, \cdot)$
- $\mu_X(\Omega \in \Pi^2(X)) = \lambda B \in \Sigma_X. \int \mathrm{ev}_B \mathrm{d}\Omega_{\Pi(X)}$.

The Kleisli category of Π

- If $T : \mathcal{C} \longrightarrow \mathcal{C}$ 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 $X \longrightarrow \mathcal{P}(Y)$ which we recognize as just binary relations.
- Here we get $h : X \longrightarrow \Pi(Y)$ or $h : X \longrightarrow (\Sigma_Y \longrightarrow [0, 1])$ or $h : X \times \Sigma_Y \longrightarrow [0, 1]$.
- These are exactly the Markov kernels.
- How do we prove associativity of compostion of Markov kernels?
- Use the monotone convergence mantra Luke!