Part II: Markov Processes

Prakash Panangaden
McGill University

How do we define random processes on continuous state spaces?

How do we define conditional probabilities on continuous state spaces?

How do we define probabilities on continuous state spaces?

Points are useless!

The probability of a set may be 1 but the probability of *every single* point could be 0.

We understand countable sums, but
uncountable sums are handled by integration.

Consider the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Of course this assumes that $P(B) \neq 0$.

But we will want to write transition probabilities like $\tau(x, A)$: the probability of landing in the set A given that the system starts at x.

We can consider a family of sets $B_1 \supset B_2 \supset \ldots$ with $\cap_i B_i = \{x\}$ and try to define some kind of "limit":

$$\lim_{i \to \infty} \frac{P(A \cap B_i)}{P(B_i)}.$$ This rarely makes sense!

Basic fact: There are subsets of **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.

What is measure theory?

We want to assign a “size” to sets so that we can use it for quantitative purposes, like integration or probability.

We could count the number of points but this is useless for the continuum.

We want to generalize the notion of “length” or “area.”

What is the “length” of the rational numbers between 0 and 1?

We want a consistent way of assigning sizes to these and (all?) other sets.

Alas! Not all sets can be given a sensible notion of size that generalizes the notion of length of an interval.

We take a family of sets satisfying “reasonable” axioms and deem them to be “measurable.”

Countable unions are the key.

σ-algebras

A σ-algebra on a set X is a family Σ of subsets of X satisfying:

- $X, \emptyset \in \Sigma$
- $A \in \Sigma$ implies $A^c \in \Sigma$

and

- $A_i \in \Sigma$ where $\{A_i | i \in I\}$ is a countable family,

  implies that $\bigcup_{i \in I} A_i \in \Sigma$.

It follows that Σ is closed under countable intersections.

Basic Facts

The intersection of any collection of σ-algebras on a set is another σ-algebra.

Thus given any family of sets $\mathcal{B}$ there is a least σ-algebra containing $\mathcal{B}$: the σ-algebra *generated* by $\mathcal{B}$.

Measurable sets are complicated beasts, we often want to work with the sets of family of simpler sets that generate the σ-algebra.

The σ-algebra generated by the intervals in **R** is called the *Borel* algebra.

There is a larger σ-algebra containing the Borel sets called the Lebesgue σ-algebra; we will not use it.

What are the "right" functions between measurable spaces?

Let $f : X \longrightarrow Y$ be a function and let Σ be a σ-algebra on Y. The sets of the form $\{f^{-1}(A) | A \in \Sigma\}$ form a σ-algebra on X.

σ-algebras behave well under inverse image.

A function f from a σ-algebra (X, Σ_X) to a σ-algebra (Y, Σ_Y) is said to be **measurable** if $f^{-1}(B) \in \Sigma_X$ whenever $B \in \Sigma_Y$.

A **measure** (**probability measure**) μ on a measurable space (X, Σ) is a function from Σ (a set function) to $[0, \infty]$ ($[0, 1]$), such that if $\{A_i | i \in I\}$ is a countable family of pairwise disjoint sets then

$$\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i).$$

In particular if I is empty we have

$$\mu(\emptyset) = 0.$$

A set equipped with a σ-algebra and a measure defined on it is called a **measure space**.

Properties of Measures

1. *If* $A \subseteq B$ *then* $\mu(A) \leq \mu(B)$.

2. *If* $A_1 \subseteq A_2 \subseteq \ldots A_n \subseteq \ldots$ *and* $\cup_i A_i = A$ *then* $\lim_{i \to \infty} \mu(A_i) = \mu(A)$.

3. *If* $A_1 \supseteq A_2 \supseteq \ldots A_n \supseteq \ldots$ *and* $\cap_i A_i = A$ *then* $\lim_{i \to \infty} \mu(A_i) = \mu(A)$, *if* $\mu(A_1)$ *is finite.*

Convexity: $$\mu(\bigcup_i B_i) \leq \sum_i \mu(B_i)$$

B_i a countable family.

For *any* subset of **R** we define *outer measure* as the infimum of the total length of the intervals of any covering family of intervals.

The rationals have outer measure zero.

This is not additive so it does not give a measure defined on all sets.

Lebesgue Measure I

We begin as follows

$$m((a,b)) = m((a,b]) = m([a,b)) = m([a,b]) = b-a$$

where a and b are real numbers. If we have a pairwise disjoint collection of intervals $\{I_k | k \in \mathcal{K}\}$ we define

$$m(\bigcup_{k\in\mathcal{K}} I_k) = \sum_{k\in\mathcal{K}} m(I_k)$$

If A is arbitrary we define $m^*(A)$ to be the **inf** of m over all families of intervals that cover A. Clearly (?) if A is countable $m^*(A)$ is zero.

m^* is not countably additive but it does satisfy convexity.

An **outer measure** μ^* on a set X is a set-function defined on *all* subsets such that:

1. $\mu^*(\emptyset) = 0$,

2. $A \subseteq B \Rightarrow \mu^*(A) \leq \mu^*(B)$ *and*

3. *for any countable family of subsets of* X*, say* $\{A_i\}$*, we have*

$$\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i).$$

Theorem *Let X be a set and let μ^* be an outer measure defined on X. Denote by Σ the collection of all subsets, say A, such that for every subset E of X we have*

$$\mu^*(E) = \mu^*(A \cap E) + \mu^*(A^c \cap E).$$

For all A in Σ define $\mu(A) = \mu^*(A)$. Then (X, Σ, μ) is a measure space.

This is how we construct the usual (Lesbesgue) measure on $\mathbf{R}$.

Consider the set of all infinite *sequences* from some alphabet $\mathcal{A}$.

Call this set $\mathfrak{S}$.

Let α be a *finite* sequence.

$\alpha\uparrow \stackrel{\text{def}}{=} \{s \in \mathfrak{S} | \alpha \text{ is a prefix of } s\}$.

We define Σ on $\mathfrak{S}$ as the σ-algebra *generated by* all sets of the form $\alpha\uparrow$.

The sets of the σ-algebra are a pain to describe, but the generating sets are easy to describe.

One can show that measures defined on these generating sets extend to a measure on the whole σ-algebra.

Integration

Two approaches:

- **Riemann:** Divide up the domain, functions have to be "well-behaved"; continuous (Lebesgue)-almost everywhere. The integral has poor convergence properties. The basis of all numerical algorithms; much more constructive in flavour.

- **Lebesgue:** Divide up the range, functions can be perverse (measurable), very good convergence properties. Hard to see how to make it constructive.

1. First define the integral of characteristic functions

$$\int \chi_A \mu =_{df} \mu(A).$$

2. Using linearity, define the integral of simple functions

$$\int s\mu = \sum_{i=1}^{n} a_i \mu(A_i)$$

where

$$s = \sum_{i=1}^{n} a_i \chi_{A_i}.$$

3. Using the fact that all positive measurable functions are sups of *increasing* sequences of simple functions we define the integral of a positive measurable function as a sup:

$$\int f\mu = \sup\{\int s\mu : s \text{ is simple and } s \leq f\}.$$

Radon-Nikodym Theorem

If μ, ν are measures and for any measurable set A, whenever $\nu(A) = 0$ then also $\mu(A) = 0$, we write $\mu << \nu$ and say that μ is *absolutely continuous* with respect to ν.

Theorem: If $\mu << \nu$ then there is a measurable function h such that for any measurable set B:

$$\int_B h \mathrm{d}\mu = \nu(B).$$

Any other function with the same property will agree with h except on a set of points with ν-measure 0. We say that h is *almost* unique. We write $h = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

Conditioning wrt a sub-σ-algebra

Suppose that (X, Σ, P) is a probability space describing a random process.

Suppose that you find out the following *partial information* about the outcome: you know in which member of a sub-σ-algebra $\Lambda \subset \Sigma$ the outcome lies.

For $B \in \Lambda$ and fixed $A \in \Sigma$ define $Q(B) = P(A \cap B)$. Clearly $Q << P$. So by the Radon-Nikodym theorem:

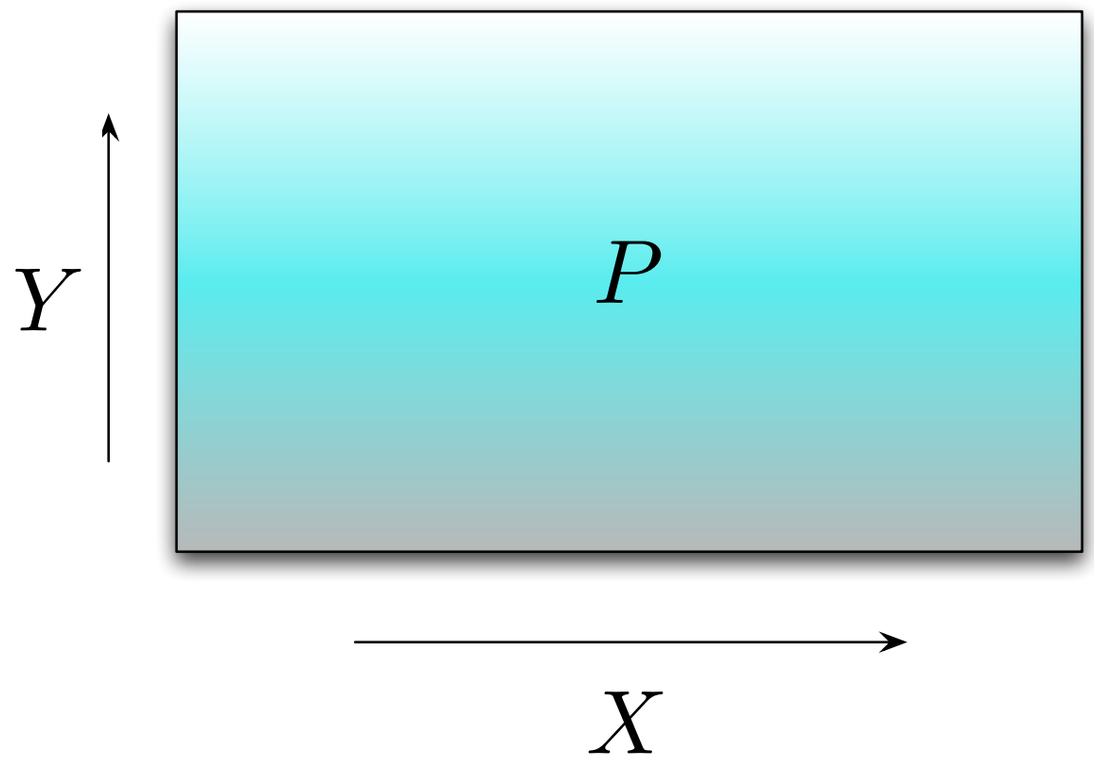
$$\exists f \text{ such that } P(A \cap B) = \int_B f \mathrm{d}P, \ \ \forall B \in \Lambda.$$

We write $P[A||\Lambda](\cdot)$ for f. This is called the conditional probability for A given Λ.

Note that it is a Λ-measurable function. Hence it is "smoother" than a Σ-measurable function.

Note that it is not unique, but defined up to sets of P-measure 0. The different functions are called *versions*.

Consider a joint distribution P of two continuous variables.



Let Σ be the usual measurable sets in the plane and let Λ be sets of the form $A \times Y$ where A is a measurable subset of X.

The conditional probability $P[A||\Lambda](\cdot)$ will be the estimate of the distribution if you know perfectly the X result. Since it is Λ-measurable, it has to be constant along the Y-axis.

Markov Kernels

The previous example is ideally suited to the case where the two variables are "the last state" and "the present state" of a transition system.

We use the notation $\tau(x, A)$ to mean the conditional probability density for the next state to be in the set A given that the present state is x.

For fixed A it is a measurable function of x and for fixed x it is a measure; *almost everywhere*.

With suitable *topological* assumptions (Polish or analytic) one can choose versions so that τ is a measure everywhere.

They are called Markov kernels or stochastic kernels.

Probabilistic relations

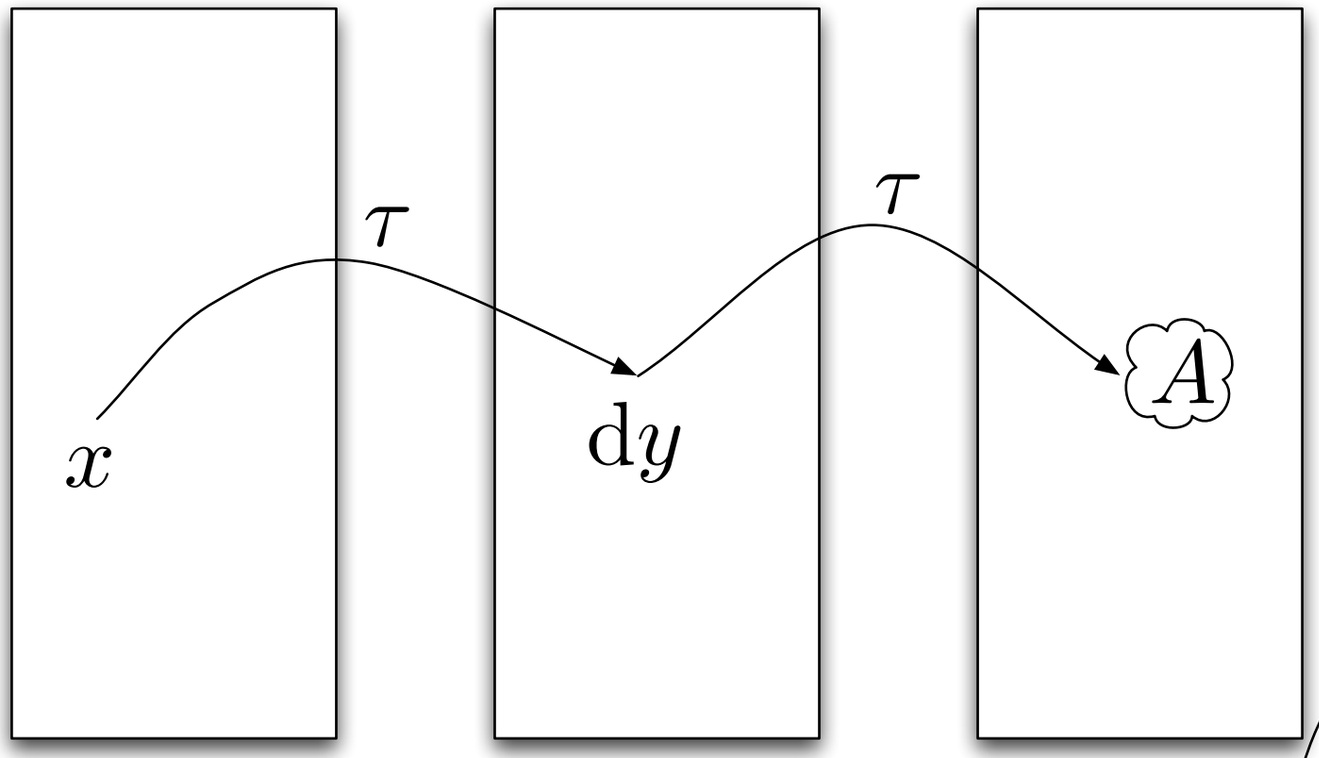
I called them "probabilistic relations" in 1998 by analogy with ordinary relations.

Even though they are not symmetric they are connected to "fuzzy" powersets in the same way that relations are connected to powersets.

They compose by integration.

They are also just like matrices.

How do you get from x to A in two steps?



$$\tau^{(2)}(x, A) = \int_S \tau(x, \mathrm{d}y), \tau(y, A).$$

for fixed A, τ is a measurable function.

for fixed x, τ is a measure.

Stochastic Processes

Definition *A* **stochastic process** *is an indexed family of random variables* $X_t : \Omega \longrightarrow \mathbf{R}$ *where* $(\Omega, \mathcal{B}, P)$ *is a probability space and* $t \in T$ *is the indexing set.*

Think of Ω as the space of *trajectories*.

More generally, instead of $\mathbf{R}$, we can have any measurable space as the state space.

How does this connect with the state-transformation view?

Joint distributions can be defined:

$$P_{t_1 \ldots t_n}(B) = P(\{x | (X_{t_1}(x), \ldots, X_{t_n}(x)) \in B\}).$$

First Kolmogorov consistency requirement

$$P_{t_1 \ldots t_n t_{n+1}}(B \times \mathbf{R}) = P_{t_1 \ldots t_n}(B).$$

The second reqirement says that the distributions behave the way they should under permutation of the variables.

Fundamental Theorem: Any family of finite-dimensional probability distributions satisfying these requirements can be realized as a stochastic process.

Markov Processes

We write $P(A_{n+1}|x_1, \ldots, x_n)$ for the conditional probability that the system is in the set A_{n+1} given that at time t_1 it was at x_1 etc.

In a Markov Process we only depend on the last state:

$$P(A_{n+1}|x_1, \ldots, x_n) = P(A_{n+1}|x_n).$$

These are precisely what can be described as matrices or stochastic kernels.